Schools, Language, and Nations: Evidence From a Natural Experiment in France

Guillaume Blanc† and Masahiro Kubo, *Brown University**

†**Job Market Paper 1**

December 27, 2021
Link to the latest version

Abstract

This paper studies nation-building. We explore the role of state-sponsored education in the adoption of a common language and the formation of a national identity in a fragmented society. At the time of the French Revolution, only ten percent of the population spoke the French language in France. We digitize a novel municipality-level dataset on spoken languages to document the process of homogenization in the nineteenth century. Using a regression discontinuity design, we demonstrate that state-sponsored education brought about the homogenization of language. Then, we study the geographical origins of the French language and the heterogeneous effects of schools. We find that elites were an important driver of homogenization. Finally, we document a persistent impact of nation-building on national identity and preferences for political centralization, with increased participation in the Resistance during World War II and votes against the 1969 referendum on regionalization.

***JEL codes**: D83,H52,I25,N43,P16,Z13*
***Keywords**: nation-building, language, schools*

1 INTRODUCTION

The emergence of modern nation-states dramatically reshaped fragmented societies and brought about the homogenization of populations in the nineteenth century. This process

fostered social cohesiveness and interactions, political stability, and the provision of public goods.

In this research, we study the making of the French nation. Using a natural experiment and drawing on a novel, detailed historical dataset, we document the process of homogenization and nation-building and empirically establish, for the first time, that state-sponsored education played a substantial role in the adoption of the French language and in the formation of a national identity in France.

At the time of the French Revolution, French was a foreign language to the vast majority of the population. The French language was one of more than forty-six dialects and languages spoken historically. Fernand Braudel (1986) argues that France was "the dazzling triumph of the heterogeneous" (p. 38). Today, French is the common language. Eugen Weber (1976) and Eric Hobsbawm (1990) provide fascinating accounts of the process of homogenization in their seminal books *Peasants into Frenchmen* and *Nations and Nationalism since 1780*; yet statistical evidence for this change is missing because of data limitations on the adoption of the French language and on the use of and the similarity between the languages or dialects spoken.

We document homogenization at a particularly granular level. We rely on a detailed municipality-level survey of the languages spoken in France in 1900, the *Atlas Linguistique de la France*. This dataset records the common (everyday) language and pronunciation for 1,920 unique words or expressions in 577 municipalities at the time. We digitize the pronunciation of a representative sample of words in the *Atlas* selected from a standard collection of words with universal meanings, the Swadesh list, and we measure linguistic distance across 166,176 pairs of municipalities using a standard measure of distance. In the bulk of the analysis, we focus on the linguistic distance from French, or francien, the dialect spoken in the region near Paris, which historians believe to be the place of origin of the French language—in the historical *langues d'oïl* region (Lodge, 1993).

Using a regression discontinuity design exploiting quasi-experimental variation in the building of schools, we establish that the provision of state-sponsored education brought about the adoption of the French language.

The *Loi Guizot du 28 juin 1833*, or Guizot Act, named after the Minister of Public Instruction François Guizot, laid the foundations of mass public primary schooling in France by requiring towns with more than five hundred inhabitants to build a school for boys. The law mandated major changes in the curriculum and in the training of schoolteachers, with the teaching of the national language and a national history, the creation and distribution of textbooks approved by the state, and the creation of both a national system of teacher-training colleges and a body of nationally recruited school inspectors. Paul Lorain (1837), a close advisor of Guizot, argued at the time that "each school shall be a colony of French language in a conquered land" (p. 29), while Guizot (1833) himself wrote that "through the teaching of the French language, primary schools will instill and spread the spirit and

unity of nationality everywhere" (p. 102).

In addition, we demonstrate that the regression discontinuity design allows us to estimate the effect of the treatment on the change in language over time and to estimate a lower bound on the speed (rate) of homogenization at the threshold from 1833 to 1900, although we only observe language in a single cross-section. We rely on the estimated percentage decrease resulting from the policy, relative to the counterfactual, to estimate the lower bound, and we find that state-sponsored education contributed to a total decrease in linguistic distance from French of at least 20 percent during that period, a particularly rapid homogenization.

Our results are robust across a wide range of standard tests and alternative or placebo specifications. We first provide evidence to support our empirical design, by testing for manipulation at the threshold and showing that observables are balanced at the threshold. Then, we show the robustness of the estimated coefficients to different methods of estimation, kernels, polynomial order, and bias correction; to using alternative bandwidths; to accounting for individual-level characteristics; and to using different distance metrics and measuring distance across the entire corpus of maps from the *Atlas* instead of only the Swadesh list. We also vary the threshold for schools and show that only the five-hundred-inhabitant cutoff, based on population at the time of the law, returns statistically significant estimates.

We also discuss the interpretation of our results and explore the heterogeneous effect of state-sponsored education to study the drivers of homogenization. We show that schools primarily affected language, not pronunciation, and gave rise to a process of language shift in the historical *langues d'oc* region. Additionally, we invert the regression discontinuity analysis across the universe of potential references for standard language to empirically trace out the geographical origins of the French language. We find that schools spread the language spoken near Paris and in the Loire Valley—that is, the language of the elites. This represents the first empirical evidence on the origins of French and confirms anecdotal evidence by historians (e.g. Lodge, 1993). Finally, we study the heterogeneous effect of school openings. We find that elites were among the most important drivers of homogenization, and we further document the role played by fiscal capacity, state legitimacy, and educational demand. Our results suggest an interaction between state-building and nation-building, and emphasize the role of the ruling class in the process of nation-building.

Finally, we document a persistent impact of schools on social interactions and national identity, suggesting that nation-building dramatically reshaped society and that the adoption of the French language was indeed "essential to the concept of France" (Hobsbawm, 1990, p. 60). We show that linguistic distance between municipalities was negatively associated with migration and trade across pairs of districts or *départements* even after accounting for gravity with geographical distance and population. Then, using municipality-level data on the places of birth of World War II Resistance heroes and on votes against the regional-

ization of political authority in a 1969 referendum, we find that state-sponsored education increased the salience of national identity and preferences for centralization in the twentieth century.[1]

Our paper makes several contributions. We identify the role of state-sponsored education in the homogenization of language and empirically document the process of nation-building described by Weber (1976). In particular, we contribute to an important and growing literature on nation-building (Alesina, Giuliano and Reich, 2021; Bazzi et al., 2019; Dell and Querubin, 2018). We also contribute to a large literature on identity, assimilation, integration, and linguistic diversity (Dickens, 2021; Michalopoulos, 2012) by digitizing a novel, detailed historical dataset on spoken language. Our paper adds evidence to a literature on the determinants of identity (Blouin and Mukand, 2019). More particularly, we support the constructivist view, in which identity is socially determined (Gellner, 1983a; Smith et al., 1991). Finally, we contribute to the literature on the effects of education and school curricula on beliefs and identity (Algan, Cahuc and Shleifer, 2013; Cantoni et al., 2017).

2   Historical Background and Related Literature

2.1   Historical Background

"France is diversity", "the dazzling triumph of the heterogeneous" (p. 38) are some of the words used by historian Braudel (1986) to describe France. More than forty different languages or dialects were spoken historically, and French was unknown to most until the twentieth century. Hobsbawm (1990) and Grégoire (1794) argue that only 10 percent of the population spoke French fluently at the time of the French Revolution. The vast majority in the south of France spoke Occitan and its dialects (*lenga d'òc*, or *langues d'oc*), while in the northern part the population spoke *langues d'oïl*, including the French language. A multitude of other languages, such as Alsatian, Basque, Breton, Catalan, Flemish, or Francoprovençal, were spoken in the rest of the country. We display the spatial distribution of historical languages and dialects in Appendix Figure A3.[2] Figure 1 displays the poem "Mirèio" in Provençal, a dialect of Occitan, and its translation into French in order to comprehend some of the differences between these languages.[3]

One of the first attempts to drive the adoption of the French language dates back to the Ordinance of Villers-Cotterêts of August 10, 1539. It called for the use of French in all legal documents including parish records, notarized contracts, and legislation. The ordinance made French the language of the government and of local administrations, and it

---

[1]Following Dehdari and Gehring (forthcoming), we digitize municipality-level data on the referendum, which is otherwise unavailable, from regional newspapers that published the results the day after the election.

[2]There is no formal distinction separating a dialect from a language (see Ginsburgh and Weber (2020)); rather, "a language is a dialect with an army and navy" (Weinreich, 1945).

[3]Similarly, Appendix Figure A4 displays street names in Provençal and Alsacien.

**Figure 1:** Mirèio (1859), Frédéric Mistral

MIRÈIO

CANT PROUMIÈ

LOU MAS DI FALABREGO.

Cante uno chato de Prouvènço.
Dins lis amour de sa jouvènço,
A travès de la Crau, vers la mar, dins li bla,
Umble escoulan dóu grand Oumèro,
Iéu la vole segui. Coume èro
Rèn qu'uno chato de la terro,
En foro de la Crau se n'es gaire parla.

Emai soun front noun lusiguèsse
Que de jouinesso ; emai n'aguèsse
Ni diadèmo d'or ni mantèu de Damas,
Vole qu'en glòri fugue aussado
Coume uno rèino, e caressado
Pèr nosto lengo mespresado,
Car cantan que pèr vautre, o pastre e gènt di mas!

(Panel A) Provençal
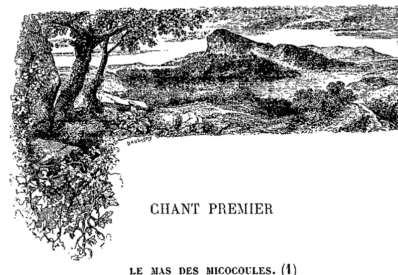
CHANT PREMIER

LE MAS DES MICOCOULES. (1)

Je chante une jeune fille de Provence. — Dans
les amours de sa jeunesse, — à travers la Crau (2),
vers la mer, dans les blés, — humble écolier du
grand Homère, — je veux la suivre. Comme c'é-
tait — seulement une fille de la glèbe, — en
dehors de la Crau il s'en est peu parlé.

Bien que son front ne brillât — que de jeunes-
se ; bien qu'elle n'eût — ni diadème d'or ni man-
teau de Damas, — je veux qu'en gloire elle soit
élevée — comme une reine, et caressée — par
notre langue méprisée, — car nous ne chantons
que pour vous, ô pâtres et habitants des *mas*.

(Panel B) French

relegated all other dialects and languages to the status of spoken only.[4] Nevertheless, the main target of the ordinance was the use of Latin, widespread among elites and in official documents. Local languages were virtually unaffected: "They naturally used the King's French in their administrative documents. However, it took a further three centuries for even the 'best families' of the town and the surrounding countryside to adopt the King's French in everyday speech" (see Lodge, 1993, p. 125).[5]

The use of the French language was still an exception for the rural population at the time of the French Revolution. More than 80 percent of the population was rural and French was only spoken by the elites, in the cities and in the region around Paris. In his *Report on the necessity and means to annihilate dialects and to universalize the use of the French language*, Grégoire (1794) argues that "we are still, for language, the Tower of Babel"; he estimates that only three million French persons, out of twenty-nine million, could speak the French language fluently. The different revolutionary governments considered the use of dialects to be a threat to the Revolution and to national unity, and they undertook efforts at school building, homogenization, and forced adoption of the *langue nationale* (Perrot,

---

[4]There were some very late and unsuccessful attempts to codify the written language—for example, the development of the Mistralian norm with *Lou Felibridge* of Frédéric Mistral, the 1904 Nobel Prize in Literature.

[5]Lodge (1993) argues that the ordinance merely continued previous policies. Also, see Binzel, Link and Ramachandran (2021) on the impact of the Protestant Reformation on vernacularization.

1997). In their reports on public instruction, Talleyrand (1791) and Condorcet (1792) proposed universal education and a system of public primary schools in order to spur the adoption of French.[6] Yet, none of the projects was carried out to its term because of the high political instability. On July 20, 1794, the decree of 2 Thermidor, Year II, extended the provisions of the Ordinance of Villers-Cotterêts to local dialects, but Robespierre was arrested and executed by guillotine a few days later, marking the end of the Reign of Terror and the suspension of the decree.

Today French is the only official language and is spoken by nearly all French nationals, while other languages and dialects are in danger of becoming extinct; according to UN-ESCO, twenty-one languages are endangered in France (Moseley, 2010).[7] Weber (1976) and Lodge (1993) argue that the French language was not widely adopted until the early twentieth century, the time of universal state-sponsored education and the development of railways (Weber, 1976). In particular, they argue that the schooling laws enacted by Guizot (1833) and Ferry (1881–82) played a major role. As noted, Guizot mandated that schools be built and that the curriculum and training of teachers be standardized (more details in Section 4), while Jules Ferry extended this and made education free, universal, and secular. Last but not least, the *plan Freycinet* (1878) mandated the construction of railways, canals, and ports everywhere in France.

2.2 Related Literature

Historians and social scientists have argued about the importance of schools for the homogenization of populations during nation-building (e.g. Alesina, Giuliano and Reich, 2021; Anderson, 2006; Hobsbawm, 1990; Weber, 1976). Governments typically invest in infrastructure (e.g., roads and railroads) and implement school-building and compulsory-education laws in order to generate a "commonality of interests, goals, and preferences" (Alesina, Giuliano and Reich, 2021, p. 2). These nation-building policies promoted homogenization in many countries historically. However, because they are often implemented together and most people are treated, evaluating the role of each policy is hard. As a result, while an important literature focuses on government incentives to homogenize the population (e.g. Aghion et al., 2019; Alesina, Reich and Riboni, 2020; Bandiera et al., 2019; Cáceres-Delpiano et al., 2021), the effects of particular nation-building policies have been less explored (e.g. Assouad, 2021; Bazzi et al., 2019). Our paper not only gives the first causal and quantitative evidence on the role of education in homogenization but also extends the discussion about the drivers of homogenization and their long-run consequences.

The second literature our paper relates to treats language as a proxy for cultural, ethnic,

---

[6]de Talleyrand-Périgord (1791) writes that "the language of the Constitution and of the laws will be taught to all; and this crowd of corrupted dialects, the last remnant of feudalism, will be forced to disappear."

[7]These twenty-one idioms are considered languages by UNESCO but are listed as dialects in Billy, Nadiras and Moufflet (n.d.) (Appendix Figure A3). Some of the dialects in Billy, Nadiras and Moufflet (n.d.) are not considered languages or are not considered endangered by Moseley (2010).

or national identity.[8] While prior research uses discrete measurements, we can evaluate the gradual nation-building process across space and time with a continuous measure of linguistic proximity at a point in time.[9] In this regard, Falck et al. (2012) is closely related to our paper since the authors exploit local dialect variations using the linguistic atlas of the German Empire (*Sprachatlas des Deutschen Reichs*). However, our dataset is more detailed since we observe spoken language and therefore capture continuous differences in not only language but also pronunciation. Our data work relates most to the Salzburg dialectometry team's, which is of particular relevance. We use some of their data to validate our measure of distance and for completeness and robustness (more details in the following section).

Third, following the seminal paper by Akerlof and Kranton (2000) on the economics of identity, the determinants of identity formation have been gradually identified in empirical research. Most previous studies show identity backlashes (Bazzi, Hilmy and Marx 2021; Carvalho and Koyama 2016; Fouka 2020) and identity formation (Clots-Figueras and Masella 2013). In the context of France, Dehdari and Gehring (forthcoming) study the relationships among oppression, annexation, and regional identity. Our focus on the institutional setting enables us to identify the causal effect of state interventions on identity. In this regard, our paper supports the constructivist view, according to which identity is socially determined (Gellner, 1983a; Smith et al., 1991), and it is related to Blouin and Mukand (2019), which shows that government can manipulate the salience of ethnicity.

Finally, an important literature studies the effects of education, in particular works evaluating school programs and labor market outcomes (Duflo, 2001, 2004). In the context of the *Loi Guizot* of 1833, two studies find it had positive impacts on fiscal capacity and development (Montalbo, 2020) and literacy in a single French village (Blanc and Wacziarg, 2020). Moreover, prior studies show that schools affect individual values (e.g., beliefs and identity) via the content of curricula (e.g. Algan, Cahuc and Shleifer, 2013; Cantoni et al., 2017; Clots-Figueras and Masella, 2013) in the short and medium runs.

3 THE DATA

3.1 Linguistic Distance

Atlas Linguistique de la France. We gather data on the language spoken around 1900 in 577 municipalities from the *Atlas Linguistique de la France*.[10] The *Atlas* was

---

[8]Prior research about ethnic identity and ethnic languages has studied the roots of ethnic and ethnolinguistic diversity (for example, Ahlerup and Olsson, 2012; Cervellati, Chiovelli and Esposito, 2019; Green, 2013; Michalopoulos, 2012) and linguistic characteristics (for example, Blasi et al., 2019; Blouin and Dyer, 2021; Dickens, 2021; Galor, Özak and Sarid, 2018). Covering all the papers in this literature is beyond our scope. See the recent review by Ginsburgh and Weber (2020).

[9]Bazzi et al. (2019) use language at home; Depetris-Chauvin, Durante and Campante (2020) use a categorical variable with five scales to measures the relative salience of ethnic identity and national identity; Assouad (2021) uses the adoption of pure-Turkish first names.

[10]The *Atlas* also contains information on sixty-two additional towns in neighboring countries, including Italy, Switzerland, and Germany. We did not digitize the data for these towns.

published in nine volumes from 1902 to 1910 (Gilliéron and Edmont, 1902-1910) and relies on a survey carried out by linguist Jules Gilliéron from 1897 to 1901 (henceforth "1900") to study Romance languages in rural parts of the country, where the majority of the population lived and did not speak French.

For four years, Gilliéron's assistant Edmond Edmont traveled across the country and asked locals for the standard, common pronunciation of 1,920 words or expressions in their municipality. The survey captures spoken language, including vocabulary and pronunciation, and paints a particularly detailed picture of the languages and dialects commonly spoken by ordinary people in their daily life at the time (Le Dû, Le Berre and Brun-Trigaud, 2005). Importantly, the Atlas was conducted to capture the common language spoken in the surveyed municipality and not the language spoken by the individuals surveyed.[11]

Figure 2: Map of *ciel* ("sky")

(Panel A) France **(Panel B)** Zoomed in
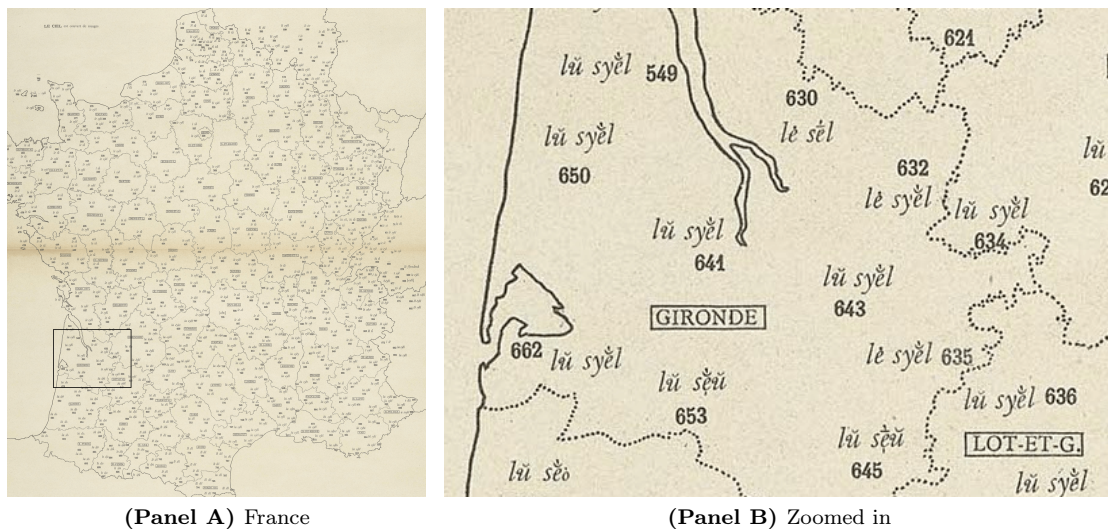
In order to assess selection into the sample, Appendix Table A1 Panel A displays summary statistics for observable variables of interest both in the sample of full municipalities in France and in the *Atlas* only. We do not find any particularly relevant differences— except from the fact that municipalities in the *Atlas* are on average less populated, but the difference is not statistically significant. In fact, Gilliéron and Edmont (1902) argue that they "never searched for the places that could have stood out in any way" (p. 4). Most of the municipalities in the *Atlas* were selected before the start of the survey with the sole

---

[11]In order to best do this, Edmond Edmont mostly asked bilingual informants (Gilliéron and Edmont, 1902; Le Dû, Le Berre and Brun-Trigaud, 2005).

purpose of covering places "at approximately equal distances from each other" (Gilliéron and Edmont, 1902, p. 4). Appendix Figure A5 displays the spatial distribution of the municipalities surveyed in the *Atlas*, and Appendix Figure A6 shows that the distance to the closest town in the *Atlas* was twenty-five kilometers on average, with very little variance.

The inclusion of a large number of words in the survey allows for a direct, word-by-word comparison between languages. While the survey of Georg Wenker in 1880s Germany biased its results by suggesting a grammatical structure to the subjects, the *Atlas* did not. The words were mostly chosen before the survey was carried out; and the large size of the corpus allows for substantial heterogeneity in their characteristics. Gilliéron carefully thought about how to not bias the survey and mostly included standard, common words.[12] Finally, only the spontaneous replies of the persons surveyed were recorded in order to further limit potential biases (Baiwir, 2019; Gilliéron and Edmont, 1902).

In order to capture the complexity of the languages surveyed, the *Atlas* relied on the *alphabet Rousselot-Gilliéron*, a system of phonetic transcription developed at the time. The pronunciations of the words are displayed in maps—one map for each word or expression.[13] Appendix A1.1 details how we digitized the phonetical representations in the *Atlas*. We display the pronunciation of *ciel* ("sky," the 285th map in the *Atlas*) in Figure 2. Panel A maps the pronunciation of the word across space in France. In Panel B, which covers an area of roughly 100x80 miles, we zoom in to the *département* of Gironde, in the historical region of Aquitaine in the southwest of France and at the frontier between the historical *langues d'oc* and *langues d'oïl* regions. The map paints a particularly detailed picture of the common language spoken at the time. We observe substantial heterogeneity, with both persistent historical differences and recent variations resulting from the diffusion of the French language.[14] Some municipalities had adopted the French word *ciel*, while others were still using dialects of *langues d'oc*, with *cel* or *ceu*.

Finally, we note that, while some of the variation is only phonological, it also arises from different structures, vocabularies, or languages. Here, the change from *ceu* to *ciel* involves both vowel breaking, or diphthongization, from "e" to "ie" and a change of the final letter of the word. Differences can be more profound and structural, and for that reason we under-estimate the variation coming purely from language and the effect of schools on homogenization in the historical *langues d'oc* region.

Measure of linguistic distance. We rely on the *Atlas Linguistique de la France* to measure linguistic distance across pairs of municipalities. We use a Levenshtein-distance algorithm, defined as the minimal number of edits by insertions, deletions, or substitutions between phonetic representation of words, to capture linguistic dissimilarity in the *Atlas*.

---

[12]Also included but more marginal were ancient words, words of recent origin, and words used only in some dialects.

[13]The scanned maps are publicly available on the website **cartodialect.imag.fr**.

[14]Appendix Figure A7 displays a similar map of *étoile* ("star"). The words used instead of *étoile* were mostly *estello* and *estella*, from Provençal—a dialect of Occitan.

Our measure of distance is then normalized to the unit interval by dividing by the greater number of letters between the two words. We provide more details on our measure in Appendix A1.2. Levenshtein distance was first introduced in dialectology by Kessler (1995) to study Irish dialects, and is used by Adsera and Pytlikova (2015); Dickens (2018, 2021) in more recent research by economists.

Table 1: Digitized maps (words) from the Swadesh list in the *Atlas*

Maps of the *Atlas Linguistique de la France*									
43	animal	142	drink	796	hand	923	new	1241	sun
210	ashes	946	ear	314	how	929	night	894	swim
1	bee	1299	earth	638	knee	908	nose	1120	tail
916	black	558	fire	1200	know	76	others	233	to sing
1187	blood	1052	fish	559	leaf	1039	rain	344	to spit
250	cat	582	flowers	120	lots	1159	river	750	tongue
461	child	178	fog	788	moon	285	sky	52	trees
612	cold	615	fruit	874	mountain	418	sleep	397	two
882	die	1031	full	151	mouth	494	star	432	water
279	dog	270	hair	328	neck	1014	stone	594	wood

In order to measure distance, we digitized the maps showing the pronunciation of fifty representative words in the *Atlas*, which we chose from a list of words that exist in most languages and cultures and correspond to a "standardized universal list of meanings" (Ginsburgh and Weber, 2020, p. 367). We relied on the list of words compiled by Swadesh (1952) and digitized words such as "cat," "dog," "drink," "fruit," "hand," "mouth," "night," "rain," "sun," "trees," and "water." Table 1 lists the words we digitized from the *Atlas* to calculate Levenshtein distance. We take the average Levenshtein distance over the fifty words to measure linguistic distance from French.[15] Since the *Atlas* does not provide any information on the standard French language, we use the language spoken in the town closest to Paris in the *Atlas* (Le Plessis-Robinson) as the reference point for French. We refer to our main measure as the linguistic distance from French in 1900.[16]

Figure 3 Panel B displays the spatial distribution of our measure of linguistic distance from French in 1900. The darker areas indicate a larger linguistic distance from the standard language. Panel A shows that our measure tracks closely the historical language regions, suggesting that it captures recent changes but also deep-rooted historical differences.[17]

[15]We also measure linguistic distance across the total of 166, 176 pairs of municipalities, but we focus in the paper on the linguistic distance from French.

[16]The French language is believed to originate, in part, from the *francien* dialect spoken in the region near Paris (Lodge, 1993, 2002). We also turn to alternative towns of reference for robustness and to attempt to locate the regions of origin of the French language.

[17]We plot the distribution of linguistic distance from French in Appendix Figure A9. The only other measure of linguistic distance in France not using the *Atlas* is cladistic distance, a measure of the theoretical, structural distance

Figure 3: Historical language regions and linguistic distance from French in 1900

(Panel A) Historical language regions

(Panel B) Linguistic distance from French (1900)

In the bulk of the analysis we use our measure of linguistic distance from French—across the representative words from the Swadesh list whose maps we digitized. We also show the robustness of our results to using alternative measures of linguistic distance and to relying on the entire corpus of words from the *Atlas*. The Salzburg dialectometry team has been studying and digitizing the *Atlas* for more than two decades (see Goebl, 2002, 2003, 2006, 2011) and made their measure of distance across the entire corpus of words in the *Atlas* available in late 2019 (Goebl et al., 2019), after we started digitizing the *Atlas* for this project.[18] Their distance metrics, Relative Identity Value (RIW) and Weighted Identity Value (GIW) (Goebl, 1984; Séguy, 1973), are closely related to ours. Appendix Figure A18 plots our measure of average Levenshtein distance across 50 maps from the Swadesh list against their measures across 1,681 maps—with a correlation of more than .96. While the Levenshtein algorithm is purely automated, the RIW and GIW measures initially rely on a qualitative analysis by linguists before the data is fed into an algorithm computing distance:

between families of languages in large delineated areas (Desmet, Ortuño-Ortín and Wacziarg, 2012; Desmet, Ortuño-Ortín and Weber, 2009; Fearon, 2003; Fearon and Laitin, 1999; Laitin, 2000). In France, Spolaore and Wacziarg (forthcoming) compute cladistic distance from the Oïl family of language using the classification and trees in the *Ethnologue* (Eberhard, Simons and Fennig, 2021). Appendix Figure A8 displays their measure for the regions where Occitan, Francoprovencal, Basque, Breton, and Flemish were spoken historically. However, this approach suffers from limitations: it gives no sense of what the timing is since cladistic distance is only a geographical attribute of the regions in which these languages were spoken historically; we don't observe variations between dialects or within dialects; finally, the spatial distribution of speakers and their usage of dialects or languages is unknown.

[18]We discovered their website in December 2021. Their measure of distance is only available across the entire corpus of words/maps, while the raw data for any given map is not available.

"the original data from a given linguistic atlas are analyzed by trained geolinguists according to traditional linguistic criteria (phonetics, morphology, vocabulary, etc.). This results in a great number of disjunctive (qualitative) units called taxates" (Goebl, 2011, p. 437). Then, the distance measure is computed using "the percentage of pairwise (qualitative) matchings of the elements (types, taxates) of two measuring point profiles"—with the GIW measure giving a greater weight than RIW to rare features.

3.2 National Identity and Other Outcomes

We rely on a set of different municipality-level measures to capture the salience of national identity, including data on participation in the French Resistance during World War II and on the vote in the 1969 constitutional referendum on regionalization. Additionally, we use data on migration and trade flows across pairs of districts and *départements* to study the effect of linguistic distance on these outcomes. We detail the process of data collection of these measures in Section 6, in particular regarding the 1969 referendum, whose outcome is not available at the municipality level.

3.3 Additional Details on the Sample

Finally, we note that, throughout the paper, we focus on the parts of mainland France that were not annexed at any point in the nineteenth century, hence excluding Corsica, Savoy, Nice, and Alsace-Lorraine. We further restrict the sample to rural municipalities with recorded population that have not split or been merged since the French Revolution.[19] Including such municipalities would bias our regression discontinuity results to zero because historical population is the running variable. Last, we focus on rural towns because of the better quality and representativeness of the *Atlas* there and because it allows us to bracket the fact that the language spoken and the schools in urban settings were likely very different from those in rural places, both before and after the Guizot Law.[20]

4 EMPIRICAL STRATEGY

4.1 The *Loi Guizot* of June 28, 1833

In many ways, primary education was in an appalling state before Guizot took over as Minister of Public Instruction in 1832.[21] The vast majority of the population was still

---

[19]Out of the 577 municipalities in France, 552 are within the parts of mainland France that were not annexed at any point in the nineteenth century; 510 had a non-missing population at the time of the law; 479 were rural, with less than five thousand inhabitants; and 375 have not been merged or split since the French Revolution.

[20]In practice, including urban municipalities in the sample does not change our results significantly since we compare municipalities near the threshold. It only (marginally) affects the selection of the optimal bandwidth.

[21]For more details, see Day (1983); Furet and Ozouf (1977); Meyers (1976); Montalbo (2021); Weber (1976) and see the 1833 Guizot survey, available at **www.inrp.fr/she/guizot**. Blanc and Wacziarg (2020) also provide a

illiterate. Most municipalities did not have schools. When they had one, the classrooms were only open part of the year, sometimes only for a couple of months, with irregular schedules and without trained teachers. The curriculum placed heavy emphasis on religious and moral education (Meyers, 1976), and Weber (1976) notes that when a teacher was not a clergy member (or a 'lay cleric'), they "could well have been a retired soldier, a rural constable, the local barber, innkeeper, or grocer, or simply a half-educated peasant's son" (p. 305).

Figure 4: Literacy and schools

Note: This figure displays the literacy rate (left) and number of primary schools per thousand inhabitants (right) over time in France. Source: Statistique Générale de la France (n.d.a).
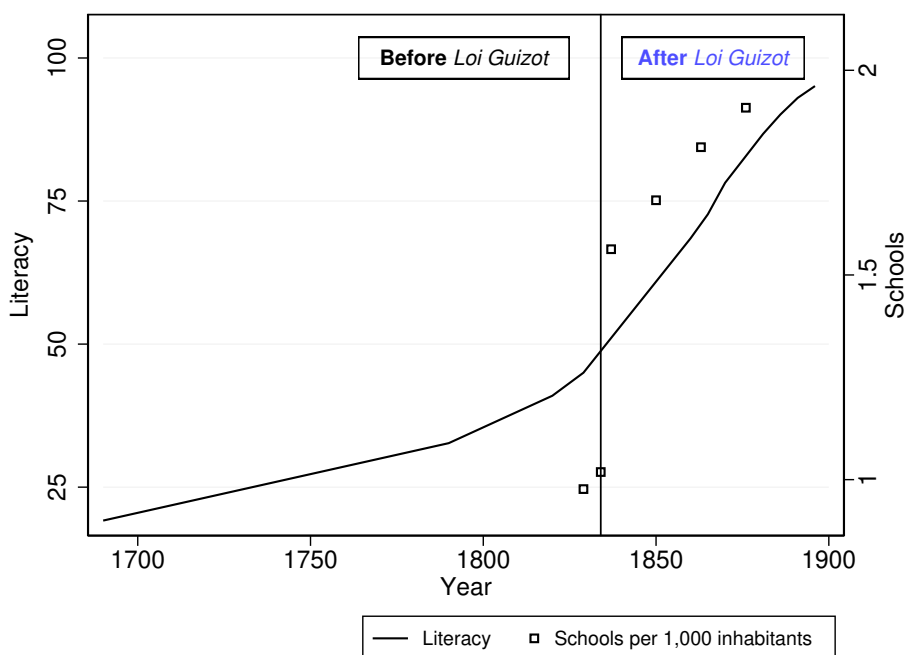
The *Loi Guizot du 28 juin 1833* laid the foundation for education for the masses in France by requiring municipalities with more than five hundred inhabitants to build a primary school for boys.[22] We exploit the five-hundred-inhabitant cutoff in a regression discontinuity design to study the effects of school openings on the homogenization of language and the formation of a national identity. The number of primary schools increased nearly twofold in a decade, from one school for every thousand inhabitants to one for every six hundred,

detailed analysis of the changes brought about by the Guizot Law.

[22]The Jules Ferry Laws of 1881–82 established free, universal, and secular education. The *Loi Guizot* prefigured the *lois Ferry* by mandating the building of schools and establishing very similar principles in practice, albeit not for the entire population. Furet and Ozouf (1977) (p. 136) provide an extensive description of the law and its subsequent decrees, including on the five-hundred-inhabitant threshold.

and universal literacy was achieved in the aftermath of the law (Figure 4).[23] In 1829, only 1.3 million children attended primary schools in France; by 1872, this number had risen to more than 4.6 million (Appendix Figure A10).

Yet Guizot radically modernized primary education well beyond the building of schools; his reforms brought about the advent of the state's control of the provision of education in France (Furet and Ozouf, 1977; Nique, 1990; Prost, 1968).[24]

The state mandated the construction of a national system of teachers schools; created a centralized body of school inspectors; required the teaching of French and required that teaching be in French; certified the textbooks authorized for use in the schools; and mandated the creation and distribution of these textbooks. The annual budget of the Ministry of Public Instruction for primary education increased sixtyfold in less than a decade, from one hundred thousand Francs to six million Francs.[25] The building of primary schools had to be financed by the municipalities, but the *département* or the state had to provide financial assistance to municipalities that did not have the resources or refused to raise additional taxes (*centimes*). Public schools came to account for more than 90 percent of schools on the eve of the *lois Ferry* of 1881–82 (Appendix Figure A11), and students who could not afford tuition had to be fully funded (Appendix Figure A12). We provide a detailed summary of the Guizot Law in the remainder of the subsection.

Standardization of curricula. The spread of the French language and the formation of a national identity through the standardization of the curriculum were some of the main purposes of the law. As noted, François Guizot (1833) himself wrote that "through the teaching of the French language, primary schools will instill and spread the spirit and unity of nationality everywhere" (p. 102). Paul Lorrain, a close advisor of Guizot, argued that "each school shall be a colony of French language in a conquered land" (Lorain, 1837, p. 29).[26] In order to undertake standardization, the state became involved in all stages of the production of education; yet its involvement in the curriculum was unmatched.

The public primary schools built after the law were required to teach "elements of French language" and French history and geography (Article 1); and teachers were to teach in French. In 1863, local dialects were the language of instruction in only 92 out of 65,338 public primary schools in France (Weber, 1976, p. 498). Meyers (1976) documents this in his study of rural teachers in nineteenth-century France: "The language they spoke and taught was French, not patois; the history they taught, even the holidays they celebrated,

---

[23]The bulk of the increase in the number of schools built took place from 1835 to 1837, when the policy took effect. Only a limited number of schools were built in 1834. After that, the increase can be accounted for by the *loi Falloux* and *loi Duruy*, which extended the provisions of the Guizot Law to schools for girls at the eight-hundred-inhabitant threshold and five-hundred-inhabitant threshold, respectively.

[24]Rosanvallon (1985) provides an interesting discussion of the motivations and political philosophy of Guizot.

[25]See Reboul (1991) (p. 170). The budget for secondary education remained roughly constant over the same period.

[26]Lorain (1837) also argues that "it is time to truly enforce this reunion through the uniformity of language and customs, and ⋯ to merge them all into national unity" (p. 29).

were national, not regional" (p. 552). Interestingly, the expression "our ancestors the Gauls", coined by Ernest Lavisse in 1878 and depicting a fabricated national ideal, can be traced out to the early nineteenth century; and a similar form is used by Guizot in his book *L'histoire de France* (Bourdon, 2017).

The creation of five main textbooks to be used in all public primary schools was mandated by Guizot (Choppin, 1986). Then the Royal Council for Public Instruction had the authority to determine which textbooks were authorized in the classrooms of public schools, and only authorized books could be used (Article 9, *Statut portant règlement des écoles primaires élémentaires* of April 25, 1834).[27] The textbooks were free for poor students. In three years, the Ministry of Public Instructions sent millions of these textbooks to schools throughout France, including a million copies of a book on the French language (*Alphabet and Reading Primer*). At the same time, Guizot created the Society of the History of France (1833) and the Committee for Historic and Scientific Works (1834) to encourage the creation of a national narrative in history textbooks (Rosanvallon, 1985). Last, he created the *Manuel général*, a pedagogical review first published in 1832–33, to provide direct instructions and up-to-date information on recommended teaching methods.[28]

Finally, the *Loi Guizot* imposed the teaching of French and standardization through state coercion. Although it was not officially mandated, schoolteachers often relied on corporal punishments and on the use of the *symbole*, a token of shame, when students were caught speaking a language other than French (Weber, 1976, p. 313). The quasi-systematic use of coercion was especially strong in Brittany and Occitanie (Calvet, 1974; Polard, 2004). Weber (1976) argues that "Breton was hunted out of the schools" (p. 313). Today, Occitan speakers use the word *vergonha* ("shame") to describe the policies of the French government intended to shame them in the nineteenth century.[29]

Training of schoolteachers. The state took steps to dramatically improve the training of schoolteachers to help standardize the curriculum in primary schools. François Guizot (1860) wrote in his memoirs that he "was trying to penetrate the soul of every schoolteacher" (p. 75).

The creation of a national system of normal schools (or teachers colleges) led to radical improvements in the training of primary school teachers. All *départements* were required to open a normal school in the departmental capital, and schoolteachers had to graduate with a *brevet de capacité* (certification) to teach in public schools. There were only three

---

[27]The Royal Council would decide which books were authorized for use in public schools only, not private or religious schools (in towns below the cutoff). See *Circulaire aux recteurs ayant pour objet de leur transmettre la liste des livres dont l'usage est autorisé dans les établissements d'instruction primaire* of June 21, 1837.

[28]For example, we find the following quote in the first edition of the *Manuel général*: "The study of the French language is paramount where provincial dialects and foreign languages have been preserved" (*Manuel général, ou Journal de l'instruction primaire, destiné à guider les instituteurs dans le choix des méthodes et à répandre dans toutes les communes de France les meilleurs principes d'éducation, publié sous la direction d'un inspecteur général des études et de plusieurs autres membres de l'Université* Novembre-Décembre 1832, Paris: Hachette, p. 12).

[29]See **Wikipedia**.

normal schools in France in 1828; by 1834, that number had risen to seventy-two (Code de l'instruction primaire, 1834, p. 42). As a result, graduates of normal schools quickly replaced older teachers. In 1846, seventeen thousand of the forty thousand primary school teachers in France had attended a normal school (Day, 1983, p. 29).

Schoolteachers became de facto civil servants with the creation of the *engagement décennal*, which required graduates of normal schools to remain on the job as schoolteachers for at least ten years (Day, 1983).[30] Their salary improved significantly with the creation of a minimum annual salary that increased sixfold by 1880.[31] According to Meyers (1976), on the eve of the *lois Ferry*, teachers "were earning considerably more even than many artisanal occupations" (p. 553). They were under the direct supervision of both the local municipal council and departmental school inspectors, who were themselves civil servants. The strong influence and control of the state therefore left schoolteachers with a very thin margin for acting outside of what was recommended by the state.

Finally, a body of school inspectors was created by the Royal Ordinances of February 26, 1835, and November 13, 1837 (Ravier, 2012).[32] School inspectors were recruited nationally and were under the aegis of the Ministry of Public Instruction. Schools had to be reviewed yearly, and inspectors attended classes to review the quality of the teaching and the teachers' conduct. In particular, they had to check that only authorized books were being used (per the *Arrêté du 27 février 1835*). Ravier (1998) argues that inspectors played an important and direct role in the imposition of the French language in the north of France (p. 23).

4.2 Regression Discontinuity Design

In order to evaluate the causal effect of state-sponsored education on linguistic distance from French, we rely on a regression discontinuity (RD) design exploiting quasi-experimental variation in school building. Our empirical specification takes the following form:

$$(1) \qquad y_i = \alpha + \tau \times school_i + f(population_i) + \varepsilon_i$$

where y_i is either the linguistic distance from French in 1900 in town i ($y_{i,1900}$) or some other variable capturing national identity; $school_i$ is a deterministic and discontinuous function of population that equals 1 if the population of town i ($population_i$) was above five hundred at the time of the law and 0 otherwise; $f(population_i)$ is a polynomial controlling for smooth functions of population in town i and allowing for different slopes on different sides of the cutoff to account for the conditional expectation of the outcome.[33] Throughout the paper,

---

[30]Only graduates on partial or full scholarships were required to do so, but they constituted the majority of graduates of the normal schools.

[31]While it was on par with the salary of agricultural workers at the time of the law; it then rose significantly relative to other occupations (Morrisson and Snyder, 2000).

[32]First departmental inspectors, then inspectors for individual schools.

[33]We follow Montalbo (2020) and use population in 1836, when the law took effect and most of the schools were built (see Figure 4 and the May 8, 1834, speech of François Guizot at the Chambre des députés (Guizot, 1863, p.

we use "schools," "school building," or "state-sponsored education" interchangeably to refer to the treatment.

Under the assumption of continuity of conditional expectation functions of potential outcomes, τ identifies the causal effect of the treatment on linguistic distance from French in 1900. Cattaneo, Idrobo and Titiunik (2020); Imbens and Lemieux (2008); Lee and Lemieux (2010) provide reviews of RD designs, and we provide below a detailed discussion applied to our framework. In particular, we show that using regression discontinuity allows us to estimate the effect of schools on the change in language and a lower bound on the speed of homogenization over time, from one cross-sectional regression.[34]

Schools and homogenization. Under the main RD assumption, we estimate the effect of schools on the homogenization of language over time. Because linguistic distance before the policy was not affected by the policy, we can account for unobserved deep-rooted historical differences at the threshold and the effect of schools on linguistic distance from French in 1900 ($y_{i,1900}$) is equal to their effect on the change in linguistic distance over time ($y_{i,1900} - y_{i,1833}$).[35] This is particularly relevant in our context since nation-building aims at homogenizing populations.

Then, we demonstrate in Proposition 1 that the percentage change relative to the counterfactual is a lower bound on the speed of homogenization over time at the threshold—although we only observe linguistic distance in a single cross-section.[36] The intuition is straightforward. If homogenization took place at the threshold and state-sponsored education contributed to this process, then the total decrease in linguistic distance over time will be larger than the decrease relative to the counterfactual in the cross-section—which would only be a portion of the total decrease. Equivalently, we assume that, absent the policy, homogenization would still have taken place at the threshold. This allows us to estimate how fast homogenization took place, a parameter of interest.

Proposition 1. *If the change over time is of the same sign as the effect of treatment in the cross-section, then the percentage change relative to the counterfactual, $|\tau|/\alpha$, is a lower bound on the speed of homogenization over time, $|\boldsymbol{g}|$: $0 \leq |\tau|/\alpha \leq |\boldsymbol{g}|$.*

Proof. See Appendix A2.2.

---

244)). Anecdotal and historical evidence also suggests that the application of the law was based on population at the time. In Appendix Table A11, we show that there was no jump in linguistic distance at the five-hundred-inhabitant threshold using the population in other years, suggesting that treatment status was assigned at the time of the law.

[34]Other papers extend the classical RD framework to incorporate variation over time. In particular, Lemieux and Milligan (2008) observe both the outcome and the treatment in intermediate years and propose a first-difference RD estimator to account for the fact that the treatment changes over time; Cellini, Ferreira and Rothstein (2010) proposes a dynamic RD design that tracks the treatment effect over time, with a dynamic treatment assignment; Grembi, Nannicini and Troiano (2016) implement a difference-in-discontinuities design in the context of treatments for the same threshold at different times.

[35]See Appendix A2.1. Alternatively, we can account for linguistic distance before the policy with historical-language or historical-dialect fixed effects, which will account for regional-level variation in the historical use of dialects.

[36]The speed of homogenization is defined as the rate of decrease in linguistic distance from the French language from 1833 to 1900.

Additionally, it is easy to see that if there was no secular change from 1833 to 1900 other than that caused by the Guizot Act, $|\widehat{\tau}|/\widehat{\alpha}$ would be an unbiased estimator of the speed of homogenization.[37] In that case, the estimator could also be interpreted as the effect of state-sponsored education on the rate of homogenization.[38] We interpret this as a suggestive benchmark, but the estimation of $|\tau|/\alpha$ is highly informative about the speed of homogenization over time regardless since it is a lower bound.

Timing of the survey. The timing of the linguistic survey is ideal for studying the effect of state-sponsored education during Guizot and to abstract from potential confounding factors resulting from the Freycinet Plan and Ferry Laws—the two other important drivers of homogenization at the time (Weber, 1976). Most adults in 1900, at the time of the *Atlas*, had not been affected by the Jules Ferry Laws of 1881 and 1882, which made education free, universal, and secular, or by the Freycinet Plan of 1878, which mandated the construction of railways, canals, and ports and was only completed in 1914.[39]

Nature of the treatment. The policy we evaluate is a bundled treatment capturing the building of schools (brick and mortar), the provision of education, the creation of a standardized curriculum, and the training of schoolteachers. Although we cannot rigorously compare the elements of the policy, we argue that the creation of a standardized curriculum and the training of schoolteachers were likely the most important dimensions. Appendix Figure A13 displays the share of municipalities with education (with or without a school building) before the *Loi Guizot*, plotted against population. The data is available in less than a quarter of *départements* in France, but it paints a detailed picture of education before the law. At the five-hundred-inhabitant threshold, one-third of municipalities had a schoolhouse and more than half provided some education. This allows us to grasp the nature of the counterfactual; and since a large share of towns already provided education before the law was passed, it suggests that the provision of education or of school buildings is unlikely to drive our result. Meanwhile, the creation of a standardized curriculum and the training of schoolteachers played important roles, as discussed in Section 4.1. Finally, we estimate a local, intention-to-treat effect. There is no data available at the municipality level on schools built after the law. The data is only available at the *département* level.[40]

---

[37]Although it is a strong assumption, historical and anecdotal evidence suggest it may hold in our context since culture is highly persistent and there was no other nation-building policy until much later in the nineteenth century (see below and Weber, 1976).

[38]When there is no homogenization below the threshold, the rate of homogenization at the threshold is also the effect of the treatment on the rate of homogenization.

[39]Children six to thirteen years old attended primary school.

[40]Appendix Table A2 estimates the *département*-level association between exposure to the Guizot Law, defined as the share of municipalities with more than five hundred inhabitants, and the growth in the number of schools per capita in the aftermath of the law—after controlling for the number of schools before, in 1829. The results suggest a strong impact of the law on school building, in line with the time series in Figure 4. In order to account for endogeneity, we instrument the overall share of municipalities with more than five hundred inhabitants with the share—with more than five hundred inhabitants—in the narrower samples of municipalities within one hundred, two

Estimator, polynomial degree, and bandwidth. Throughout the paper we report conventional RD estimates, following Hahn, Todd and Van der Klaauw (2001), and we use the bias-corrected estimator of Calonico, Cattaneo and Titiunik (2014) in robustness analysis. Second, we rely on three different ways of specifying the polynomial fit—mean comparison (order 0), local linear regression (order 1), and quadratic polynomial (order 2)—and we avoid higher-order polynomials in order to limit bias and overfitting (Gelman and Imbens, 2019). Last, we follow Calonico, Cattaneo and Titiunik (2014) and Imbens and Kalyanaraman (2012) in using mean squared error (MSE) optimal bandwidths and a range of other bandwidths for robustness.

Identification and balance at the threshold. In order to assess the validity of our design, we test the main assumption of continuity of the conditional expectation function of potential outcomes. We show that observable geographic, climatic, institutional, cultural, and economic factors vary smoothly at the threshold, using data on altitude, distance from Paris, distance from coastlines, temperature, precipitation, subscriptions to the *Encyclopedie* in 1776–79, and distance from railways in 1870. Figure 5 plots linguistic distance, predicted linearly using these covariates of interest, against population—allowing us to aggregate observables into a single conditional expectation function. We do not find any discontinuity at the threshold.[41]

Finally, additional threats to identification may come from sorting and migration. Appendix Figure A15 plots the density of the population of municipalities in France and in the *Atlas*. We do not find any jump in the density at the threshold and, using the Cattaneo, Jansson and Ma (2020) density test, we fail to reject the null that there is no discontinuity and conclude that manipulation is unlikely—in both samples.[42] Migration itself is only

---

hundred, three hundred, and four hundred inhabitants of the five-hundred-inhabitant threshold for school building. We report first-stage, reduced-form, and two-stage-least-squares results, and note that the instrument is weaker when using smaller bandwidths. We find a large effect of exposure to the law on the building of schools across specifications.

[41] Appendix Table A1 Panel B compares each of these variables below and above the threshold for school building. We restrict the sample to municipalities that fall within different bands around the 500-inhabitant cutoff in order to evaluate the balance in observables for different bandwidths. We report coefficients on the difference in means above and below the cutoff for each variable of interest. In Panel A, we rely on the sample of all municipalities in France in order to compare a large number of places near the cutoff—for example, the 601 municipalities with between 490 and 510 inhabitants. In Panel B, we solely rely on the sample of municipalities surveyed in the *Atlas*. We do not report the results for the ten-inhabitant bandwidth in Panel B because the sample size would be too small. We obtain similar results. In particular, we find no jump at the threshold in distance from railways, which is one the main competing explanations for homogenization (see Weber, 1976). Finally, we note that municipalities near Paris and in the historical *langues d'oïl* region are less populated near the threshold, both in the full sample and in the *Atlas* only. We also find this in Appendix Figure A14, which displays the spatial distribution of towns in the *Atlas* that are near (above or below), or away from the threshold.

[42] If anything, positive sorting is particularly unlikely to affect our results since, according to François Guizot, most municipalities did not want to build the schools mandated by the law—instead, negative sorting would bias our results toward zero. In his May 8, 1834, speech at the Chambre des députés, Guizot (1863) noted that "there are 21,000 municipalities in France that do not feel the need for primary instruction or that do not dare to do what it takes to satisfy this need" (p. 244). These municipalities had to be compelled or financially helped by the state. Moreover, there was very little room for towns to manipulate census counts at the time of the law (Biraben, 1963, p. 309). They had to keep a nominal list of all their inhabitants updated yearly, and this list had to be sent to the prefect of each department to be checked by the central administration. The requirement to send the list out for verification only existed after the passage of the law and was abolished in 1837.

Figure 5: Balance in observables at the threshold: CEF of predicted linguistic distance using a set of geographic, climatic, cultural, institutional, and economic covariates

Note: This figure plots predicted linguistic distance from French against population around the discontinuity introduced by the school-building policy. Linguistic distance from French is first regressed on the set of geographic, climatic, cultural, institutional, and economic covariates used in Appendix Table A1. We use the predicted values from this regression. Each point plots the average value within a bin, partialled out of the full set of controls used in Appendix Table A4. Linguistic distance from French is defined as the Levenshtein distance from Le Plessis-Robinson. We apply a local-polynomial fit of order 2 and a mean-squared-error optimal bandwidth for local-polynomial estimation. We rely on an evenly spaced mimicking-variance optimal number of bins. Observations are at the municipality level.
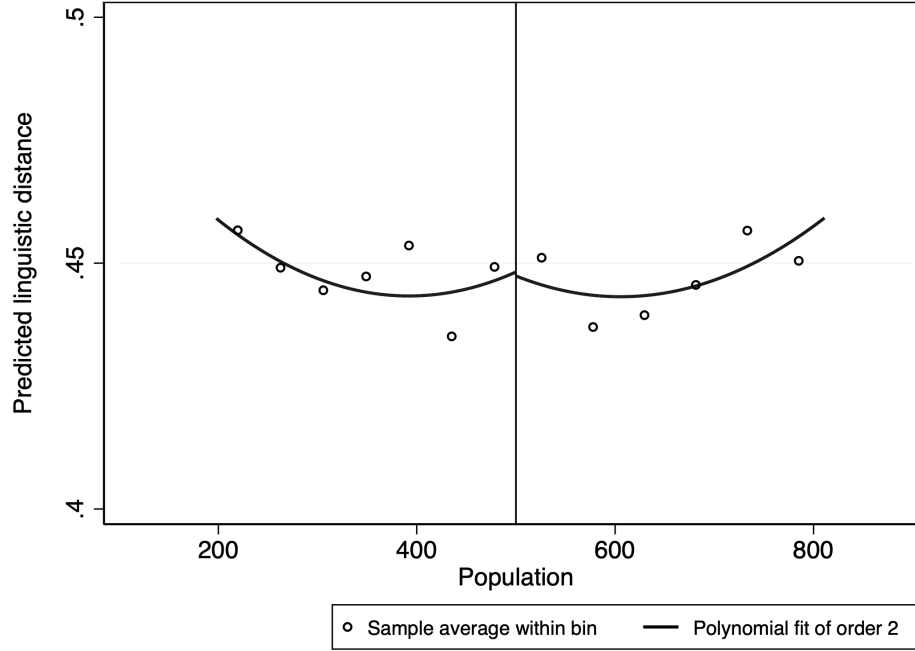
an issue if the building of schools led to in-migration, since out-migration would bias our results toward zero. There is no data on migration flows at the municipality level, yet we argue it is very unlikely to be an important factor. Appendix Figure A16 displays evidence supporting this: there was no jump in the population of treated municipalities after the policy. Moreover, because we estimate a local effect in municipalities with five hundred inhabitants, and because of gravity forces, we expect more out-migration than in-migration following the building of schools, if anything.

5   Schools and Homogenization

5.1   Correlates of Linguistic Distance

Appendix Table A3 shows the correlates of linguistic distance from French as found through ordinary least squares (OLS) regressions. We report standardized beta coefficients and cluster standard errors at the *département* level. In column (1), we show that municipalities in

the historical *langues d'oc* region are much more distant from French than those in the historical *langues d'oïl* region (where French originates from). Hence, our measure of linguistic distance captures both recent variations and, to a large extent, deep-rooted differences. We also show, in column (2), that a one-standard-deviation increase in geographical distance from Paris is associated with three-fourths of a standard deviation increase in linguistic distance. In column (3), we show a positive association between log population at the time of the Guizot Law and linguistic distance. This surprising association likely stems from a combination of factors. The most important being that municipalities in the historical *langues d'oïl* region, around Paris, were on average smaller than those in the *langues d'oc* region in southern France.[43] In column (4), we report a large negative association between linguistic distance and disposable income per capita. In column (5), we show that, surprisingly, the increase in the number of schools built following the law is positively associated with linguistic distance and not statistically significantly. The sign of the correlation likely underscores that the building of schools was not random but was correlated with historical linguistic distance. Indeed, we expect that municipalities in which the language spoken was initially very different from French built more schools, hence confounding the effect of the treatment when using OLS. However, when we include all variables in column (6), in particular controlling for dummies for historical language regions and for geodesic distance from Paris, we find that the building of schools had a negative and statistically significant effect on linguistic distance from French, suggesting we might be capturing a causal effect.

5.2   Main results

Case study.   We first present a detailed preview of our data, the historical context, our empirical strategy, and our results in the historical region of Aquitaine, in the southwest of France, in Figure 6. Panel A plots the map showing the pronunciation of *ciel* ("sky") in the *Atlas*, as in Figure 2. We observe that French was spoken in municipalities located in the middle of the map. Others, in the lower part of the map, were still speaking the Gascon language, one of the *lenga d'òc* languages spoken in the south of France, with the word *cèl* or *cèu* used in place of *ciel*. Panels B and C superimpose the boundaries of municipalities and historical language regions. Panel D displays the treatment status of municipalities near the cutoff following the *Loi Guizot* of 1833, and Panel E displays the Levenshtein distance from the pronunciation of *ciel*. Finally, Panel F displays our measure of linguistic distance across all fifty maps. Untreated municipalities were less likely to speak French. For example, the linguistic distance from French was 0.7 in Saint-Côme (town 645), which was not treated, but only 0.63 in Vélines (town 634), which was treated, even though the

---

[43]Moreover, the *Atlas* was likely a less representative sample in larger municipalities. Additionally, the fact that population is a relatively poor proxy for income after the demographic transition (Galor and Weil, 2000) could play a role, as France had already experienced the demographic transition 100 to 150 years earlier (Blanc, 2020, 2021). Finally, it is very likely that both representativeness and selection into the sample change at higher levels of population (see Section 3.1).

two municipalities are only fifty kilometers, or thirty miles, from each other and had very similar populations.[44]

**Figure 6:** Case study in Aquitaine

Note: This figure constitutes a case study of the historical region of Aquitaine and is designed to understand our data collection (Panels A and B), the historical framework (Panel C), the school-building policy (Panel D), our measure of linguistic distance (Panel E), and our regression-discontinuity results (Panel F). We use the map of *ciel* ("sky") from the *Atlas Linguistique de la France* (Gilliéron and Edmont, 1902-1910). In Panels D and F, "no school" refers to a town with fewer than 500 inhabitants, "school" a town with more than 500 inhabitants, and "school (away from the cutoff)" a town more than 327 inhabitants (the optimal bandwidth, as reported in Table 2) away from the cutoff at the time of the *Loi Guizot*.
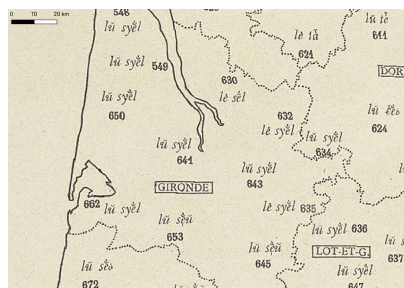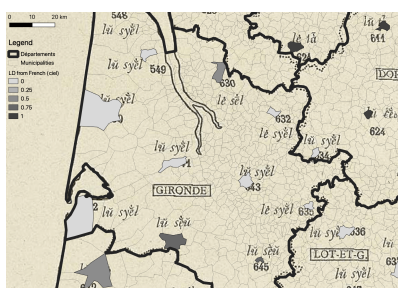

(Panel A) Map of *ciel* ("sky")
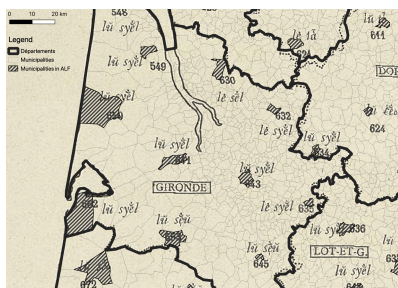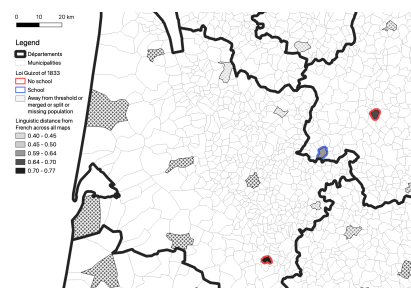

(Panel B) With municipalities
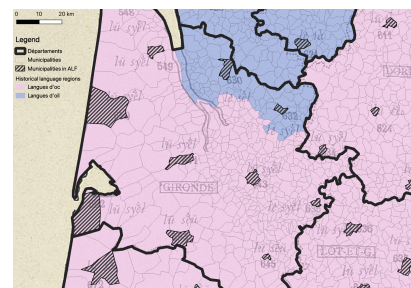
(Panel C) With historical language regions
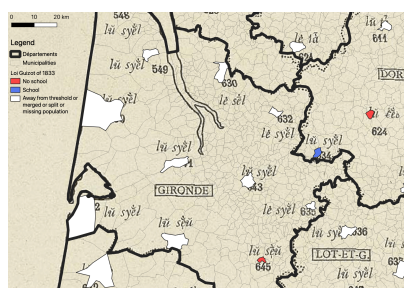
(Panel D) With treatment status after *Loi Guizot* (state-sponsored education)

(Panel E) Linguistic distance from French (*ciel*)

(Panel F) Average linguistic distance from French across 50 maps

Regression discontinuity results. We rely on the discontinuity introduced by the *Loi Guizot* to study the effect of schools on the spread of the French language in Table 2. In column (1), we report results from controlling for a zero-th degree polynomial in population and compare the average linguistic distance with and without schools in a particularly narrow bandwidth around the threshold. The regression in column (2) accounts for linear trends in population and relies on a larger optimal bandwidth. Column (3) accounts for nonlinear, quadratic functions of population. We report the MSE optimal bandwidth and effective number of observations throughout.

We find that the provision of state-sponsored education significantly decreased linguistic distance from French by 0.1 at the threshold. The effect is statistically significant at the

---

[44]When the *Loi Guizot* was passed, there were 446 inhabitants in Saint-Côme and 801 in Vélines.

1 percent level across specifications and is particularly large. A one-standard-deviation increase in schools is associated with 32 to 41 percent of a standard deviation decrease in linguistic distance. We also estimate that the policy decreased linguistic distance by 20 percent relative to the counterfactual mean.

Table 2: Schools and homogenization

	dep var: Linguistic distance from French (1900)		
	(1) Poly. degree 0	(2) Poly. degree 1	(3) Poly. degree 2
RD Estimate of $\mathbb{1}$(School building)			
τ	-0.09***	-0.11***	-0.11***
	(0.03)	(0.03)	(0.04)
Estimated rate of homogenization (perc. decrease from 1833 to 1900)			
τ/α	-0.19***	-0.22***	-0.22***
	[-0.28,-0.10]	[-0.34,-0.11]	[-0.35,-0.09]
	⟨-0.28,-0.10⟩	⟨-0.34,-0.11⟩	⟨-0.36,-0.09⟩
Cutoff	500	500	500
Bandwidth selection	mse	mse	mse
Standardized beta coefficient			
$\tau * sd(school)/sd(y)$	-0.32***	-0.40***	-0.39***
Optimal bandwidth (inhab.)	139	287	327
Effective observations	60	123	141

Weber (1976) documents a process of homogenization in the nineteenth century, but its speed remains unknown.[45] Our data provide the first detailed snapshot of observed linguistic similarity across space at a point in time, yet we do not observe the languages spoken before the policy. We rely on our cross-sectional estimation of $|\tau|/\alpha$, the percentage change relative to the counterfactual, to estimate a lower bound on the rate of homogenization from 1833 to 1900 at the threshold. Our results suggest that linguistic distance from French decreased by at least 20 percent from 1833 to 1900, and that state-sponsored education played a major role in homogenization.

Since τ/α is a nonlinear combination of the expectation of two random variables, we can-

---

[45] We only know that about 10 percent of the population spoke the French language fluently at the time of the Revolution and that French is the common language today.

not easily provide traditional 90 percent confidence intervals (CIs). We report approximate CIs using the delta method, which finds the standard error of the first-order Taylor expansion of the lower bound. We also rely on the Fieller method, which yields asymmetric CIs for the estimate of the lower bound—not its linear approximation—derived from the fact that $\widehat{\alpha}g - \widehat{\tau}$ is normally distributed with mean zero and known variance (Fieller, 1954). The Fieller method is less intuitive and more computationally complex than the delta method, yet it has been shown to be superior in a variety of contexts (Hirschberg and Lye, 2010), in particular when the sample size is small.

Figure 7: Schools and linguistic distance from French

Note: This figure plots linguistic distance from French against population around the discontinuity introduced by the school-building policy. Each point plots the average value within a bin, partialled out of the full set of controls used in Appendix Table A4. Linguistic distance from French is defined as the Levenshtein distance from Le Plessis-Robinson. We apply a local-polynomial fit of order 2 and a mean-squared-error optimal bandwidth for local-polynomial estimation. We rely on an evenly spaced mimicking-variance optimal number of bins. Observations are at the municipality level.
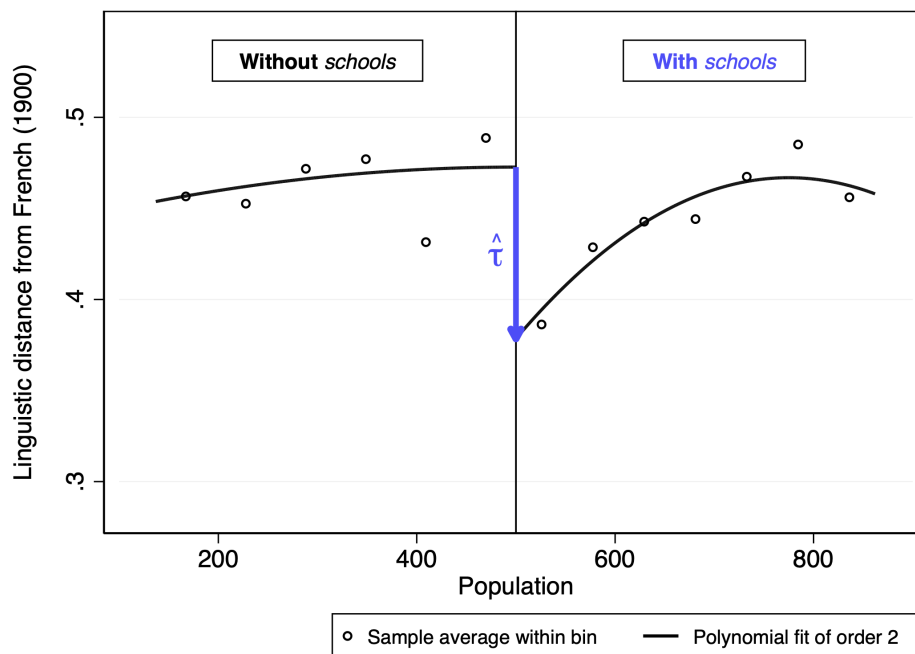
We display our results in Figure 7, which plots linguistic distance from French against population after accounting for both distance from Paris and regions and historical dialects fixed effects.[46] In order to account for pre-existing differences in linguistic distance, and to capture egional institutional and cultural differences, we present estimates of the effect of

[46]We include the full set of controls from Appendix Table A4. Appendix Figure A17 plots the same figure using different polynomial functions of population. We include dialect fixed effects only for the dialects with more than five observations in the sample. Alternatively, we can use *département* fixed effects instead of historical-dialect and region fixed effects. The results (available upon request) do not change.

state-sponsored education on linguistic distance after accounting for covariates in Appendix Table A4 following Calonico et al. (2019). The first column corresponds to the coefficients estimated above. We sequentially add covariates in the subsequent columns. As predicted by the fact that the RD framework allows us to control for these factors, which we expect vary smoothly at the cutoff, the results remain unchanged after.

Our empirical results demonstrate that the common language spoken by those living in treated municipalities became closer to French. This suggests an assimilation into the French culture instead of a backlash, which contrasts with the findings of Fouka (2020) in the US and Bazzi, Hilmy and Marx (2021) in Indonesia. Additionally, because the *Atlas* captured the common language spoken by ordinary people in each municipality, our results are not driven merely by the forced use of French at school; if they were, the students would speak French at school and the other language outside of school. For that reason, our results suggest that the use of the French language reflects the adoption of a national identity.

5.3 Robustness and Placebo

We perform a range of standard robustness and placebo analyses in Appendix 3. We show the robustness of the estimated coefficients to (a) different methods of estimation, kernels, and bias correction in Appendix Table A5; (b) alternative, non-optimal bandwidths in Appendix Table A6; (c) accounting for individual-level characteristics in Appendix Tables A7 and A8; and (d) dropping municipalities near the national border to account for bilingualism in Appendix Table A9. We also show, in placebo tests, that (e) only the five-hundred-inhabitant threshold returns statistically significant estimates (Appendix Table A10) and (f) only population at the time of the *Loi Guizot* mattered (Appendix Table A11). Finally, our results are robust to the use of different distance metrics and to measuring distance across the entire corpus of maps from the *Atlas* instead of only the 50 words from the Swadesh list (Appendix Table A12).

5.4 Discussion and Interpretation

Origins of French. Here, we attempt to empirically trace out the geographical origins of the French language; in addition, we validate our reliance on the town of Le Plessis-Robinson, ten kilometers from Paris, as the reference point for French. While historical and anecdotal evidence suggests that the region around Paris was the center of the origin of the French language (see Lodge, 1993; Weber, 1976), no empirical evidence exists. We take advantage of the *Atlas* to compute linguistic distance across any pair of towns in order to understand what language was adopted following the policy and where that language came from. We evaluate the effect of schools on linguistic distance from all the

municipalities in the *Atlas* in order to locate the regions in which French originated.[47] We essentially estimate, for each municipality of origin j, the following modified version of Equation 1, where $LD_{i,j}$ is the linguistic distance between i and j and j is fixed:

$$(2) \qquad LD_{i,j} = \alpha^{j} + \tau^{j} \times school_i + f^{j}(population_i) + \varepsilon_{i,j}$$

In Appendix Table A13, we display RD estimates of the effect of schools on linguistic distance from each of the six towns closest to Paris: Le-Plessis-Robinson (10 km), Sartrouville (14 km), Liancourt-Saint-Pierre (51 km), Ormoy-la-Rivière (52 km), Droue-sur-Drouette (54 km), Gommecourt (58 km). To ease comparability, we rely on the MSE optimal bandwidth estimated in the baseline regression. We find that schools brought the observed language closer to the language spoken in all of these towns, with the largest coefficient on Le-Plessis-Robinson. We extend the analysis to the rest of the municipalities in the *Atlas* in Appendix Figure A19. Panel A plots the estimated coefficients against the distance between the reference town and Paris, while Panel B plots the t-statistics. We first remark that schools predominantly spread the language spoken in the municipalities within four hundred kilometers from Paris—that is, in the historical *langues d'oïl* region. Second, there is a positive association between the estimated coefficient and the distance of the reference town from Paris, suggesting that the language spoken near Paris (in particular the francien dialect), was adopted as schools were built.

Finally, our framework enables us to empirically estimate the geographical origins of the French language, instead of assuming that the reference point was Paris. Appendix Figure A20 maps our estimates of the effect of schools on linguistic distance from different reference towns across space. We plot each coefficient τ^{j} in the location of the reference town j in the regression. We find that state-sponsored education predominantly led to the adoption of the language spoken near Paris and Touraine, in the Loire Valley, since the largest statistically significant estimated coefficients are for the reference towns in these regions. Our results suggest that French originated in Paris and the Loire Valley. The castles of the Loire Valley housed the royal court during the Renaissance, and the region (along with Paris) is sometimes believed to be one of the historical geographical centers of French (for example Lodge, 1993). Hence, our results lend credence to this hypothesis and confirm the anecdotal and historical evidence suggesting that French was the language of the elites.

Language and pronunciation. In this section, we compute a measure of linguistic distance accounted for by language only. We abstract from the phonetic accents in the *Atlas*, using the same Levenshtein-distance algorithm as in our measure of total linguistic distance across 50 representative words. While our baseline measure of linguistic distance

---

[47]Barjamovic et al. (2019) undertake a similar attempt in estimating the geographical locations of historical cities based on Assyrian trade links. However, they use a structural-gravity model to back them out.

Figure 8: Share of linguistic distance from French accounted for by language

Note: This figure displays the share of linguistic distance accounted for by language (excluding phonetic accents), plotted against total linguistic distance from French, including accents. Linguistic distance from French is defined as the Levenshtein distance from Le Plessis-Robinson. Observations are at the municipality level.

○ % of linguistic distance explained by language
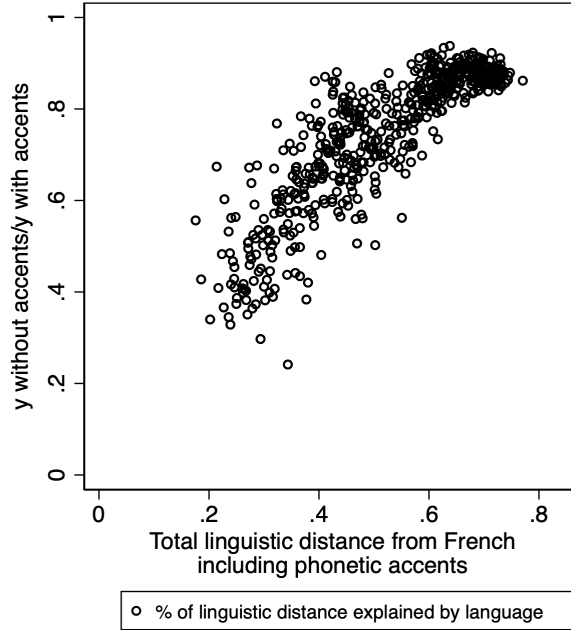
considers phonetic accents and captures both language and pronunciation, including sound replacements, the measure of distance excluding accents captures only differences in language and vocabulary. For example, the "e" of *seu* (*ciel* in French, "sky" in English, see Figure 2) is pronounced as a long vowel in town 645 and as a short vowel in town 653 (see Gilliéron and Edmont, 1902, p. 19). The total linguistic distance between 645 and 653 is therefore one third, including phonetic accents. However, the linguistic distance excluding phonetic accents is zero, capturing the fact that both of these municipalities used the word *seu* and spoke dialects from the same family of language: *langues d'oc*. Figure 8 plots the share of total linguistic distance accounted for by language (not pronunciation). We remark that, for municipalities speaking a language similar to French, almost half of the total measure of distance is accounted for by pronunciation. For municipalities speaking a very different language, often in the historical *langues d'oc* region, 90 percent of the difference is explained by language and less than 10 percent by differences in pronunciation.

Table 3 reports estimates of the effect of schools, or state-sponsored education, on linguistic distance including and excluding phonetic accents. Throughout the table we use the MSE optimal bandwidth that was selected in the baseline regression to ease comparability. Odd columns report the baseline estimates from the main table. We report our

estimates of the effect of schools on language only in the even columns. We estimate that about three-fourths of the effect of schools on linguistic distance from French comes from the effect on language. The remaining one-fourth comes from pronunciation. The effect on language is statistically significant, suggesting that schools spread not only the standard pronunciation but also, and predominantly, the standard language.

Table 3: Schools and linguistic distance from French: language and pronunciation

Note: This table displays regression-discontinuity estimates of the effect of schools on linguistic distance from French (including and excluding phonetic accents) using the discontinuity introduced by the school-building policy. Linguistic distance from French is defined as the Levenshtein distance from Le Plessis-Robinson. By excluding phonetic accents (in even columns), we compute linguistic distance without taking into account different pronunciations. School is defined as a dummy variable that equals 1 if population at the time of the law was above five hundred. Columns (1), (3), and (5) show our baseline estimates. We use local-polynomial fits of orders 0, 1, and 2 (mean, linear, quadratic) and, across all specifications, rely on the baseline mean-squared-error optimal bandwidths estimated in the baseline (including accents) for local-polynomial estimation. Observations are at the municipality level. $^{+}$ $p < 0.2$, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

	dep var: Linguistic distance from French (1900) with and without phonetic accents					
	Poly. degree 0		Poly. degree 1		Poly. degree 2	
	(1)	(2)	(3)	(4)	(5)	(6)
RD Estimate of $\mathbb{1}$(School building)	-0.09***	-0.07**	-0.11***	-0.09**	-0.11***	-0.09*
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Cutoff	500	500	500	500	500	500
Bandwidth selection	mse		mse		mse	
Without phonetic accents: language only		Yes		Yes		Yes
Bandwidth (inhab.)	139	139	287	287	327	327
Effective observations	60	60	123	123	141	141

Finally, Lodge (1993) documents a process of language shift, or abandonment, in the historical *langues d'oc* region and a gradual convergence of vernacular dialects toward the common language in the *langues d'oïl* region in the north (p. 190). We find similar results in Appendix Table A14, which reports estimates of the effect of schools on phonetic distance from French—defined as the difference between linguistic distance including and excluding phonetic accents—in the different historical language regions. The coefficient on the interaction between school and historical *langues d'oïl* region is negative and statistically significant, suggesting a process of language convergence took place in that region following the adoption of state-sponsored education. We report similar results in Appendix Table A15. We find that schools mostly led to the homogenization of language, not pronunciation, in the historical *langues d'oc* region, suggesting a language-shift process in this region.

5.5 Mechanisms and Drivers of Homogenization

Here, we explore the heterogeneous effects of the treatment to better understand the drivers and mechanisms of homogenization. We discuss possible mechanisms that suggest a va-

riety of channels were at work. Throughout the section, using the threshold for state-sponsored schools, we estimate Equation 3 with different variables of interest x_i and we cluster standard errors at the appropriate level since some variables are only available at the *département* level.

$$(3) \qquad y_i = \alpha + \tau_1 \times school_i + \tau_2 \times school_i \times x_i + f(population_i) + \beta \times x_i + \varepsilon_i$$

Local elites. Because nation-building was conducted by elites and standard French was the language of the elites (Lodge, 1993; Weber, 1976), we expect that elites were instrumental to homogenization. Furet and Ozouf (1977) argue that, for the elites, "school was first and foremost an instrument of control, designed to moralize and discipline the masses" (p. 120). We now exploit variation in the distribution of elites at the municipality level to understand their role in homogenization.

Figure 9: Heterogeneous effect of schools: local elites

Note: This figure displays the effect of schools on linguistic distance from French—at low and high levels of presence of local elites—using the discontinuity introduced by the school-building policy. Linguistic distance from French is defined as the Levenshtein distance from Le Plessis-Robinson. School is defined as a dummy variable that equals 1 if population at the time of the law was above five hundred. The presence of local elites is proxied by the log of subscriptions to the *Encyclopedie* per capita within a 50 km buffer around each municipality (available at the municipality level) and is coded as high if it is above the median value. We apply a local-polynomial fit of order 0, for ease of interpretation of the interaction term, and a mean-squared-error optimal bandwidth for local-polynomial estimation. We present the results for other polynomial degrees in Appendix Table A16. The regression-discontinuity coefficient is estimated with ordinary least squares (see Appendix Table A5). Observations are at the municipality level, and standard errors are clustered at the municipality level. Source: Darnton (1973) for presence of local elites.
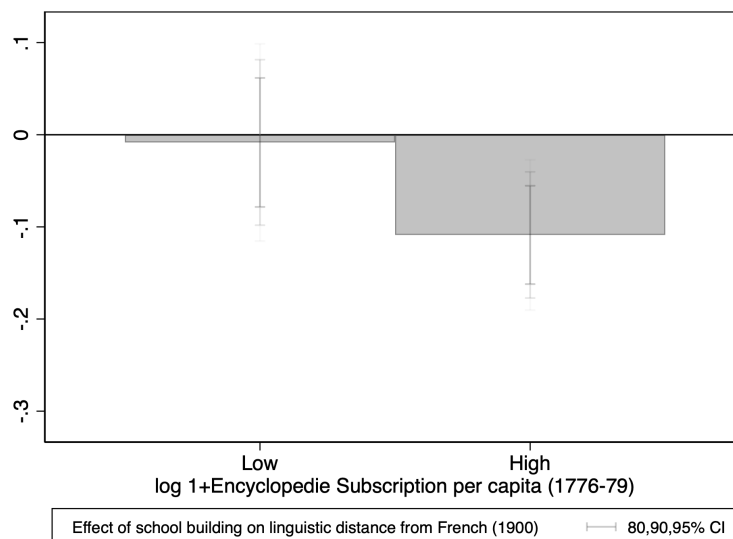

Figure 9 shows that the effect of state-sponsored schools is more than twice as large in places with many elites nearby, as proxied by *Encyclopedie* subscriptions per capita in the period 1776—79 (Darnton, 1973).[48] Subscriptions per capita are associated with

[48]We use a municipality-level measure of *Encyclopedie* subscriptions per capita within a fifty-kilometer radius of

the presence of scientific societies and famous people, and more generally they capture the presence of local elites (Squicciarini and Voigtländer, 2015). Darnton (1973) (p. 1350) discusses the occupations of 137 subscribers in the city of Besançon. The vast majority were political elites: parliamentarians, members of the military, lawyers, local administrators. These people had strong incentives to build schools, and Squicciarini and Voigtländer (2016) show that the presence of elites is indeed associated with school building under the Guizot Law.

The interaction between schools and elites is likely the outcome of two distinct processes. First, at the intensive margin, local elites were directly involved in the administration of public schools and in the curriculum. The schoolteachers were supervised by local municipal councils, and district councils (*comités d'arrondissement*) had the authority to influence the curriculum, within the limits of the law, by choosing—from the list of textbooks authorized by the state—the textbooks schoolteachers would use (Choppin, 1986, p. 37). Appendix Figure A21 displays evidence supporting this; namely, the effect of schools was smaller in municipalities farther away from departmental capitals (*préfectures*). Second, elites built more schools. Local municipal councils voted for the building of schools and arranged for the collection of additional resources in accordance with the law.[49] This is particularly relevant to our context because the estimated coefficient is an intention-to-treat effect.

Finally, one concern is that we might be capturing not the effect of elites but rather that of income. For example, Squicciarini and Voigtländer (2015) document that subscriber density is positively associated with disposable income and various measures of development. Appendix Figure A22 shows that the effect of schools was the same across all levels of development, which alleviates this concern.

State capacity and legitimacy. We also expect that the government wanted to homogenize regions where state capacity was weak and where the legitimacy and authority of the state were contested. Alesina, Giuliano and Reich (2021) argue that the threat of democratization is the strongest determinant of homogenization. The threat was particularly high at the time, following the French Revolution of 1789 and the July Revolution of 1830 and before the February Revolution of 1848. When the legitimacy of the state is contested, the population is more likely to revolt and overthrow a ruler seen as illegitimate. This relationship also suggests an interaction between state building and nation-building. We document the relationship using data on historical language regions as well as data on fiscal capacity in the Ancien Régime and on the refractory clergy—that is, clergy who opposed the secular revolutionary state in 1791.

Appendix Figure A23 shows that schools had a larger effect in the historical *langues d'oc* region. However, as already explained, since towns in the south were larger, there were few

each municipality. Squicciarini and Voigtländer (2015) use a *département*-level measure and discard rural towns from their sample.

[49]See Blanc and Wacziarg (2020) for a detailed account of this in the municipality of Saint-Germain-d'Anxure.

towns in that region near the threshold for school building and so we may lack sufficient statistical power to identify a meaningful effect. Appendix Figure A24 documents a similar pattern and shows that state-sponsored schools had a stronger effect in the regions outside of the Cinq Grosses Fermes (CGF) in Ancien Régime France. The CGF was a system of tax farms and a unified customs union including provinces in the region around Paris and ambitiously created in 1664 by Jean-Baptiste Colbert, Louis XIV's controller-general of finances. Johnson (2019) argues that the CGF strengthened national institutions and legitimacy, increased fiscal capacity and tax collection, and undermined the feudal system as a whole. This result suggests that nation-building targeted and was more successful in poorly integrated regions, where state capacity and national identity were initially weak.

Last, Appendix Figure A25 shows that schools had a stronger effect in places with a high share of refractory clergy. The Civil Constitution of the Clergy, passed in July 1790, required clergy members to take an oath of allegiance to the secular state (Tackett, 1986). The presence of refractory clergy is commonly used as a proxy for devotion at the time of the French Revolution (Blanc, 2021; Squicciarini, 2020), after the process of secularization that took hold in the mid-eighteenth century. We argue that the state targeted places that had not secularized as a way to assert its legitimacy and authority in opposition to the Catholic Church. Our results also suggest that a backlash from the clergy may have taken place, similar to what Bazzi, Hilmy and Marx (2021) find in Indonesia, since the difference between secularized and non-secularized regions is small. Koepke (1995) documents, for example, that the clergy in Morbihan "did not want the children to learn French, as opposed to their native Breton, for such knowledge would open the young to dangerous books and even more dangerous conversations and liaisons with the irreligious French" (p. 602).

Demand for schooling. Finally, we show that returns to education and educational demand played an important role in shaping the impact of the policy. Appendix Figure A26 finds that the effect of schools on linguistic distance was larger in *départements* where the number of students in public schools increased the most. This could be interpreted as the effect of the demand for schooling or as the effect of the positive externality of adopting a common language. In Appendix Figure A27, we further show that schools had a larger effect in places where returns to education were high, as proxied by the horsepower of steam engines (see Franck and Galor, 2021). These results suggest that where returns to education and educational demand were important, more children attended the primary schools built under the Guizot Law and, as a result, their effect was larger since compulsory schooling was not mandated by the law.[50] Because we were concerned that these results

[50]Note that it would only be a threat to identification, by introducing selection at the cutoff, if educational demand systematically impacted the building of schools and not only school attendance. However, we exploit quasi-exogenous variation in the supply of schools. While the supply of schools may have been impacted by educational demand in a limited number of cases, it probably was not a widespread phenomenon, since the law made *départements* or the state pay for the building of the school when municipalities would not or could not raise additional taxes to finance it. Hence, educational demand is very unlikely to introduce endogeneity and affect our results.

could be driven by certain geographical features, we investigated this possibility and report in Appendix Figure A28 that the effect of schools was the same in places both near and far from coal fields.

6 PERSISTENT EFFECTS OF NATION-BUILDING

This section explores the persistent effects of the policy on society. We provide evidence strongly suggesting that the adoption of a standard language also promoted the exchange of ideas and goods, as captured by migration and trade, and we document a persistent effect on the salience of national identity and preference for the centralization of political authority.

Migration and trade. We use data on migration flows across *départements* and trade links across districts to estimate the relationship between linguistic proximity and the exchange of ideas and goods. There is unfortunately no data on migration and trade flows across municipalities, neither historically nor in contemporary times, hence we can only provide suggestive evidence of the role of the policy on these variables. We estimate Equation 4 with ordinary least squares.[51]

$$(4) \qquad y_{d(i),d(j)} = \alpha + \beta \times LD_{i,j} + \boldsymbol{X}'_{i,j}\delta + \varepsilon_{i,j}$$

with $LD_{i,j}$ is the linguistic distance between municipalities i and j; $\boldsymbol{X}_{i,j}$ a vector of control variables to account for gravity factors (namely, the geographical distance between i and j, the log of the population of $d(i)$, and the log of the population of $d(j)$); and $y_{d(i),d(j)}$ a measure of migration or trade flows from administrative unit $d(j)$ to unit $d(i)$.[52] Migration is observed at the pair-of-*départements* level while trade is observed at the pair-of-districts level.

Table 4 presents our results. We report standardized beta coefficients for ease of interpretation, and we cluster standard errors at the pair-of-*départements* or pair-of-districts level. We first consider migration, defined as $log(1+$percentage of population living in $d(i)$ but born in $d(j))$, in 1891, 1896, and 1911, using data from the census on eighty-two *départements* (Statistique Générale de la France, n.d.b). We find that a one–standard deviation increase in linguistic distance between municipalities is associated with a decrease in migratory flows of more than half of a standard deviation. We account for gravity factors in column (2). The coefficient on geographical distance is particularly large, yet the effect of linguistic distance remains meaningful and statistically significant, though smaller, at the

---

[51]We cannot use RD because the level of observation (pairs of districts or pairs of *départements*) is more aggregated than the municipality level and because using pairs creates estimation issues that, to the best of our knowledge, have not been theoretically discussed.

[52]Because we observe migration in 1891, 1896, and 1911, we additionally control for census-year fixed effects in the migration regressions.

Table 4: Linguistic distance, migration, and trade across pairs of towns

Note: This table displays ordinary least squares estimates of the effect of linguistic distance on migration and trade across pairs of towns. Linguistic distance between towns i and j is defined as the Levenshtein distance between i and j. Migration and trade are observed between the *département* or district of i, $d(i)$, and that of j, $d(j)$. Migration is defined as log(1+percentage of population living in $d(i)$ born in $d(j)$) in 1891, 1896, and 1911 and is available at the pair-of-*départements* level. Trade is defined by a dummy variable that equals 1 if district $d(i)$ is being supplied in goods by district $d(j)$ in 1794. Coefficients are standardized and standard errors clustered at the pair-of-administrative-units level; that is, $d(i)$ and $d(j)$ on the one hand and $d(j)$ and $d(i)$ on the other hand are considered to be the same cluster since the distance measure is symmetric. We control for the log of the population of i and the log of the population of j in even columns, and we include census-year fixed effects in the first two columns. Sources: Statistique Générale de la France (n.d.b) for population Daudin (2010) for trade. $^+$ $p < 0.2$, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Standardized beta coefficient (clustered se)	*dep var* is:			
	Migration		Trade	
	(1)	(2)	(3)	(4)
Linguistic distance between i and j (1900)	-0.55***	-0.08***	-0.19***	-0.04***
	(0.02)	(0.01)	(0.01)	(0.01)
Geographical distance between i and j		-0.82***		-0.26***
		(0.01)		(0.01)
log population of i and j		Yes		Yes
Observations	914,112	914,112	55,752	53,400
Clusters (pairs of *départements*)	3,403	3,403		
Clusters (pairs of districts)			24,321	22,803
Adjusted R^2	0.30	0.78	0.04	0.08

1 percent level.

Then we study the determinants of trade links, defined by a district-level dummy variable that equals 1 if district $d(i)$ was supplied with goods from district $d(j)$ in 1794. Hence, these regressions leverage the deep-rooted, historical component of linguistic distance and we estimate a lower bound on the strength of the association at the time of the *Atlas*. Daudin (2010) provides data on 88 receiving districts and 551 supplier districts; less than 20 percent of pairs were linked by trade.[53] We match 67 supplier and 396 consumer districts to municipalities in the *Atlas*. In column (3), we report that a one–standard deviation increase in linguistic distance is associated with decrease in trade of 19 percent of a standard deviation. After accounting for geographical distance in the last column, the coefficient on linguistic distance diminishes but remains statistically significant. Geographical factors appear to play a smaller role in trade links than in migratory flows.

Our results suggest that language shapes migration and trade and that the homogenization brought about by state-sponsored education likely had far-reaching and persistent impacts. Although we cannot rely on the discontinuity introduced by the *Loi Guizot* to account for confounding and pre-existing factors—because of the high level of aggregation of the data and because of the nature of the regression (namely, the use of pairs)—the fact

---

[53]When we aggregate the existence of links across pairs of *départements*, we find that 50 percent were linked by trade. The coefficient on linguistic distance is larger at this level, likely because there are fewer zeros in the data.

that the effect of linguistic distance is large and survives the inclusion of controls for gravity is remarkable and suggests a causal interpretation. Additionally, our study is one of the first to document an association between linguistic distance and migration and trade flows at the subnational level. Falck et al. (2012) report very similar results and magnitudes in the context of Germany.

National identity. This section explores the persistent effects of the policy on the salience of national identity and preferences for the centralization of political authority. We leverage data on a set of different outcomes to proxy for national identity and political preferences, in particular data on the birthplace of heroes resisting Nazi Germany during World War II and data on votes against the regionalization of political authority in 1969.[54] Figure 10 demonstrates that state-sponsored education increased the salience of national identity and political preferences for centralization. We rely on the discontinuity introduced by the school-building policy, with the adoption of the French language now part of the bundled treatment.

Figure 10: Schools and national identity

Note: This figure plots measures of the salience of national identity against population around the discontinuity introduced by the school-building policy. Each point plots the average value within a bin, partialled out of the full set of controls used in Appendix Tables A17 (Panel A) and A19 (Panel B). In Panel A, French Resistance during World War II is defined by a dummy variable that equals 1 if at least one individual born in the municipality was awarded a Médaille de la Résistance for documented "remarkable acts of faith and courage that contributed to the resistance of the French people against the enemy." In Panel B, votes on the 1969 constitutional referendum on regionalization are defined as the vote share in favor of the referendum. We apply a local-polynomial fit of order 2 and a mean-squared-error optimal bandwidth for local-polynomial estimation. Appendix Figures A29 and A33 display additional polynomials. We rely on the evenly spaced mimicking-variance optimal number of bins using spacing estimators from Figure 7. Observations are at the municipality level. Sources: Site Mémoire des hommes (n.d.) for French Resistance and Dehdari and Gehring (forthcoming) and Archives départementales du Var (1969) for the 1969 referendum.
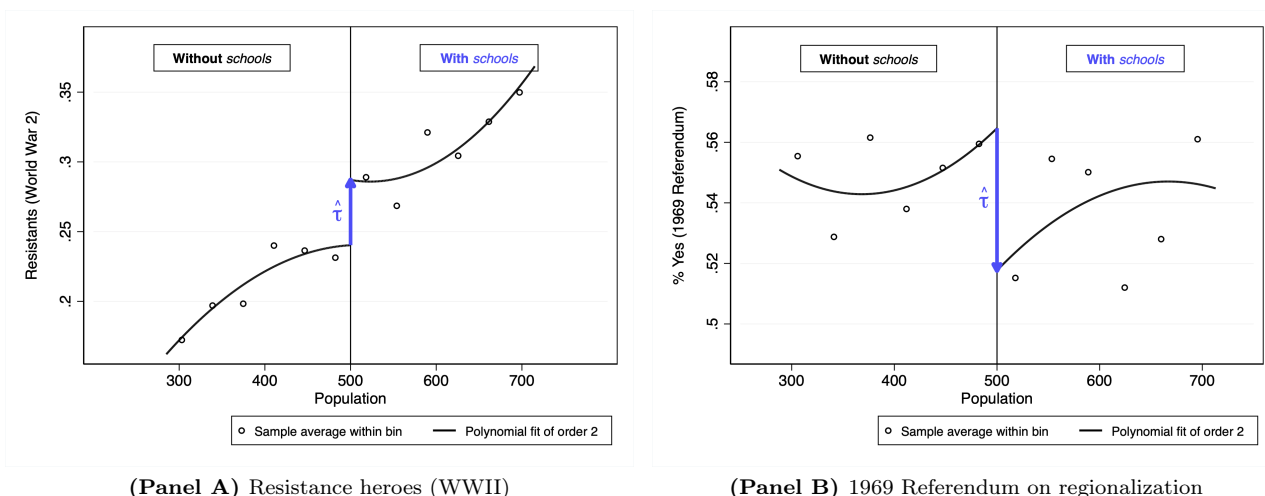
(Panel A) Resistance heroes (WWII)

(Panel B) 1969 Referendum on regionalization

We first show, in Panel A of Figure 10, that the policy significantly increased the likeli-

[54]Note that we are limited by the granularity of our empirical strategy, which relies on municipality-level data on towns around five hundred inhabitants. Standard surveys that could capture national identity do not include individuals living in such small municipalities.

hood that individuals born in treated municipalities were heroes of the French Resistance and were awarded a Médaille de la Résistance for documented "remarkable acts of faith and courage that contributed to the resistance of the French people against the enemy."[55] We carefully match 42,441 medals to 11,783 of the 33,814 municipalities in our sample and report the results of our estimation for different polynomial degrees in Appendix Table A17. We control for geographical distance from Paris and for *département*, historical-dialect, and Vichy-regime (the French government that chose the path of collaboration with the Nazis) fixed effects in order to account for local institutional and cultural factors.[56]

Then, we use data on votes against the regionalization of political authority in the 1969 referendum on regionalization, where Charles De Gaulle proposed to return significant power and autonomy to regional authorities. However, the outcome of the vote is not readily available at the municipality level. Dehdari and Gehring (forthcoming) study this referendum in the context of Alsace-Lorraine and digitize the municipality-level results in that region from a local newspaper that published the results the day after the election. Similarly, we obtain municipality-level results in Provence from another local newspaper, *Le Petit Varois*. We searched for the edition published the day after the election in the Archives départementales du Var (1969) and digitized it. Appendix Figure A31 displays scanned pages of the newspapers. We use the data from both Lorraine and Provence and display the spatial distribution of our data in Appendix Figure A32.

In Panel B of Figure 10, we establish that state-sponsored education significantly reduced the vote share in favor of regionalization by 2 to 6 percentage points. We detail the results in Appendix Table A19. This is a particularly important effect since the referendum was rejected by 52.4 percent of voters. The referendum would have been adopted and France would look more like Spain today had nation-building not taken place in the 1830s. We break down the effect of the treatment by region in Appendix Table A20. The magnitude is similar, but, because of the small number of observations in each region, we barely reach conventional levels of statistical significance.

Finally, we perform a number of additional analyses.[57] First, in Appendix Table A22, we

---

[55]Similarly, Cage et al. (2020) write a fascinating account of the determinants of collaboration and support for authoritarian leaders during World War II using declassified intelligence data on collaborators, which they digitized. Unfortunately, the digitized data is not yet publicly available.

[56]Additionally, we use data on 375,346 Resistance fighters from a corpus of reported acts of resistance, also taken from Site Mémoire des hommes (n.d.). Most of these individuals were not awarded Resistance Medals. This data arguably provides more variation but includes less heroic acts that are often unverified and unconfirmed. According to Site Mémoire des hommes (n.d.), "This corpus is not a list of all resistance fighters. Indeed, many people who engaged against the occupier during World War II never made themselves known. In addition, certain forms of resistance have not given rise to administrative procedures leading to the constitution of an individual file. Conversely, the mention of an individual in the database does not presume his capacity as a resistance fighter because the corpus contains all files opened by the Service historique de la défense, including those of individuals who, after examination, have not been recognized or approved for acts of Resistance." We display the results in Appendix Figure A30 and Appendix Table A18. Our results are rather imprecise but suggest an effect of one to ten additional Resistance fighters per ten thousand inhabitants in treated municipalities. After controlling for *département*, historical-dialect, and Vichy-regime fixed effects to account for pre-existing institutional and cultural differences and to account for differences in the reporting of acts of resistance, we improve the precision of the estimator.

[57]Appendix Table A21 also reports a placebo test and shows that the number of individuals *Morts pour la France*—individuals who died in action or from an injury or an illness contracted during service during World War I or II—is

find no statistically significant effect of the treatment on votes for Jean-Marie and Marine Le Pen in the 2002 and 2017 presidential elections, respectively. This null result could be explained by a number of factors, in particular it is possible that persistence broke down as a result of subsequent policies. Yet the view that the emergence of nation-states gave rise to nationalism or ethnic cleansing is widely held among historians (see Gellner, 1983b; Hobsbawm, 1990; Ther, 2014). Our null results are therefore surprising and suggest otherwise, that nation-building did not have any impact on nationalism broadly defined. Second, we report in Appendix Table A23 a small but positive effect of the policy on votes in favor of the 1992 referendum on the Maastricht Treaty and the 2005 French European Constitution Referendum. This suggests that nation-building not only affected national identity but also led to greater moral universalism.

7   Concluding Remarks

Using a novel, detailed dataset on spoken languages in France and a natural experiment, our paper documents the adoption of a common language and the formation of a national identity in the process of nation-building in a fragmented society.

References

Adsera, Alicia, and Mariola Pytlikova. 2015. "The role of language in shaping international migration." *The Economic Journal*, 125(586): F49–F81.

Aghion, Philippe, Xavier Jaravel, Torsten Persson, and Dorothée Rouzet. 2019. "Education and military rivalry." *Journal of the European Economic Association*, 17(2): 376–412.

Ahlerup, Pelle, and Ola Olsson. 2012. "The roots of ethnic diversity." *Journal of Economic Growth*, 17(2): 71–102.

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and identity." *The quarterly journal of economics*, 115(3): 715–753.

Alesina, Alberto, Bryony Reich, and Alessandro Riboni. 2020. "Nation-building, nationalism, and wars." *Journal of Economic Growth*, 25(4): 381–430.

Alesina, Alberto, Paola Giuliano, and Bryony Reich. 2021. "Nation-Building and Education." *The Economic Journal*. ueab001.

Algan, Yann, Pierre Cahuc, and Andrei Shleifer. 2013. "Teaching practices and social capital." *American Economic Journal: Applied Economics*, 5(3): 189–210.

Anderson, Benedict. 2006. *Imagined communities: Reflections on the origin and spread of nationalism.* London:Verso.

Archives départementales du Var. 1969. *Le Petit Varois, la Marseillaise.* Cote 326 PRS 79.

Assouad, Lydia. 2021. "Charismatic Leaders and Nation Building." Working paper.

not different in treated and untreated municipalities. The results suggest that, on average, soldier death is likely to be more the product of idiosyncratic forces than of endogenous forces somewhat related to national identity.

Baiwir, Esther. 2019. "Francisation des dialectes d'oïl : de l'usage des atlas linguistiques comme termes de comparaison." *Langages*, 215(3): 27–42.

Bandiera, Oriana, Myra Mohnen, Imran Rasul, and Martina Viarengo. 2019. "Nation-building through compulsory schooling during the age of mass migration." *The Economic Journal*, 129(617): 62–109.

Barjamovic, Gojko, Thomas Chaney, Kerem Coşar, and Ali Hortaçsu. 2019. "Trade, Merchants, and the Lost Cities of the Bronze Age." *The Quarterly Journal of Economics*, 134(3): 1455–1503.

Bazzi, Samuel, Arya Gaduh, Alexander D Rothenberg, and Maisy Wong. 2019. "Unity in diversity? How intergroup contact can foster nation building." *American Economic Review*, 109(11): 3978–4025.

Bazzi, Samuel, Masyhur Hilmy, and Benjamin Marx. 2021. "Islam and the State: Religious Education in the Age of Mass Schooling." Working paper.

Billy, Pierre-Henri, Sébastien Nadiras, and Jean-François Moufflet. n.d.. "Carte linguistique de la France, Archives Nationales (www.archives-nationales.culture.gouv.fr/fr_FR/web/guest/cartographie-historique)."

Binzel, Christine, Andreas Link, and Rajesh Ramachandran. 2021. "Vernacularization, Knowledge Creation, and Growth: Evidence from the Protestant Reformation." Centre for Economic Policy Research Working Paper 15454.

Biraben, Jean-Noël. 1963. "Inventaire des listes nominatives de recensement en France." *Population*, 2.

Blanc, Guillaume. 2020. "Demographic Change and Development from Crowdsourced Genealogies in Early Modern Europe." Working paper.

Blanc, Guillaume. 2021. "The Cultural Origins of the Demographic Transition in France." Working paper.

Blanc, Guillaume, and Romain Wacziarg. 2020. "Change and Persistence in the Age of Modernization: Saint-Germain-d'Anxure, 1730-1895." *Explorations in Economic History*, 78.

Blasi, Damián E, Steven Moran, Scott R Moisik, Paul Widmer, Dan Dediu, and Balthasar Bickel. 2019. "Human sound systems are shaped by post-Neolithic changes in bite configuration." *Science*, 363(6432).

Blouin, Arthur, and Julian Dyer. 2021. "How Cultures Converge: An Empirical Investigation of Trade and Linguistic Exchange." University of Toronto, Department of Economics Working Papers tecipa-691.

Blouin, Arthur, and Sharun W Mukand. 2019. "Erasing ethnicity? Propaganda, nation building, and identity in Rwanda." *Journal of Political Economy*, 127(3): 1008–1062.

Bourdon, Étienne. 2017. *Forging the nation ahead with the Gauls: Ernest Lavisse and the making of ancestors.* Lyon:ENS Éditions.

Braudel, Fernand. 1986. *The identity of France: History and environment ; translated from the French by Sian Reynolds.* Vol. 1. 1990 ed., London:Collins.

Cáceres-Delpiano, Julio, Antoni-Italo De Moragas, Gabriel Facchini, and Ignacio González. 2021. "Intergroup contact and nation building: Evidence from military service in Spain." *Journal of Public Economics*, 201: 104477.

Cage, Julia, Anna Dagorret, Pauline A. Grosjean, and Saumitra Jha. 2020. "Heroes and Villains: The Effects of Combat Heroism on Autocratic Values and Nazi Collaboration in France." Stanford University Graduate School of Business Research Paper 3923.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*, 82(6): 2295–2326.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2019. "Regression Discontinuity Designs Using Covariates." *The Review of Economics and Statistics*, 101(3): 442–451.

Calvet, Louis-Jean. 1974. *Linguistique et colonialisme : Petit traité de glottophagie.* Paris:Éditions Payot.

Cantoni, Davide, Yuyu Chen, David Y Yang, Noam Yuchtman, and Y Jane Zhang. 2017. "Curriculum and ideology." *Journal of Political Economy*, 125(2): 338–392.

Carvalho, Jean-Paul, and Mark Koyama. 2016. "Resisting education."

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association*, 115(531): 1449–1455.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations. Elements in Quantitative and Computational Methods for the Social Sciences*, Cambridge University Press.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *The Quarterly Journal of Economics*, 125(1): 215–261.

Cervellati, Matteo, Giorgio Chiovelli, and Elena Esposito. 2019. "Bite and Divide: Malaria and Ethnolinguistic Diversity."

Choppin, Alain. 1986. "Le cadre législatif et réglementaire des manuels scolaires. I, De la Révolution à 1939." *Histoire de l'éducation*, 29.

Clots-Figueras, Irma, and Paolo Masella. 2013. "Education, language and identity." *The Economic Journal*, 123(570): F332–F357.

Code de l'instruction primaire. 1834. *Code de l'instruction primaire contenant l'historique de la législation primaire depuis 1783.* Paris:Librairie Normale de P. Dupont.

Coveney, Joseph. 2004. "FIELLER: Stata module to calculate confidence intervals of quotients of normal variates." Statistical Software Components, Boston College Department of Economics.

Darnton, Robert. 1973. "The Encyclopédie Wars of Prerevolutionary France." *The American Historical Review*, 78(5): 1331–1352.

Daudin, Guillaume. 2010. "Domestic Trade and Market Size in Late-Eighteenth-Century France." *The Journal of Economic History*, 70(3): 716–743.

Day, C. R. 1983. "The Rustic Man: The Rural Schoolmaster in Nineteenth-Century France." *Comparative Studies in Society and History*, 25(1): 26–49.

Dehdari, Sirus H., and Kai Gehring. forthcoming. "The Origins of Common Identity: Evidence from Alsace-Lorraine." *American Economic Journal: Applied Economics*.

Dell, Melissa, and Pablo Querubin. 2018. "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies." *The Quarterly Journal of Economics*, 133(2): 701–764.

Depetris-Chauvin, Emilio, Ruben Durante, and Filipe Campante. 2020. "Building nations through shared experiences: Evidence from African football." *American Economic Review*, 110(5): 1572–1602.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. 2012. "The political economy of linguistic cleavages." *Journal of Development Economics*, 97(2): 322–338.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Shlomo Weber. 2009. "Linguistic Diversity and Redistribution." *Journal of the European Economic Association*, 7(6): 1291–1318.

de Talleyrand-Périgord, Charles-Maurice. 1791. "Rapport sur l'instruction publique." *Comité de Constitution, Assemblée nationale.*

Dickens, Andrew. 2018. "Ethnolinguistic favoritism in african politics." *American Economic Journal: Applied Economics*, 10(3): 370–402.

Dickens, Andrew. 2021. "Understanding Ethnolinguistic Differences: The Roles of Geography and Trade." *The Economic Journal.*

Duflo, Esther. 2001. "Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment." *American Economic Review*, 91(4): 795–813.

Duflo, Esther. 2004. "The medium run effects of educational expansion: Evidence from a large school construction program in Indonesia." *Journal of Development Economics*, 74(1): 163–197.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World (24th edition).* Dallas, TX:SIL International.

Falck, Oliver, Stephan Heblich, Alfred Lameli, and Jens Südekum. 2012. "Dialects, cultural identity, and economic exchange." *Journal of urban economics*, 72(2-3): 225–239.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth*, 8(2): 195–222.

Fearon, James D, and David D Laitin. 1999. "Weak states, rough terrain, and large-scale ethnic violence since 1945." 2–5, Citeseer.

Fieller, E. C. 1954. "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2): 175–185.

Fouka, Vasiliki. 2020. "Backlash: The unintended effects of language prohibition in US schools after World War I." *The Review of Economic Studies*, 87(1): 204–239.

Franck, Raphaël, and Oded Galor. 2021. "Flowers of evil? Industrialization and long run development." *Journal of Monetary Economics*, 117: 108–128.

Furet, François, and Jacques Ozouf. 1977. *Lire et écrire : l'alphabétisation des Français de Calvin à Jules Ferry.* Paris:Les Editions de Minuit.

Galor, Oded, and David Weil. 2000. "Population, Technology, and Growth: From the Malthusian Regime to the Demographic Transition and Beyond." *American Economic Review*, 90(4): 806–828.

Galor, Oded, Ömer Özak, and Assaf Sarid. 2018. "Geographical roots of the coevolution of cultural and linguistic traits." *Available at SSRN 3284239.*

Gellner, Ernest. 1983*a*. "Nationalism and the two forms of cohesion in complex societies." British Academy London.

Gellner, Ernest. 1983*b*. *Nations and Nationalism*. Ithaca:Cornell University Press.

Gelman, Andrew, and Guido Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics*, 37(3): 447–456.

Gilliéron, Jules, and Edmond Edmont. 1902. *Atlas linguistique de la France: notice servant a l'intelligence des cartes*. Paris:Champion.

Gilliéron, Jules, and Edmond Edmont. 1902-1910. *Atlas linguistique de la France*. Vol. 1-9, Paris:Champion.

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

Goebl, Hans. 1984. *Dialektometrische Studien. Anhand italoromanischer, ratoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Bande. Beihefte zur Zeitschrift fur romanische Philologie*, Tubingen:Niemeyer.

Goebl, Hans. 2002. "Analyse dialectométrique des structures de profondeur de l'ALF." *Revue de linguistique romane*, 66: 261–262.

Goebl, Hans. 2003. "Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur." *Estudis Romanics*, 25: 59–120.

Goebl, Hans. 2006. "Recent Advances in Salzburg Dialectometry." *Literary and Linguistic Computing*, 21(4): 411–435.

Goebl, Hans. 2011. "22. Dialectometry and quantitative mapping." *Volume 2 Language Mapping: An International Handbook of Linguistic Variation*, , ed. Alfred Lameli, Roland Kehrein and Stefan Rabanus, 433–464. De Gruyter Mouton.

Goebl, Hans, Salzburg dialectometry team, Edgar Haimerl, Pavel Smečka, Bernhard Castellazzi, and Yves Scherrer. 2019. "Dialectometry ALF (www.dialektkarten.ch/dmviewer/alf/index.en.html), retrieved December 13, 2021."

Green, Elliott. 2013. "Explaining African ethnic diversity." *International Political Science Review*, 34(3): 235–253.

Grégoire, Henri. 1794. "Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française." *Convention nationale*.

Grembi, Veronica, Tommaso Nannicini, and Ugo Troiano. 2016. "Do Fiscal Rules Matter?" *American Economic Journal: Applied Economics*, 8(3): 1–30.

Guizot, François. 1833. "64. 2 janvier 1833. Exposé des motifs de la loi sur l'instruction primaire."

Guizot, François. 1860. *Mémoires pour servir à l'histoire de mon temps: La Révolution de 1830 (Tome 3)*. Paris:Michel Lévy Frères.

Guizot, François. 1863. *Histoire parlementaire de France: recueil complet des discours prononcés dans les Chambres de 1819 à 1848 (Vol. 2)*. Paris:Michel Lévy Frères.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69(1): 201–209.

Hirschberg, Joe, and Jenny Lye. 2010. "A Geometric Comparison of the Delta and Fieller Confidence Intervals." *The American Statistician*, 64(3): 234–241.

Hobsbawm, E. J. 1990. *Nations and Nationalism since 1780: Programme, Myth, Reality.* Cambridge:Cambridge University Press.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies*, 79(3): 933–959.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics*, 142(2): 615–635.

Johnson, Noel D. 2019. "Taxes, National Identity, and Nation Building: Evidence from France." Working paper.

Kessler, Brett. 1995. "Computational Dialectology in Irish Gaelic." *EACL '95*, 60–66. San Francisco, CA, USA:Morgan Kaufmann Publishers Inc.

Koepke, Robert L. 1995. "Cooperation, Not Conflict: Curés and Primary School Inspectors in July Monarchy France, 1833-1848." *Church History*, 64(4): 594–609.

Laitin, David D. 2000. "What Is a Language Community?" *American Journal of Political Science*, 44(1): 142–155.

Le Dû, Jean, Yves Le Berre, and Guylaine Brun-Trigaud. 2005. *Lectures de l'atlas linguistique de la France de Gilliéron et Edmont: du temps dans l'espace ; essai d'interprétation des cartes de L'Atlas linguistique de la France de Jules Gilliéron et Edmond Edmont.* Paris:Éditions du Comité des Travaux Historiques et Scientifiques.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355.

Lemieux, Thomas, and Kevin Milligan. 2008. "Incentive effects of social assistance: A regression discontinuity approach." *Journal of Econometrics*, 142(2): 807–828.

Lodge, Raymond Anthony. 1993. *French, from Dialect to Standard.* London:Routledge.

Lodge, Raymond Anthony. 2002. "« Francien » et « français de Paris »." *Linx*, 12: 149–172.

Lorain, Paul. 1837. *Tableau de l'instruction primaire en France: d'après des documents authentiques, et notamment d'après les rapports adressés au ministre de l'instruction publique par le 490 inspecteurs chargés de visiter toutes les écoles de France, à la fin de 1833.* Paris:Hachette.

Meyers, Peter V. 1976. "Professionalization and Societal Change: Rural Teachers in Nineteenth Century France." *Journal of Social History*, 9(4): 542–558.

Michalopoulos, Stelios. 2012. "The origins of ethnolinguistic diversity." *American Economic Review*, 102(4): 1508–39.

Montalbo, Adrien. 2020. "Education supply and economic growth in nineteenth-century France." Working paper.

Montalbo, Adrien. 2021. "Schools without a law: Primary education in France from the Revolution to the Guizot Law." *Explorations in Economic History*, 79: 101364.

Morrisson, Christian, and Wayne Snyder. 2000. "The income inequality of France in historical perspective." *European Review of Economic History*, 4(1): 59–83.

Moseley, Christopher. 2010. *Atlas of the World's Languages in Danger (3rd ed.).* Paris:UNESCO Publishing.

Nique, Christian. 1990. *Comment l'école devint une affaire d'Etat (1815-1840).* Paris:Nathan.

Perrot, Marie-Clémence. 1997. "La politique linguistique pendant la Révolution française." *Mots/Les langages du politique*, 52.

Polard, André. 2004. *Cet obscur sentiment qu'est la honte, in: L'épilepsie du sujet.* Paris:Harmattan.

Prost, Antoine. 1968. *Histoire de l'enseignement en France, 1800-1967.* Paris:Armand Colin.

Ravier, Joël. 1998. "L'imposition du français dans le département du Nord : de la Monarchie de Juillet au Second Empire." *Histoire de l'éducation*, 77.

Ravier, Joël. 2012. "L'évaluation des enseignants du primaire en France de 1835 à 1850. L'inspection primaire de Guizot et Falloux." *Spirale - Revue de recherches en éducation*, 49.

Reboul, Fabienne. 1991. *Guizot et l'instruction publique, in François Guizot et la culture politique de son temps, Colloque de la Fondation Guizot-Val Richer.* Paris:Seuil.

Rosanvallon, Pierre. 1985. *Le moment Guizot.* Paris:Gallimard.

Séguy, Jean. 1973. "La dialectométrie dans l'Atlas linguistique de la Gascogne." *Revue De Linguistique Romane*, , (1-24).

Site Mémoire des hommes. n.d.. "Portail culturel du Ministère des Armées, données téléchargées sur **www.memoiredeshommes.sga.defense.gouv.fr**, mise à jour du 21 juin 2021."

Smith, Anthony D, et al. 1991. *National identity.* Vol. 11, University of Nevada press Reno.

Spolaore, Enrico, and Romain Wacziarg. forthcoming. "Fertility and Modernity." *The Economic Journal.*

Squicciarini, Mara P. 2020. "Devotion and Development: Religiosity, Education, and Economic Progress in Nineteenth-Century France." *American Economic Review*, 110(11): 3454–91.

Squicciarini, Mara P., and Nico Voigtländer. 2015. "Human Capital and Industrialization: Evidence from the Age of Enlightenment." *The Quarterly Journal of Economics*, 130(4): 1825–1883.

Squicciarini, Mara P., and Nico Voigtländer. 2016. "Knowledge Elites and Modernization: Evidence from Revolutionary France." Working paper.

Statistique Générale de la France. n.d.*a. L'enseignement primaire de 1829 à 1897.* Atelier du Centre de recherches historiques, données traitées par l'INSEE, mis en ligne le 07 février 2011, consulté le 21 juin 2021.

Statistique Générale de la France. n.d.*b. Les recensements de 1851 à 1921.* Atelier du Centre de recherches historiques, données traitées par l'INSEE, mis en ligne le 06 février 2011, consulté le 21 juin 2021.

Swadesh, Morris. 1952. "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos." *Proceedings of the American philosophical society*, 96(4): 452–463.

Tackett, Timothy. 1986. *Religion, Revolution, and Regional Culture in Eighteenth-Century France: The Ecclesiastical Oath of 1791.* Princeton:Princeton University Press.

Ther, Philipp. 2014. *The Dark Side of Nation-States: Ethnic Cleansing in Modern Europe.* New York City:Berghahn Books.

Weber, Eugen. 1976. *Peasants into Frenchmen: The Modernization of Rural France, 1870–1914.* Stanford:Stanford University Press.

Weinreich, Max. 1945. "Der YIVO un di problemen fun undzer tsayt." *YIVO Bleter,* 25(1): 1–13.

The online appendix for this paper can be found **here**.