Precocious Enough to Rationalize Culture? Explaining the Success and Failure of Nation-building in Europe, 1400–2000

Yu Sasaki

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Anthony Gill, Chair

Edgar Kiser

Victor Menaldo

Steven Pfaff

Program Authorized to Offer Degree:
Department of Political Science

University of Washington

Abstract

Precocious Enough to Rationalize Culture?
Explaining the Success and Failure of Nation-building in Europe, 1400–2000

Yu Sasaki

Chair of the Supervisory Committee:
Professor Anthony Gill
Political Science

Why do some ethnic groups consolidate their cultural practices earlier than others? Extant scholarship in ethnicity, nations, and state-building hypothesizes that the state is the most important determinant. In my dissertation, I argue that it is not the only channel and there are other factors that matter. In three standalone essays, I investigate the role of (1) geography, (2) technology, and (3) public goods provision at the ethnic-group level. I provide a simple conceptual framework of how each of these determinants affects cultural consolidation for ethnic groups. I argue that geographical conditions and technology adoption can have a positive impact on ethnic groups' ability to develop unique cultural attributes without an independent state. Although they may be politically incorporated by stronger groups in the modern period, they still demand self-rule or standardize their vernacular. I also argue that, in contrast with the expectation from the political economy research on ethnicity, cultural consolidation does not always yield public goods provision at the ethnic-group level. I show that leadership in a representative organization plays a key role in making the connection between cultural consolidation and public goods provision. I document evidence by constructing two new data sets on 171 European ethnic groups that exist for 1400–2000 and presenting findings from statistical analysis. In one essay, I also use process-tracing methods to uncover a causal variable that is not captured in quantitative analysis. Findings from my histor-

ical analysis have implications for understanding why some ethnic groups can use their culture as a political instrument in contemporary politics more effectively than others. My research demonstrates that the past choices that ethnic groups made prove critical and that these choices were made hundreds of years ago.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I have accumulated intellectual debts over the course of my time at the University of Washington, Seattle. First and foremost, I would like to acknowledge my committee members: Tony Gill, Edgar Kiser, Victor Menaldo, and Steve Pfaff. I thank their unflagging support for my wading through the uncharted waters. My research is unique in that it is unconventional and ambitious in the standards of the literature, but they remained positive and encouraging in all phases of my research. I would also thank other colleagues at UW's Department of Political Science and other institutions: Jeff Arnold, Darryl Holman, Margaret Levi, James Long, Kyle Marquardt, Tristan James Mabry, Jared Rubin, and Susan Whiting. They taught me formally and informally, gave me opportunities for research and professional development, provided critical feedback to my work, and greatly contributed to my work. Finally, my deep thanks goes to my comrades-in-arms from my cohort, Aaron Erlich and Crystal Pryor. Our collaborative efforts on the exams, grant applications, and our respective research projects shaped my thinking, underpinned my work, and offered tangible and intangible support, all of which were priceless.

Chapter 1

INTRODUCTION

Why do some ethnic groups consolidate their language and other cultural attributes earlier than others? Why do some invest in their organizational and institutional capacity to consolidate culture more than others? Why do some minority ethnic groups resist incorporation by a majority group and demand autonomy more than others? What explains ethnic groups' provision of public goods more than others? These are some of the big puzzles in the study of ethnicity, nations, and state-building. In my dissertation I examine these puzzles in three empirical essays. Each essay offer a simple conceptual framework for addressing a big question and documents empirical evidence by using both quantitative and qualitative methods. My quantitative analysis draws on new data sets that I construct on 171 ethnic groups in Europe for the period between 1400 and 2000. I also employ process-tracing methods to uncover evidence that is hard to quantify. This introduction provides an overview of my dissertation and briefly discusses my conceptual and empirical contributions to the literature.

The overarching goal of my dissertation research is to establish long-term perspectives on the consolidation of cultural attributes that matter to ethnicity and nations and to explain its variation across groups. In my dissertation, I demonstrate that this analytical approach is useful for systematically understanding the big puzzles and important outcomes in contemporary politics, such as why some ethnic groups are able to use their culture as political instruments more effectively than others. Addressing this specific puzzle, for instance, has implications for public policy on linguistic diversity and multilingualism at the state level (Laitin and Ramachandran 2016; Albaugh 2015;

Cardinal and Sonntag 2015; Liu 2014; Ginsburgh and Weber 2011; Safran and Laponce 2005) and multiculturalism which involves which culture of the demographic minority should be officially sanctioned or allowed in education, street signs, and public administration (Kymlicka 2009). Extant scholarship offers a few answers. First, the state plays the most important role. Ethnic groups that comprise the state, such as the English, Spanish, Germans, and Russians, have the disproportionate ability to shape the policies that represent their cultural attributes in the language of school instruction, the "official" interpretations of the country's past, social norms, and other issues. These groups constitute the nation-states, and the literature on nations and nationalism typically undertake research using the nation-state as the unit of analysis (Breuilly 1993; Gellner 2006; Hechter 2000; Roeder 2007; Wimmer 2013). This research shows that institutional capacity based on independent statehood is probably the single most important variable explaining the variation in ethnic groups' ability to use their culture as a tool in politics.

A second hypothesis to address those puzzles focuses on the means of communications such as printed material, census, map, museum, and on symbols that uniquely appeal to specific groups, such as ancient artifacts and landscapes. Communications technologies are crucial as they are liable to copying; as Benedict Anderson (2006, ch. 5) points out, technologies like the movable type or the census are subject to easily copied—or "pirated" to use Anderson's word—once they are proven to be useful for leaving the concrete record of cultural practices in one ethnic group. Similarly, ethnic groups that have cultural relics and human remains excavated in what they perceive to be their homeland can use these objects as "proofs" of continuity from ancient times as distinct communities that existed prior to the age of the nation-state (Smith 1986). One observable implication is that although these ethnic groups with an access to these communications tools and ancient symbols may not have their own state, they are able to use culture as political tools by highlighting differences with a majority group and demanding autonomy (Hansen 2015; Sahlins 1989; Thongchai 1994).

These hypotheses from extant scholarship yield generalizable insights in terms of a set of capacities that ethnic groups ideally possess. The state, for instance, denotes the capacity of an eth-

nic group to *mobilize* its fellow members or to *enforce* formal rules including not just taxation but also conscription, citizenship criteria, and schooling and informal rules about behavioral norms (Brubaker 1992; Colley 2005; Hirsch 2005; Martin 2001; Sahlins 2004; Weber 1976; Weil 2008). By contrast, the thesis on ancient symbols refers to the capacity to *appeal* broadly. In politics, symbols can serve as "information shortcuts" that signal messages about political goals and preferences of the sender, which are contained in easy-to-carry means like flags, drawings, maps, and books.[1] Finally, communications technologies give ethnic groups a capacity to *represent* them in a way that is distinguishable from others and *disseminate* this information. The ability of these tools to replicate on a mass scale vernacular books and maps on which an ethnic group's homeland is depicted can help construct a shared identity and mobilize the group for a political cause. These capacities are not mutually exclusive, and some ethnic groups have more than others.

These capacities are what I call "proximate" causes of cultural consolidation in ethnicity and nationality. They are proximate in the sense that political and economic outcomes for ethnic groups are endogenous to them. Through empirical analysis, I show in my dissertation that the focus on proximate causes leaves three crucial empirical questions not widely investigated. The first comes from a simple observation that not all ethnic groups are equal in terms of the proximate causes: only a subset of ethnic groups in the world comprise independent states, and the degree of organizational capacity varies widely in the rest. For example, some groups in the latter category like the Scots are developed enough to have their own parliament that endows them with a taxation authority, while the Bretons, a Celtic-speaking minority in France, won the right to use bilingual road signs only in the late 1980s. How does the variation in proximate causes explain the big puzzles mentioned above, such as ethnic groups' ability to use culture in politics? There is a growing body of quantitative research (cf. Wimmer and Feinstein 2010; Wimmer 2015), but the scope of analysis has not yet included those groups that lacked proximate determinants.

The second empirical question that has not been well-researched is the varying effect of proxi-

[1] Contemporary elections typically use bumper-stickers, television advertisements, and social-networking services.

mate causes. Acquiring enforcement capacity or capacity to mass-appeal does not have a uniform impact; it is an empirical question as to whether these proximate causes function as intended. In my second dissertation essay, for instance, I use the acquisition of the printing press and examine its impact on language standardization, a crucial dimension in cultural consolidation for ethnic groups. I document statistical evidence that conditional on getting print technology, not all ethnic groups in Europe codify their vernacular. I also show that the *timing* of access matters—early adopters of the press have greater chances of standardization than latecomers. The variation of effects may depend not just on *when* but also *what* communications tool or symbol is in question. Since a variety of tools and symbols can determine the political salience of ethnic and national identity, it seems that there is much room for empirical contributions in this area.

Finally, the third empirical question that has been untapped is the *magnitude* of determinants. There is little research that investigates whether an ancient symbol or a census would alone promote cultural consolidation or ethnic mobilization or whether its impact depends on some other conditioning variables like the enforcement capacity of the state.[2] In Xinjiang, the northwestern region of China and the homeland of the Turkic-speaking Uighurs who are a Muslim minority population, the 1980 excavation of ancient mummies that dated as far back as 3,800 years made the Uighur identity salient in that Uighurs saw them as "scientific" evidence that they belong to a different group from the Chinese (Wong 2008). But its impact was not strong enough to mobilize Uighurs for greater autonomy. Although the Uighurs have been nominally granted the *de jure* self-governing authority in their Xinjiang Uighur Autonomous Region, the *de facto* power lies in the hands of Beijing. Similarly, Eric Hobsbawm (1990) highlights the importance of "government perspective" to understand why some cultural attributes become politically salient (1990, ch. 3). Yet this hypothesis brings attention back to that of the presence or absence of enforcement capacities at the state level, which leaves the questions about proximate determinants' magnitude unanswered.

---

[2]One strand of the economic-history literature, which seeks the "deep" historical determinants of modern variation in income and inequalities across countries, offers a similar debate about whether geographical variables have a direct or indirect impact on the variation of contemporary economic growth. See Spolaore and Wacziarg (2013) for review.

To fill these gaps, I investigate fundamental determinants of why some ethnic groups consolidate their earlier than others. In three standalone empirical essays. I seek to make two broad contributions. The first is to bring innovations in conceptual framework in the literature on ethnicity and nations. I show that there are routes other than the state that matter in addressing this question. More specifically, I focus on (1) geography, (2) technology, and (3) public goods provision at the ethnic-group level, draw a testable hypothesis, and examine it using both quantitative and qualitative methods. My essays document evidence that these variables systematically determine ethnic groups' ability to deploy their culture as political tools and whether they consolidate their cultural practices. In addition, my research suggests that the choices that these ethnic groups made hundreds of years ago still matter and shape their ability to use culture in politics effectively in contemporary politics.

My second contribution is empirical. My innovation is to use the ethnic group as the unit of analysis. To gain purchase on empirical variation, I identified as many observations as possible, not just those groups that today comprise the state but also those that do not. In extant scholarship, the assumption of constructivism holds that terms such as "ethnicity," "nations," and "ethnic or national identity" are *constructed*, in that groups' cultural attributes and identity evolve over time and their political salience varies conditional on institutional and other circumstances.[3] However, the literature also shows that ethnic groups, once settled in a certain locale, call their place of residence "homeland" and rarely move, even though some homeland locations are relatively inhospitable, such as a remote island. I take advantage of this fact to identify ethnic groups' homeland and systematically collect information on the homeland's political, economic, social, geographical, and other dimensions.

In my first two essays I assemble two new data sets on 171 existing European ethnic groups for 1400–2000 by drawing on various social-science fields, including political science, sociology, cultural and economic history, geography, linguistics, and anthropology. These data allow me to

[3]See Anderson (2006) and Gellner (2006) for prominent works on constructivism. On the political salience of ethnic or national identity, see, for instance, Chandra (2004), Kasfir (1979), Laitin (1998), and **?**.

provide statistical evidence on the impact of geography and technology on the degree of autonomy and language standardization, respectively. In the third essay, I use process-tracing methods to document evidence on the failure of public goods provision among the Occitans, a minority ethnic group in France. This study shows that Occitans did not invest in public goods despite the fact that they had developed the codified tongue. Therefore, it constitutes a counterintuitive or "near-miss" case of public goods provision. Combining the conceptual reorientation with new evidence, I argue for a comparative historical turn in the study of ethnicity, nations, and states that explore long-run determinants of contemporary ethnic groups' political and economic outcomes.

It is important to address the limitations of my data. My data sets only includes *existing* ethnic groups in Europe for the period of 1400–2000.[4] My choice is based on two reasons. First, there is virtually no data on the "extinct" or "dead" ethnic groups.[5] There are few investigations that systematically consider existing ethnic groups, and to my best knowledge, my dissertation is one of the first to do so.

Second, when the analyst bases ethnicity solely on language, there is little research as to why historical ethnic groups went extinct.[6] The case of the Kashubians is instructive.[7] They are an existing ethnic minority living in the coastal region of northern Poland, where the area is historically called Pomerania. It is named after the Pomeranians, a group of tribes speaking West Slavic languages. The Pomeranians later branched into four groups: one is the Kashubians, which are alive with an estimated 50,000 speakers, another is the Slovencians, which are now extinct, and the two other unindentified groups, which are assumed to have become extinct as well. Little is known about the

---

[4] As I discussed in my empirical essays in more detail, I primarily use Minahan (2000) and Lewis, Simons, and Fennig (2013) for sources of ethnic groups in Europe.

[5] There is a body of research on language deaths, and one can use it to examine the extinction of ethnic groups if one is to equate language groups with ethnic groups (or ethnolinguistic groups). For data, see, for instance, the World Language Mapping System database that includes some extinct languages.

[6] For the analysis of the decline of contemporary linguistic diversity, see Austin and Sallabank (2011) for overview and Amano et al. (2014) for a recent analysis.

[7] This anecdote's sources come from Minahan (2000, 375–77) and the Kashubian entry in Lewis, Simons, and Fennig (2013).

causes of extinctions, although one can infer this fate from the fact that neighboring great powers, namely the Germans and Poles, frequently overran this area for war. This illustration suggests that the extinction of these groups might be partly attributed to war, although data are scarce and the ultimate causes would remain indeterminate. It would be ideal to include dead ethnic groups to specify a "denominator" for quantitative analysis, but identifying them is beyond the scope of this dissertation. For these reasons, I focus on existing ethnic groups whose information is richer and can be collected more consistently.

The exclusion of extinct ethnic groups from analysis is unlikely to affect my empirical findings. First, the number of known dead groups is few. For instance, the World Language Mapping System database, which offers the geographical distribution of existing language groups across the world, identifies about a dozen of them in Europe. Even when attributes of these groups are uncovered and added to my data set of 171 European ethnic groups, this addition will constitute approximately 7 percent of an expanded data set. I do not expect it to change my findings substantially. Second, the causes of now-extinct ethnic groups seem independent of those I employ as hypotheses in my analysis. For instance, in one essay I hypothesize that the printing press is positively associated with language standardization. I suggest that the acquisition of print technology bear no impact on extinction, so my hypothesis should not be affected by the groups that are omitted from my data set. Third and finally, even when it is plausible that the inclusion of extinct groups could reject my hypotheses, my statistical analysis controls for the variables that are likely to determine the life and death of ethnic groups, such as war frequency and economic growth. I also use the exogenous variation of my main causal variables as instrumental variables to assess the robustness of my claim.

My dissertation is composed of three standalone empirical essays. The first essay, "Ethnic Autonomy" (Chapter 2), investigates the impact of geography on ethnic groups' demand for autonomy in France. There are broadly two approaches—*direct* and *indirect* effects of geography—to constructing hypotheses. One direct and positive impact is that geographical isolation, caused by oceans, terrain ruggedness, and high elevation above the sea level, keeps physical access and communica-

tions to the outside world high. Geography could serve as a buffer and keep the costs of invasion by militarily stronger communities prohibitively high, even when these isolated communities did not develop strong political institutions. Isolation can also aid the development of a unique language or a set of cultural code and norms that are distinct from others. When these geographically isolated ethnic groups are subject to a neighbor's political incorporation in modern times, they may resist centralized rule and demand self-governance.

By contrast, geography can directly and adversely affect the consolidation of culture. In these places, transportation, infrastructure, and communications technologies are expensive or hard to obtain, leaving interactions difficult not just within group but also with neighbors. As Fredrik Barth (1969) points out, regular communications with other communities play a crucial role in enabling one's community to perceive differences and strengthen the within-group identity. High costs of communications reduce the incentive to invest in consolidating cultural attributes for the whole group, which would leave the community's unstandardized culture vulnerable, particularly when it is politically incorporated by a large state. Alternatively, geography can work indirectly, most likely through institutions. Isolation is likely to make the scaling of economic growth or political centralization difficult. In this circumstance, technology adoption or investment in standardizing culture would be deemed unnecessary or costly due to the short of demand from limited growth or underdeveloped institutions.

In my first essay, I provide a simple conceptual framework about the impact of geography on the development of cultural practices and attributes among minority ethnic groups in France between 1400 and 1900. I hypothesize that the more distant an ethnic group is from Paris, the center of political authority in France, the more likely that unique cultural practices are to emerge. I also hypothesize that distant groups are expected to resist the French policy of cultural unification, which is part of the modernization process under the Third Republic that began in 1870. I test my hypotheses by drawing on a new data set of 214 French cities for eight ethnic groups, nested within province-level attributes, for 1400–1900. Findings from the OLS estimation support my hypotheses

and they are robust to the inclusion of a battery of controls and alternative hypotheses. I conduct further robustness checks using the onset of the Neolithic Revolution as an instrumental variable.

The second essay (Chapter 3: "Publishing Nations") considers the impact of technology adoption on the development of vernaculars among European ethnic groups.[8] More specifically, I explore the relationship between the acquisition of the printing press and vernacular codification. The literature on economic history shows that technology is knowledge and printing technology dramatically reduced the cost of access to knowledge in Europe (Bernstein 2013; Dittmar 2011; Mokyr 2006). Joel Mokyr (2006) calls tools like the printing press "propositional knowledge," the type of knowledge with which to define, describe, and characterize human activity more concretely and cheaply. It is a medium to produce innovations or new "goods and services."[9] With respect to cultural attributes, print gives language consistent spelling as it transformed spelling from phonetics-based to written-based. Although the technology substantially cut man-hours to copy manuscripts, casting metal types of letters remained quite laborious (Bernstein 2013, ch. 5). When one caster set a type of words with a new spelling, others would use it to reduce their workload of printing. In addition, print technology plays an important role in leaving the written record of cultural attributes and practices for ethnic groups. This is crucial especially for those ethnic groups that lacked the resources or state institutions. When scholars and intellectuals of the nineteenth century wanted to construct a distinct ethnic or national identity for their minority community, they often turned to the old writings to revive and develop their vernacular by writing novels, poems, and history. The significance of this analytical approach is that the state is not necessarily the main actor to standardize language; the printing press in Europe spread primarily through skilled workers (Cipolla 1972). This chapter, therefore, provides an alternative channel through which the consolidation of culture can take place.

---

[8] A version of this essay is to be published in *The Journal of Economic History* in the Spring of 2018.

[9] In later periods, newspapers and the radio had similar effects of sharing information more effectively and helping to instigate mass mobilization for political causes. See, for instance, Bernstein (2013, ch. 7) and Yanagizawa-Drott (2014).

In my second essay, I make two hypotheses about the the impact of the printing press on language standardization as the outcome. First, I hypothesize that the acquisition of print is positively associated with language standardization. As the technology reduced the cost of access to information in the vernacular, the cost of becoming literate also declined and innovations in language became easier. My second hypothesis is that the timing of print adoption matters. This is not just because vernacular codification took time but also because it was not a priority for subsistence or political survival in medieval through early-modern Europe. Early adopters had a chance to leave the written record of their language and to start a long process of codification. To test these hypotheses, I construct a new data set on 171 European ethnic groups for 1400–2000. Findings from Cox proportional hazards models support my hypotheses. The statistical results are consistent when using logistic regression models and robust to the inclusion of a host of covariates, including war frequency and economic growth, and to alternative hypotheses, including the human capital channel. I conduct further robustness checks by exploiting the exogenous variation of the geographical distance to Mainz, the birthplace of the Gutenberg press, as an instrumental variable for the diffusion of the printing press.

The third and final essay (Chapter 4: "Does Cultural Consolidation Lead to Public Goods Provision?") explores the "failure" of public goods provision after cultural consolidation. I have two rationales for this research. First, the stylized fact in the political economy literature is that there is a negative relationship between ethnic or cultural diversity in society and the degree of public goods provision (Alesina, Baqir, and Easterly 1999; Habyarimana et al. 2007, 2009; Miguel 2004; Miguel and Gugerty 2005). It implies that within-group homogeneity should be positively associated with public goods. In a country where multiple ethnic groups compete for power, the strong provision of public goods would give an ethnic group advantage over others in terms of political mobilization. This happens especially in developing countries where the resources are highly scarce. Second, the literature on ethnicity and nations tends to leave out "negative" cases in favor of "positive" ones. Here, *positive* refers to the cases in which ethnic groups successfully standardize their vernacular,

build strong institutions, and provide public goods, while these realizations do not occur in *negative* cases. One implication from this scholarship is that success in one area goes hand in hand with another area such that the successful standardization of language, for instance, should lead to the provision of public goods.

In my third essay, I provide empirical evidence for these questions by exploring on the case of the Occitans in France. It is an ideal case for two reasons. First, the Occitans established an active literary society in the mid-nineteenth century and successfully standardized their vernacular in 1875. My data set used in Chapter 3 indicates that only 27 out of 171 European ethnic groups were able to standardize their vernacular by 1900. The Occitans are certainly early on this dimension. The literary society sought to gain popularity in their activities by advocating vernacular instruction at the university level and producing widely-read publications, but they remained hesitant to mobilize the lay Occitans for further political causes such as autonomy or federation. Second, this essays demonstrates that success in one area (i.e., language standardization) does not necessarily lead to success in another (i.e., public goods provision). Instead of mass mobilization, the leadership of the literary society seemed content with language preservation as the goal by encouraging lay Occitans to use Occitan. This case study, therefore, constitutes a "near miss" case—the failed provision of public goods despite early vernacular codification. I use matching methods to select cases and process-tracing methods to uncover the causes of the near-miss outcome.

In short, my dissertation essays not just demonstrate non-state routes to explaining the variation in cultural consolidation among European ethnic groups but also address that variation in ethnic groups' ability to use their cultural attributes as tools of politics today may be determined hundreds of years ago. My analysis yields implications for other outcomes in contemporary politics, such as variation in public goods provision, economic growth, and redistribution.

Chapter 2

ETHNIC AUTONOMY

2.1 Introduction

The year 1300 was a time when Europe was territorially most fragmented. It had 238 political entities that include sovereign states, dependent states, and autonomous areas, the highest number between 0 and 2000 C.E. (Nüssli 2011). At the turn of the twentieth century, this number dropped to just 24 independent and five non-sovereign states, a 90-percent decline. What explains this trajectory from fragmentation to consolidation of political authority?

Two strands of scholarship in the social sciences have addressed this question. First, political scientists and sociologists examine how states acquire territorial sovereignty through developing hierarchical organizations, particularly modern bureaucracy, in the process of the centralization of authority (Tilly 1975, 1992; Weber 1978). The literature on state-building focuses on the determinants of the establishment of sovereignty as the ultimate form of political authority and of the variation of this process across time and space (Abramson 2017; Ertman 1997; North 1981; Spruyt 1994, 2009). Second, historians and area specialists highlight the social-cultural dimension of consolidation by studying the rise and spread of nationalism in the modern period. Building on the state formation literature, these scholars explain how centralizing states achieve greater cultural unity through industrialization (Gellner 2006), literacy and capitalism (Anderson 2006), conscription (Colley 2005), and citizenship (Brubaker 1992; Wimmer 2002). Variation in the degree or timing of cultural rationalization depends not just on state capacity but also on how politicians build institutions that define and enforce the "standard" cultural practices (Hobsbawm 1990; Hobsbawm and Ranger 1983). These two strands of literature account for an "ideal" form of the development of the nation-state; they suggest that lack of access to sovereign territorial states lead to the disappearance

from the map that displays the process from fragmentation to consolidation.

However, this research often misses an "intermediate" form between territorial sovereignty and lack thereof, political autonomy. In the literature on the state, autonomy is considered an ineffective form of political authority since historically it was not as adept at building bureaucracy, as shown in city-states and federative alliances (Nexon 2009; Spruyt 1994; Stasavage 2014). In contemporary politics, competition over sovereignty becomes rarer yet the race for autonomy seems to get more intense. The main actor in this paper is substate entities, especially ethnic groups. The Scottish referendum of September 2014 was a prominent example, where the Scots voted on the final stage of self-rule, independence. It is not hard to find similar cases of ethnic groups' demand for greater autonomy in Europe and beyond, from the Catalans in Spain and the Albanians in Kosovo to the Pashtuns in Afghanistan and the Moros in the Philippines. Some have obtained the right to teach in their vernacular within the designated area of residence, while others are conferred limited degrees of legislative activity.

In this paper, I argue that a focus on ethnic groups is crucial to understand the development of autonomy as a form of political authority. First, I describe a broad theoretical framework that the historical trajectories of autonomy may be quite distinct from those of state sovereignty. In the case of Europe, the state was not strong enough to administer laws or implement policy throughout its domain until around the nineteenth century. When authority from the state "center" was loosely enforced, local autonomy in the "periphery" was likely to thrive and become predominant, leading to the different patterns of governance than in the center. When the capacity of the ruling state grew to enforce rule more consistently, autonomous regions in the periphery may resist it and insist on continued self-rule. Prolonged autonomy in the state periphery, I suggest, gives an opportunity to form a distinct group category in terms of "ethnicity." I call this specific type of political authority "ethnic autonomy." Second, I offer a theoretical framework centered on geographical scale that captures the institutional development of autonomy and draw testable implications. Third, I examine this argument based on a sample of 214 cities in France over the period of 1400–1900 C.E. Using

the proportion of non-French-language communes and population as proxies for ethnic autonomy, I test my hypothesis with ordinary least squares (OLS) estimators. Findings show that one geography scale, the distance from Paris, is positively associated with ethnic self-rule. At the mean level of distance (373 kilometers), the proportion of users in non-French tongues is expected to increase by 0.3 percentage points—and by approximately 0.56 percentage points in the countryside where historical ethnic groups are found. Analysis is robust to inclusion of a host of covariates that are potentially correlated to autonomy.

This paper makes two contributions. First, it aims to build a theoretical foundation toward political autonomy. By staying away from the preoccupation with state sovereignty, it is now possible to address another form of political authority. When it comes to explaining contemporary autonomy and independence movements by ethnic groups, scholarship tends to look at recent developments of state institutions or relative deprivation and discuss whether demands for self-rule can be understood as organically "constructed" or a "rational choice" of ambitious leaders (Brubaker and Laitin 1998; Fearon and Laitin 2000). The theoretical framework here rests on a simple yet heretofore under-examined argument that geographical conditions put constraints on rulers who would reach out to the countryside and maximize state revenue and that geographical scale shapes institutional development for stateless ethnic groups in the countryside. Under this framework, premodern states had limited capacity and were quite apart from the "ideal type" as defined by Weber (1978) and were unable to demarcate the boundaries territorially and project power far afield from the capital. Thus it is not surprising to see the demands for self-rule come from the peripheral regions of sovereign territorial states.

Second, my empirical findings help identify the variation of demand for ethnic autonomy. In extant scholarship, it is not clear why some ethnic groups develop enough to be able to seek independence while others are limited to holding on to the right to use their vernacular in school instruction. This paper suggests that the geographical distance to the capital of the ruling state affects the ability of peripheral communities to cultivate institutions distinct from those of the state

center. Long distance to the center means that minority ethnic groups are less likely to interact with state agents who demand taxation or conscription on the regular basis. De facto independence allowed these communities to develop distinct institutional practices that differed from those of the center. These groups may be prone to demanding greater degrees of self-rule when they fail to obtain their own state and their owner states begin to impose authority. By contrast, proximity to the center makes minority groups more susceptible to influence by the ruling state, which reduces the chances of consolidating unique cultural practices. These places are predicted to see less demands for autonomy.

My findings complement the literature on the contemporary independence or separatist movements by minority ethnic groups. It has generated various hypotheses, including relative deprivation (Gurr 1971), the power relations set up under colonialism (Horowitz 1985), preexisting institutions (Brubaker 1996; Roeder 2007), weak state capacity (Wimmer 2002), and unassimilable cultural attributes (Gellner 2006). These works seek to identify the causes by focusing on the conditions at the time of the mobilization. Yet how such conditions form in the first place is not always theorized and, more specifically, it remains unclear why some ethnic groups have the capacity to engage an independence movement while others stay dormant[1]. Evidence in this paper provides implications that geographical scale plays an important role in shaping the conditions in the long-run. In particular, it suggests that it is important to take into account the historical development of the spatial dimension in the political relationship between peripheral ethnic groups and the ruling state in the center. Empirical findings also help explain the variation of the occurrence of ethnic mobilization.

This paper is organized as follows. The next section briefly reviews the relevant literature. The third section lays out a theoretical framework on political autonomy for ethnic groups and describes theoretical frameworks. Then I discuss my empirical strategy for testing hypotheses. The fifth section documents main findings from statistical analysis. The paper concludes with a discussion about future research.

[1] Fearon and Laitin (2003) offers statistical evidence on the variation of ethnic insurgency and civil war for the second half of the twentieth century.

2.2 Literature

The imperative of ethnic autonomy as an analytical category of political authority rests on the realization that the centralized, bureaucratized state was not the only relevant and viable mode of political organization in the European past. Although this type proved to be predominant, as I will show later, how other types of organizations acquired and retained political autonomy is not necessarily the same as states did.[2] In this section, I will briefly discuss two theories of political authority closely related to ethnic groups: the homeland and sovereignty. The concept of ethnic autonomy is not analytically separate from these theories; rather, it is useful for encompassing a larger subset of political organizations.

2.2.1 The Homeland

Research on the homeland is closer of the two theories of political authority to ethnic autonomy, as it stresses "ethnic" dimensions in its theoretical construction.[3] The homeland is defined as a territory which an ethnic group believes to be its own and over which it claims an exclusive ownership for historical residence and the birthplace of its group identity (Smith 1986, 93, White 2000, 25, Yiftachal 2001). It has intrinsic values to ethnic groups, the literature puts forward, because one group's possession of land is central to its group existence so that it is perceived to be indivisible and unsharable with others (Smith 1986, 28, 29). This attribute about territory as a primary source of group identity is theorized to generate incentive for ethnic groups to engage in protracted and often violent conflicts (Connor 1994; Ross 2007; Shelef 2016). According to this literature, whether ethnic groups can secure and command political authority depends on the ability to hold on to exclusive control over the claimed territory.

What accounts for the varying degrees of political authority that different ethnic groups possess?

---

[2] See, for instance, Stasavage (2014), Tilly and Blockmans (1994).

[3] Section 2.3.2 of this article provides a definition of *ethnicity* and *ethnic groups* and a discussion of the development of ethnicity during the early-modern period.

No systematic data exists today to answer this question (Shelef 2016). Yet the literature offers two hypotheses to explain demand for exclusive control over territory—"the homeland." The first is the state. Existing works model on historical ethnic groups that had already possessed an independent state or would eventually obtain one (Sahlins 1989; Thongchai 1994; Zerubavel 1995). The ability to administer rule and exercise authority is, in turn, conditional on the extent of bureaucratization for the given state. This hypotheses suggests that only "strong" states with a well-developed bureaucracy are able to enjoy exclusive authority over a certain territory and establish a homeland. The second hypothesis is access to the cartographic technology based on the method of triangulation.[4] The technology provides users with the capacity to organize space into clearly demarcated units and to visualize it in a concrete fashion. It played a crucial role in enabling the transformation of political authority from based on the subjects to territory (Sahlins 1989).

These hypotheses have two limitations. One is that the capacity to achieve political authority in the name of a "homeland" is a function of having an independent state. Although references are made for ethnic groups, the literature uses the state as the unit of analysis and draws evidence from ethnic groups that already possessed a state. Likewise, access to the cartographic technology depends on independent statehood. One implication is that only under a strong state could cartographic surveys or ethnographic studies be carried out to delimit territory and establish authority.[5] The second limitation is that the "homeland" is observationally equivalent with sovereignty. Discussion about how actors acquire the capacity to command authority over a given territory or demand the right to do so internally and externally is built on with theories of sovereignty, as shown in the following section.

---

[4]See Lindgren (2007); Woodward (2007) for the development of the cartographic technology in early-modern Europe.

[5]There is large specialized literature on this theme. See, for instance, Biggs (1999), Hale (2007), Kagan and Schmidt (2007), Murphy (2002), Steinberg (2009). On the state's role in conducting population surveys and ethnographic studies, see Hansen (2015), Hirsch (2005), Martin (2001), Scott (1998).

2.2.2 Sovereignty

The scholarship on sovereignty lays out the centralization of political authority. It uses the state as the unit of analysis and focuses on territory-based authority as a chief outcome of interest. The convention in the international relations field holds that territorial sovereignty, which determines the exclusive right to and extent of rule based on states' territorial possession, began with the Peace of Westphalia of 1648. The argument is that through recurring and devastating religious wars, European states acquired the capacity to set the religion of a society on its own ("external sovereignty") and to command authority over matters within the state ("internal sovereignty"). In this account, the Peace constituted a sharp discontinuity from the past practice, because it established the territorialization of rule, where the domain of state authority would be defined by territory. Recent works in International Relations provide theoretical and empirical challenges to this model.[6] Yet as in the homeland, either side of the debate presupposes that sovereignty would arise as a function of having an independent, "strong" state; only with the well-developed ability to survive in a war-prone region did political organizations acquire this type of authority.[7]

The sovereign territorial state model has implications for how ethnic groups obtain political authority. In the study of ethnicity and nationalism, the *direct rule hypothesis* is derived from the literature on territorial sovereignty and state formation. It is a contractarian model based on the political history of nineteenth-century Europe, in which rulers agreed to institute state-wide citizenship, a policy of unitary language in state administration, and public schooling in exchange for their subjects' willingness to pay greater taxes and service for the state.[8] This model intends to

---

[6]For representative works, see Branch (2014), Krasner (1999), Krasner (2001), Nexon (2009), Osiander (2001a,b). For the competing view, see Philpott (2001).

[7]Branch (2014) argues that access to the grid-based cartographic technology allowed states to substantiate territorial authority was a key variable that explained the emergence of territorial sovereignty. It became a useful tool of statecraft after Westphalia in the late seventeenth through the mid eighteenth centuries. He maintains that only toward the mid-nineteenth century did territorial sovereignty become solidified in Europe. However, like the case of Thailand, the argument still assumes that access to this technology depends on having an independent state.

[8]Hechter (2000), Hobsbawm (1990), Tilly (1992), Tilly (1994), Wimmer (2002).

account for how one ethnic group comes to be predominant within a state and how its cultural practices are standardized as a "nation."[9] This theory, again, assumes that states would be strong enough to eliminate the political autonomy of other ethnic groups and assert authority over their subjects.

A brief overview of the homeland and sovereignty as models of political authority for ethnic groups shows that whether ethnic groups attain political authority is a function of having a state in the first place. However, these theories rest on the assumption of the Weberian ideal state that refers to a small subset of ethnic groups with a state, while political authority outcomes for a *larger* subset of ethnic groups without a state are omitted from analysis. Theoretical implications from empirical analysis of a small subset tends to be limited, in that they are not equipped to accounting for why ethnic groups with an independent state remain unable to achieve cultural homogeneity or why a greater number of ethnic groups remain stateless. Today, many if not most sovereign territorial states have a sizable "minority" ethnic group that calls for varying degrees of autonomy, ranging from the right to exercise cultural practices like language use to full-fledged political independence. How do such minority groups remain resistant to political and cultural integration by the state? There is a theoretical gap as to why some ethnic groups enjoy a greater degree of autonomy or what accounts for the variation of demands ethnic groups make for political authority.

2.3 Theory

2.3.1 Political authority in premodern Europe

Following Weber, I define political authority as the ability of an organization (typically a state) to command rule and obtain compliance (Weber 1978, 212). The path in which states achieved sovereignty by concentrating authority in greater breadth and scope through centralization and bureaucratization has been well trodden in the literature of state formation, especially with respect to Europe.[10] Although these theories are correct to point out that the sovereign territorial state even-

[9] Hechter (2000, 14) argues that possession of territory distinguishes between nations and ethnic groups.

[10] See Tilly (1975, 1992), Nexon (2009, ch. 3).

tually outperformed other types of political organizations, how authority earned support within the state, other than through coercion, is not as well understood. Bureaucratization does not necessarily mean compliance, because otherwise it is difficult to understand why substate entities like ethnic groups could demand self-rule. In this section, I sketch the broad conditions under which autonomy as a form of political authority can arise in the European context and how it becomes a basis for forming ethnicity. In organizing this theory, I refer to the state that obtained sovereignty as the "center" and substate communities that failed to do so as the "periphery" for the purpose of analytical simplicity.

The predominant system of political authority in premodern Europe was one of "composite," where authority was fragmented and decentralized throughout the Middle Ages and early modern times (Elliott 1992; Koenigsberger 1987; Pounds 1951). When, for instance, strong states absorbed weaker ones in a war or diplomatic settlement, they did not necessarily install an authority structure on the acquired lands; instead, incorporated areas were left with their own devices whose jurisdictional "limits" remained largely under-specified (Febvre 1973). In this environment, two features merit attention. First, there were overlapping authorities. In the Middle Ages, it was not uncommon to observe that the same feudal lords paid obligations to multiple kings of multiple states at the same time (Pounds 1951, 150; Febvre 1973, 213). Second, the jurisdiction of authority was determined based not on territory but on the subjects, using such units as fiefs, bishoprics, towns, and villages (Sahlins 1989, 6, 28; Kagan and Schmidt 2007, 662). These conditions suggest that there regularly were difficulties implementing policy and enforcing rule throughout the country.

2.3.2 Ethnicity formation

Under the condition of composite states, how does ethnicity form? Similarly, how does ethnicity develop autonomy? In this section, I provide a simple theoretical framework to understand the development of ethnicity through early modern times. In the following section, I will discuss how geography can affect the formation of autonomy for ethnic groups within states.

I define ethnicity as a category that delimits membership to a group and an ethnic group as a collectivity determined by a set of attributes that its members share. Following many scholars, I use the myth of common descent as the most essential attribute for my definition.[11] By contrast, I define the nation as a collectivity with *standardized* cultural attributes that command loyalty from members of its community. Standardization implies institutionalization, because a nation is required to have a capacity to promote and enforce certain cultural practices as "official" to ensure continuity from one generation to the next. Ethnicity and nation are distinct in terms of the degree of standardization.

My definition of ethnicity is appropriate for understanding ethnolinguistic diversity in early-modern Europe. Given that authority was defined in terms of subjects rather than territory under composite states, it is not realistic to hypothesize the existence of a unitary ethnic identity that was clearly delimited by state or provincial boundaries. Group identity was rather more localized; even within a nominally single group, identification was fragmented so that village- or parish-level affiliation seemed more salient than an identity that cuts across these units. The French countryside presents evidence. Loyalty at the local level was so persistent even after the Revolution that locals gave strangers a suspicious look and was uncooperative of requests from government officials (Weber 1976, 44–5). Thus, it is reasonable to think that the labels of collective identity such as "French," "Bretons," or "Occitans" may have commonality in language and ways of life but do not necessarily denote the homogeneity of identity among constituting members and that the establishment of a collective identity evolved over time.

In Europe, the consolidation of identity at the ethnic-group level is most likely to take place in the nineteenth century forward. Many scholars in ethnicity and nationalism argue that this is a time when cultural attributes of ethnic groups became standardized and ethnic groups became consolidated into nations through the combination of industrialization (Gellner 2006; Weber 1976) and the bureaucratization of the state (Hechter 2000; Hobsbawm 1990; Tilly 1994; Wimmer 2002). As

[11] My definition draws on Weber (1978). A thorough discussion of ethnicity as a concept and its implication for political analysis is found in Chandra (2006).

part of the centralization process, the state played a major role in cultural standardization by promulgating new institutions such as compulsory primary education, conscription, and citizenship. I hypothesize that when centralizing states sought to promote the identity of majority ethnic groups, the identity of today's minority groups in the periphery began to consolidate as a way to highlight distinctness and resist cultural incorporation.

To be clear, the rationalization of cultural practices began before the modern period. In the case of France, the effort of language standardization began in the 1530s when François I issued the Edict of Villers-Cotterêts that required French to be used as the language of the court in the conquered area of the kingdom (Ager 1999, 21; Oakes 2001, 55, 91). However, it is crucial to distinguish between the start of the process and its consolidation. It was not until the French Revolution that matters of linguistic unity attracted attention in the state center (Bell 2001; Weber 1976, 70–2). Had the rationalization effort been well enforced, non-French tongues would have been less likely to persist to the extent that the French government in 2008 officially recognized some non-French vernaculars, such as Corsican, Breton, and Occitan, as part of the country's "heritage" ("patrimoine de la France") via a constitutional revision. This example demonstrates that cultural standardization is difficult and a time-consuming process.

This section offered a theoretical framework of the evolution of ethnic identity. It argues that there was great heterogeneity in group identity within the same group in premodern times. It then hypothesizes that the consolidation of group identity for today's demographic majority and minority ethnic groups likely took place in the modern period. The remaining question is: what explains the development of a collective identity for minority groups in the periphery which is distinct from that of majority groups in the center? The following section provides a hypothesis.

2.3.3 Role of geography in the formation of autonomy

I argue that simple geographical conditions shape both the ability of the state center to project power into the periphery and the ability of peripheral groups to retain autonomy in the absence of an inde-

pendent state. Specifically, I hypothesize that geographical scale serves as a plausible proxy for institutional development for stateless ethnic groups. To begin, it is critical to recognize that overland transportation was underdeveloped and unreliable during the medieval and early modern period. Scholars point out, for example, that it took two weeks to travel from the Mediterranean coast to Paris (Stasavage 2010). It was not until the nineteenth century that major improvements in transportation occurred in Europe, owing to enabling technologies such as steam engines and railroads (Weber 1976). Moreover, geographical features like high elevation above the sea level and terrain ruggedness can impose additional costs of transportation and communications (Nunn and Puga 2012).

State leaders and monarchs partially took advantage of these limitations in organizing their polity. In premodern times, geography constituted natural borders that loosely demarcated states' sphere of influence. Oceans, rivers, mountains, and other natural terrain features provided obvious and convenient markers that helped rulers define the "frontiers" and "natural limits" of their domain, the concepts used to capture the extent of authority (Febvre 1973; Pounds 1951, 1954). Although rulers would want to maximize tax revenue in the territory under control (Friedman 1977), these geographical limits present regular obstacles to bureaucratic centralization. The resulting institutional compromise was fiscal fragmentation, where tax farming became predominant as an indirect form of taxation (Dincecco 2011). The ruling governments also devised alternative institutions, such as the sale of offices to local elites (as in France and Italy) or the nonpecuniary appointment of local officials (as in England), to buy acquiescence (Koenigsberger 1987, 44–7).

If geography imposes constraints on the ruling state, it worked relatively in favor of stateless ethnic groups to establish and retain self-rule. Put simply, geography provides an easy exit to them. For instance, when there is an access to the oceans and navigable rivers, ethnic groups can take advantage of it and establish self-rule by long-distance trade (Bosker, Buringh, and van Zanden 2013). Similarly, rough terrain features can give them leverage against being directly governed by ruling states and against heavy-handed policies like conscription or taxation. Despite the rise of

the centralized and bureaucratized state, mountainous terrain and high elevation allow peripheral communities to be shielded from the external environment and resist integration by the state center. In contemporary civil war and insurgency, these physical features work in favor of rebels because rebels have an easy time hiding themselves and use the terrain to fight back (Fearon and Laitin 2003).

Seclusion from bureaucratic centralization may create a condition in favor of a group identity in the periphery—and the institutionalization of it—that is distinct from that of the state center. In early modern Europe, the norm was the absence of popular allegiance to the state and there was the great diversity in the languages spoken (Marx 2003). Perhaps the dominant attitude of ordinary people toward political autonomy was one of indifference, as most countryside peasants were chronically impoverished and concerned primarily about subsistence in one year to the next (Weber 1976). The political body with which those people had direct and regular interactions was a much smaller unit than the state, such as "cantons," "villages," and "communes." Within these institutions, local courts and administration regulated public life, constituted the interface with the state, and retained autonomy over communal affairs (Sahlins 1989, 12–3). The development of a collective entity as an "ethnicity" in the periphery may rest on many conditions, but the key theoretical assumption is to identify a factor that grants these communities de facto freedom of action and keeps the center away from directly ruling them. This isolation affords an opportunity for language and other cultural practices in the periphery distinct from those of the center to flourish and become *institutionalized*. By the time the ruling state's capacity grows and the centralization of authority takes place in the modern period, these peripheral communities may have consolidated as ethnic groups. In short, geographical scales play a crucial role in understanding these developments.

2.3.4 Alternative arguments

The propositions thus far have established a plausible link between geographical conditions and ethnic autonomy—ethnic groups' ability to resist the ruling state. However, it is possible to theorize other mechanisms in which autonomy can arise. Specifically, I discuss two alternative hypotheses:

the local parliament and urbanization.

The first argument is the local parliament, an alternative institutional path to ethnic autonomy. It has occupied a pivotal place in European political development. Beginning in the late twelfth century, it established a structure of interactions between monarchy and subjects and determined the extent of power and privileges monarchs could enjoy (van Zanden, Buringh, and Bosker 2012). Representative assemblies played numerous functions, including constraints on the executive (Blaydes and Chaney 2013; North and Weingast 1989) and an intermediary between the monarch and the population (Poggi 1978; Stasavage 2010; van Zanden, Buringh, and Bosker 2012). This paper focuses on the local (or regional), rather than national, assemblies within state that represented the interest of the periphery. Yet the bodies on both levels had shared attributes and created a win-win situation for the rulers and ruled. For the former, parliament served as a reliable means to generate revenue and raise credit for the state; for the latter, it helped guarantee local autonomy and protect local privileges, rights, and liberties (Swann 2012). Representative bodies provided additional values to the subjects. In particular, local assemblies constituted a permanent place for accumulating political skills and experience on the regular basis. Attendees discussed, *inter alia*, legislation, taxation, budgetary allocation on public goods, and administration (Swann 2012). Some communities whose area of residence coincide with representative bodies can acquire institutional capacity and develop a distinct group identity. An exposure to and experience in such institutions can give ethnic groups an incentive to bargain for greater political demands, including recognition, autonomy, and independence.[12]

The second alternative hypothesis is urbanization. Cities of 5,000 or more population influenced not just economic growth but also political development in European history. Until 1500 C.E., these cities were relatively rare and there were many small independent city-states. This environment put city oligarchs in a politically highly influential position in determining duties on trade, taxation, monopoly rights, and protection of trade routes and cities themselves (Tilly 1994b). Some pursued

[12]Roeder (2007) examines this point using modern cases. Campbell (2012) notes that in absolutist-era France, the subjects used local assemblies to oppose demands from the monarch.

political autonomy by creating an exclusive property rights regime and shielding themselves from cash-hungry rulers through enforcement mechanisms (Stasavage 2014). Others had to bargain with rulers if they faced competition with monarchs and feudal lords (Blockmans 1994, 233). Whether cities would enjoy autonomy and grow depend on multiple forces, such as the relative strength to the governing states and the types of regime with which they had to contend,[13] but they remained viable political and economic players by taking advantage of the extraordinary ability to concentrate capital. Implications for ethnic autonomy are to consider specifically the competition between the cities and states. In places with a cluster of cities, as in northern Italy through the North Sea, states confronted a stiff competition with them (Rokkan 1999). Cities could deflect ruling states' attempt to absorb them and remain isolated, giving them freedom to grow unique institutions and cultural habits, even though they eventually had to submit to ruling states with a more coercive apparatus. By contrast, regions with less urbanization tilted the competition in favor of states (Rokkan 1999). Ethnic groups that emerged in these areas were more vulnerable to state centralization and ethnic autonomy would be less likely to be sanctioned.

In this section, I have described three plausible mechanisms through which peripheral communities without an independent state could maintain autonomy and develop distinct cultural practices as ethnic groups. These suggest that the so-called "national states," or sovereign territorial states with a coercive administrative and military capacity, are not the only route for acquiring political authority. At the same time, one hypothesis is not necessarily superior to another. For instance, geography may enhance or undermine ethnic autonomy in comparison to urbanization. While sovereign states may be able to outpace cities under their rule in the ability to raise revenues and wield authority, nature can present too high costs to build infrastructure and enforce rules on the regular basis toward ethnic groups living in a rugged terrain. Yet if state capacity is strong enough, these geographical constraints may be overcome. I examine these empirical questions in the remainder of the paper.

---

[13] Spruyt (1994), DeLong and Shleifer (1993).

2.4 Empirical Strategy

Since the issue of ethnic autonomy has not attracted scholarly attention, systematic data is yet to be compiled. Empirical evidence on the rise of the sovereign territorial state is available and expanding,[14] but independent statehood is not the outcome of interest in this paper. Given the limitations, an alternative approach is to sample ethnic groups that share similar circumstances but diverge on a range of political autonomy outcomes. These include some groups that were once independent in premodern times but no longer so today and other groups that have been successfully able to resist full incorporation by the ruling state. This strategy allows me to exploit the variation of these outcomes while controlling for factors that potentially confound the effect of the explanatory variables. Thus, to test the above hypotheses, it is necessary to collect data systematically on the attributes that are not only theoretically relevant but capture the background conditions for the observed groups.

In this paper, I present evidence from France for three reasons. First, France offers a wide range of variation in ethnic autonomy outcomes.[15] While it is known for a history of the government with highly centralizing tendency, today's France houses a number of ethnic groups. Some, like the Bretons, Occitans, and Corsicans, have been established to the extent that their vernaculars have earned an official status.[16] They can be called "recognized" ethnic groups. Others, like the Burgundians, Normans, Savoyards, and Alsatians, do not enjoy such a status but are known to have historically resided in this part of the continent. They can be labeled as "unrecognized" ethnic groups. While none of them now have an independent state, there is a diverse history of political autonomy. For instance, the Burgundians once were independent; the Bretons sought autonomy until recently. This is a defensible mix in the range of outcomes: a single "nation-state" (the French) and many ethnic groups within it.

---

[14] See, e.g., Abramson (2017), Boix, Codenotti, and Resta (2011), Wimmer (2015), Wimmer and Feinstein (2010).

[15] Gerring (2007) calls this approach a "diverse case" method, in which researchers deliberately draw a subset of cases, based on the chosen outcome variable, explanatory variable, or some theoretical relationship between the two, from the population of cases with a full range of variation.

[16] This is part of the constitutional revisions enacted in July 2008.

Second, there is great variation in political and economic power within France throughout its history. Contrary to the popular imagination, the process of state centralization in France was never certain, steady, or systematic. As indicated above, France was one of those "composite" states. When it expanded the frontier, the system of indirect rule was installed, where jurisdiction was asserted but the enforcement of rule was ad hoc and more de jure than de facto. This was particularly the case in the south, bordering on the Mediterranean and Spain (Given 1990; Sahlins 1989). There, too, is within-country variation in wealth. Although trade was less active in medieval France than in its eastern neighbors, scholars observe a variation where the north was, on average, more prosperous than in the south (Spruyt 1994, ch. 5). Did urbanization in the peripheral provinces promote autonomy by being able to resist centralization or, conversely, was it counterproductive by inviting Paris' further administrative integration for more revenues? A historical analysis which capitalizes on these variations offers a rich context for addressing this paper's hypotheses.

Finally, France presents a tough case. Extant research suggests that France lay outside the highly urbanizing areas in the east. Rokkan (1999) hypothesizes that France had early state-building due to few serious competitors from urban, autonomous communities within it and that state officials did not have much difficulty projecting authority from Paris and undermining autonomy in the periphery. Scholarship on state formation reinforces this view, by arguing that France is one of the early and most effective developers of the state.[17] It follows that this is a very challenging environment for non-French ethnic groups to establish or maintain political autonomy. For these three reasons, France constitutes a good case for my empirical strategy.

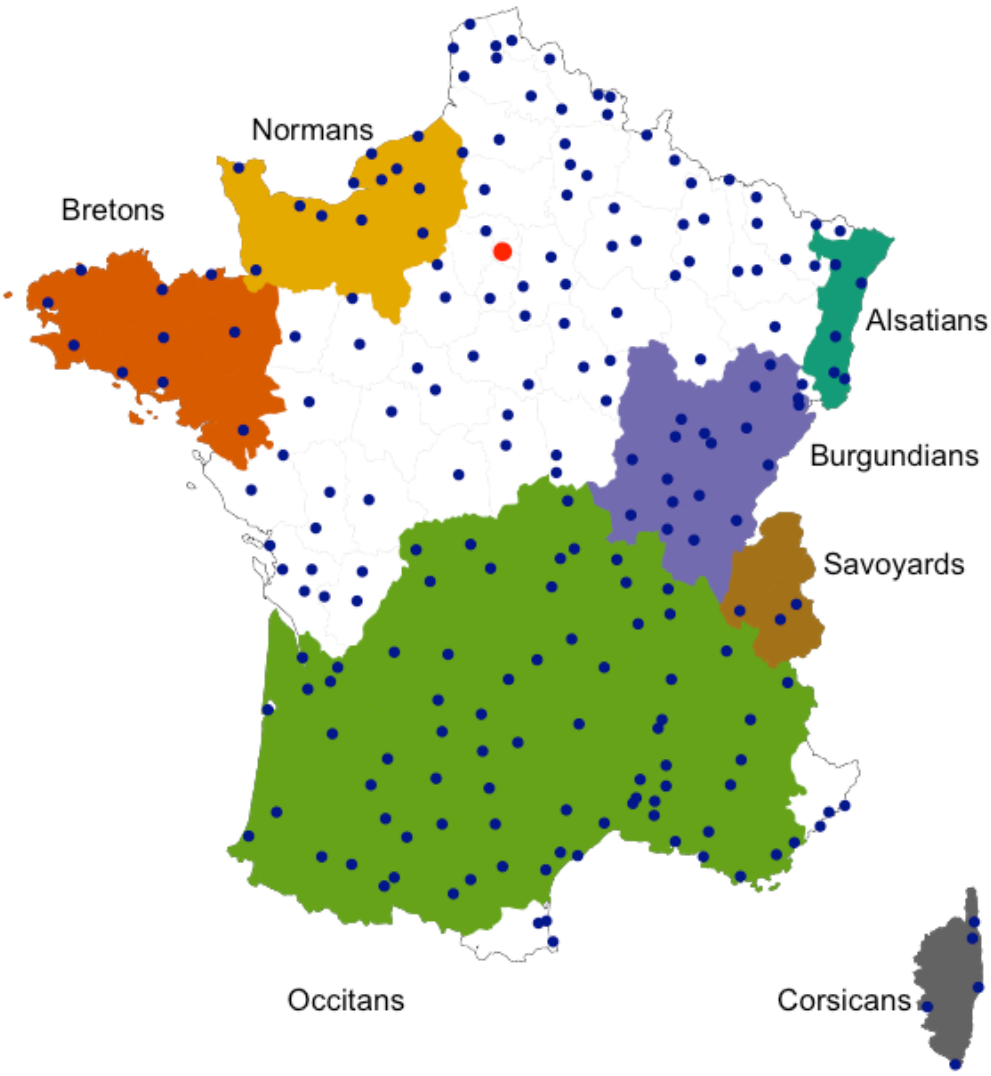
To test my argument, I have built a new data set from a variety of sources. It is composed of 214 cities that come to be under the French jurisdiction today, for the period between 1400 and 1900 C.E. The city data draws from Nüssli (2011), which offers GIS (geographical information system)-based information on the location, administrative divisions, and political status for the states and cities within them that existed at the final year of each given century. The choice of the city as a prox-

[17] See Ertman (1997), Tilly (1992), among others.

imate unit for ethnicity rests on empirics. Ethnic groups predominantly live in a clustered fashion and designate a hub city as their "capital," in which economic activity flourishes and institutional development occurs. This historical pattern allows me to consider city-specific attributes as a useful starting point of collecting information at the ethnicity level. I draw on Minahan (2000) to specify the historical location of ethnic groups. Figure 2.1 shows the location of seven ethnic groups housed historically in France (excluding, the French). The data set is time-series and cross-sectional, organized in fifty-year periods with time-varying and time-invariant variables, but as mentioned below, because of the limitations on the outcome variable, I limit statistical analysis to the cross-section.[18]

[18]I also collected the information on the cities in the adjacent states that were once under French rule over this period. There are at least 256 cities in the data set. However, I have dropped those that eventually fell outside of French authority today for the theoretical and methodological reasons. The theoretical rationale is that these cities were nominally "French" and never developed long-term, institutional connections with Paris, such as a provincial estate. The methodological reason is that limitations of data availability due to the lack of institutional ties make systematic analysis difficult.

Figure 2.1: Ethnic groups in France and the cities. The red dot indicates Paris.



The inclusion of these seven ethnic groups is based on the fact that their city (or "capital") of historical importance is located in today's French territory. To be sure, there are more than seven

ethnic or linguistic groups in France. There are Dutch speakers in the north; some Basques and Catalans live across the southwestern border. For the Basques and Catalans, however, their primary city is found on the other side of the Pyrenées (Vitoria and Barcelona, respectively). Although they have historical ties to France, they make demands for greater autonomy or independence on the Spanish, not French, government. For these reasons, I keep the seven ethnic groups shown in Figure 2.1.

I construct the *ethnicity* indicator in two steps. First, I create an indicator whose value equals 1 when a given city is part of the historical territory for one of the seven non-French ethnic groups. Second, based on this coding, I make an umbrella category of ethnicity that takes the value of 1 by aggregating these indicators so that any given city is part of non-French ethnic groups. In both steps, the reference category is the French and is not included in the statistical analysis.

The outcome variable is ethnic groups' autonomy over their affairs as captured by their use of non-French languages. I draw the data from Eugen Weber's classic work, *Peasants into Frenchmen* (1976). While typically cited as evidence of change in peasants' loyalty to France as a culturally unitary group, it also contains a wealth of untapped data before the change. More specifically, Weber (1976) illustrates how various ethnic groups in the French countryside were isolated from the metropolitan regions and resisted the ruling state. I use this data to construct two language-related proxies for resistance. Language use is a very useful measure, because it represents both the extent to which the state is capable of exercising its authority on its subjects' behavior at a microlevel and the extent to which subjects comply or refuse to follow rule. In addition, language is one of the most important dimensions with which to define ethnicity in the literature on ethnicity and nationalism, and scholars often use language to explain political conflict between ethnic groups.

I employ two measures to tap different dimensions of language use. The first is the proportion (0–100 percent) of non-French-speaking communes in a province. The French commune is an administrative unit roughly equivalent of township or municipalities. This is a small enough unit to understand the variation of French-language use on some aggregate level. The second is the pro-

portion (0–100 percent) of the population speaking languages other than French. Though closely related to the first, it captures another dimension of language use, as each commune certainly has a varying population size. Together, these two variables offer a consistent assessment with regard to the ruling state's exercise of political authority on ethnic groups in the periphery and to the degree that such exercise is accepted. The data comes from the population survey conducted by the Ministry of Public Instruction in 1863.

It is important to note the choice of these variables. Weber (1976) contains a host of other data related to ethnic autonomy. These include, *inter alia*, the extent to which non-French language were entrenched, the cost of tax collections, draft evasions, and electoral abstentions. Ideally, I should use this information to examine aspects of ethnic autonomy other than language. For instance, high costs of tax collection may indicate the lack of loyalty to the French government and resistance to French rule. However, these data unfortunately have highly limited variability in the cross-section and are inadequate for statistical analysis. The two outcome variables chosen here are time-invariant but offer a great variability within each and capture variation on different dimensions on a single outcome, language use. Since the outcome variable is time-invariant, statistical analysis is based on the subsetted data for the period in which the outcome measure is recorded.[19]

There are three sets of explanatory variables. The first is a set of geography hypotheses. I calculate the geodetic distance (measured in kilometer) between Paris and a given city i for $N = 214$ cities. For land elevation above the sea level and terrain ruggedness, I use the GLOBE (Global Land One-kilometer Base Elevation project) database (GLOBE Task Team and others 1999).[20]

The second is the institutions hypothesis on the local parliament. The impact of French regional assemblies, provincial estates, is measured in the aggregate number of years that estates were held during the *ancien régime*. It comes primarily from Kiser and Linton (2002) and is supplemented by Blockmans (1976) and Swann (2012). Not all provinces had an assembly, but some had as many

[19]The data for all other variables is available longitudinally but not used in empirical analysis.

[20]The terrain ruggedness index (TRI) is originally proposed by Riley, DeGloria, and Elliot (1999).

as almost 400 years of experience in the representative body between 1400 and 1789 when these institutions were abolished along with other Old Regime bodies in the wake of the Revolution.

The third explanatory variable is a set of economic hypotheses on urbanization. Following the convention in economic history, I use a population-related measure as a proxy for economic growth. Specifically, I use "urban potential," originally devised by de Vries (1984). It captures a city's urbanization rate relative to its neighbors, calculated by the sum of the population divided by the geographical distance between the given city and all other cities. It shows if a city is surrounded by small, rural towns and villages or by competing urban centers. The underlying population data comes from Bairoch, Batou, and Chévre (1988) but I follow Bosker, Buringh, and van Zanden (2013) which updates the Bairoch et al. data.

In addition, I include a battery of controls that can plausibly affect ethnic groups' ability to develop and retain autonomy. The first is the measure on the cumulative years of French rule since the year 1400, based on sources such as Nüssli (2011) and Darby and Fullard (1970). It ranges from 0–500 years (i.e., 1400–1900 C.E.). The territorial shape of France expanded and contracted over time to where it is today, and accordingly this affected the political status of cities especially around today's borders. I take care to consider the period in which the "owner" of the given city was other than French or some areas were independent and exclude it from the aggregate years. This variable captures the duration of the interaction between the French and minority ethnic groups and partly accounts for the longitudinal dimension.

The second is linguistic distance as a proxy for culture. One can argue that ethnic groups' willingness to preserve self-rule is driven not by urbanization or institutional development captured by distance or local assembly but by cultural differences between them and the French. A deep division in cultural practices presents high costs for non-French users to switch, especially in language. When, as in premodern times, the ruling state's capacity to administer laws in the periphery is limited, a small circle of provincial intellectuals may have an incentive to make such a costly investment to be a profitable interlocutor useful for both bureaucrats in the state center and population in the

countryside (Laitin 1988). Yet most in the periphery do not have such resources to switch and are likely to remain culturally isolated and thus "distant." To control for this possibility, I include a "linguistic distance" measure, originally proposed by Fearon (2003).[21] It is designed to capture the "proximity" of two given languages based on the extent to which they share the classifications of the language tree as defined by linguists. When two tongues diverge early in a tree, this signifies greater linguistic difference, on average, than those that branch out later. Following Fearon, I first use Lewis, Simons, and Fennig (2013) to construct the linguistic distance measure for all sets of the language pairs among the sample. Yet as with most measures on ethnicity and ethnic diversity, Fearon (2003) uses the state as the unit of analysis. Since this paper examines within-country variation, I multiply Fearon's formula by each ethnic group's population size for each time period as the weight to obtain variation across ethnic groups. This variable proxies one ethnic group's cultural distance to all other groups in France, including the French.

The third is the printing press. Empirical research shows the press's long-term impacts in various fields, but the main ones involve technological innovations for more efficient trading practices and the lowered cost of becoming literate through the dramatic reduction in the cost of access to information.[22] Cities that hold presses can become a business and intellectual hub, facilitating economic and institutional capabilities for ethnic groups. This is an indicator taking the value of 1 if a city gets the printing press, and I aggregate the number at the provincial level for each period.

The fourth is the university. Since its emergence in the Middle Ages, universities constituted prominent sites where individuals acquired human capital and systematic training in law. As economic historians have shown, access to university education may be associated with the development of legal institutions and state administration (Cantoni and Yuchtman 2014).[23] This serves as a

---

[21] Fearon calls this measure "cultural fractionalization."

[22] Mokyr (2006), Dittmar (2011), Buringh and van Zanden (2009), Febvre and Martin (1976). Data sources include Clair (1976), Pettegree (2007), Conner (2001), Walsby (2011a,b), Bouchot (1890), Reske (2007), Burke (2004).

[23] Some scholars theorize that university training in law promoted state bureaucratization and the rise of absolutism in Central and Eastern Europe. See Ertman (1997).

proxy for an institutional variable other than those captured by geography and the regional parliament. I use Frijhoff (1996), Rüegg (2011), and Darby and Fullard (1970) to obtain the foundation date for universities. This is an indicator taking the value of 1 if a city gets a university, and I aggregate the number at the provincial level for each period.

The fifth is bishoprics. In premodern times, the Catholic church played an important role in state administration in France, such as providing the church's administrative boundaries to determine the extent of the state's jurisdictional reach and identify taxable subjects (Sahlins 1989). The presence of diocese and archdiocese can provide an organizational infrastructure for ethnic groups. Like the local assembly, residents can learn not just organizational skills but have an easier access to resources for mobilization, especially in a religious hub. This is another institutional proxy focusing on Catholicism. I collect information of Catholic bishoprics based on Chaney (2015).[24] This is an indicator taking the value of 1 if a city has either a diocese or archdiocese, and I aggregate the number at the provincial level for each period.

The sixth is the consequence of the Protestant Reformation. The historic religious struggle in the fifteenth century, some scholars argue, had a discernible impact on institution-building related to the state and how actors sought political autonomy throughout Europe (Gorski 2003; Nexon 2009). During this tumultuous period, political elites were often under pressure to declare explicitly to take either side in public and implement necessary institutional changes, including the education system (Gorski 2003). Those who underwent these changes would likely acquire knowledge and a capacity to organize, and the impact is predicted to be stronger as cities are closer to the center of the Reformation. I account for this institutional variable by taking the geographical distance from each city to the German city of Wittenberg where the movement had initiated.

Finally, I include a set of variables on Roman influence. The purpose is to account for early institution-building experience. The cities reached by the Romans had access to roads, which in turn gave them an opportunity to develop institutional capacity in economic activity and political

[24]Since the status of some dioceses can change through status elevation, suppression, or incorporation by others, I also account for this change over the period of analysis time.

organization. I use Talbert (2000), Hammond (1981), Åhlfeldt (2015), and *Pleiades* (2015) to collect information about the Roman legacy. I create an indicator taking the value of 1 if a city had major or minor Roman roads and a variable measuring the years under Roman rule.[25] I also used these sources to calculate the number of years the given city was believed to under Roman rule. See Appendix A for the summary statistics.

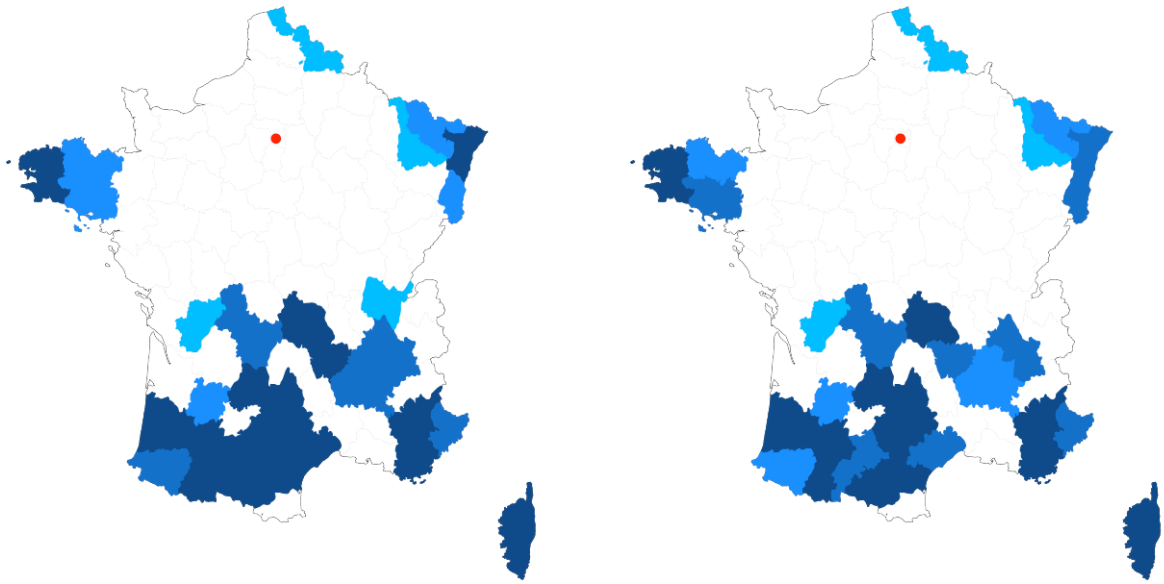
2.5 Evidence

2.5.1 Descriptive evidence

I start with descriptive statistics with a focus on the outcome variables and main explanatory variables. Figure 2.2 is the set of plots of the dependent variables, where the observed data at the provincial level is plotted on the map. The data is shown in quantile, where a darker color indicates higher values in percentage. The darkest blue indicates the top quantile (75–100 percent), next darkest the second quantile (50–74 percent), medium dark the third quantile (25–49 percent), and light blue the bottom quantile (0–24 percent). The provinces in white thus indicate that the use of the French language was well established before the start of the Third Republic. The distribution of the data looks similar, but it is crucial to distinguish the two because of the assumption that each commune has a varying population size. Figure 2.2 establishes a broad pattern that ethnic groups' autonomy measured in the nineteenth century tended to occur in provinces far away from Paris. This is seen in particular in the east, the west, and the south, where the Alsatians, Bretons, Occitans, respectively, claimed their historical residence.

[25] Following Bosker, Buringh, and van Zanden (2013), those cities with two or more major roads are coded "Roman hub" and those with one major road or one or more minor roads are coded cities with "Roman roads."

Figure 2.2: Geographical distribution of dependent variables (the red dot indicates Paris). Darker colors indicate higher proportions. The data is from 1863.



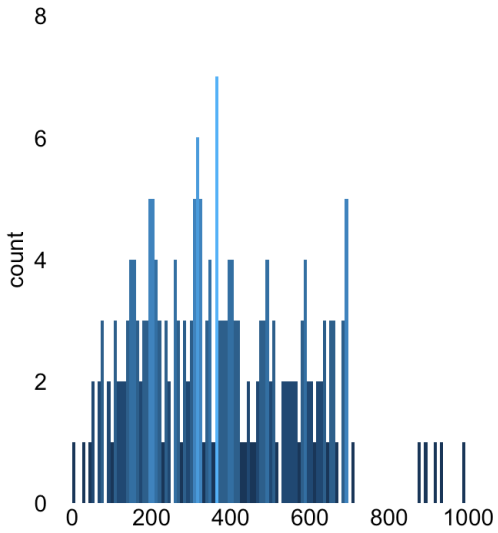
(a) Proportion of non-French-speaking communes. **(b) Proportion of non-French-speaking population.**

I have argued that geographical conditions shape ethnic groups' ability to resist French rule. To evaluate this hypothesis, I present three measures that capture different dimensions of geography: (1) distance from Paris; (2) elevation above the sea level; and (3) terrain ruggedness. Distance to Paris is a proxy for how hard the ruling government is expected to project power over the remote areas. Although it is believed that land-based transportation was unreliable until the modern age, systematic, concrete data is absent. This variable serves as an alternative. Limited access to distant locales, in turn, allows them to foster unique methods of regulating life, including cultural practices and speech distinct from those of the French.

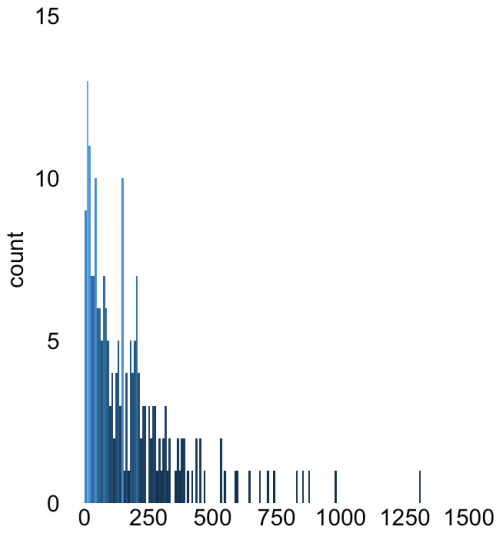
The second geographical scale is elevation. It gauges another aspect of the extent of power projection from the central government. The higher communities are located above the sea level, the

more difficult it is for Paris to enforce rules directly. The third measure is terrain ruggedness. When land is uneven, the cost of travel and casting power is expected to mount. This feature also provides natural protection for local groups when they choose to resist heavyhandedness from the ruling state. Figure 2.3 plots the distribution of these geographical measures. They exhibit wide variability, especially the data on distance to Paris.

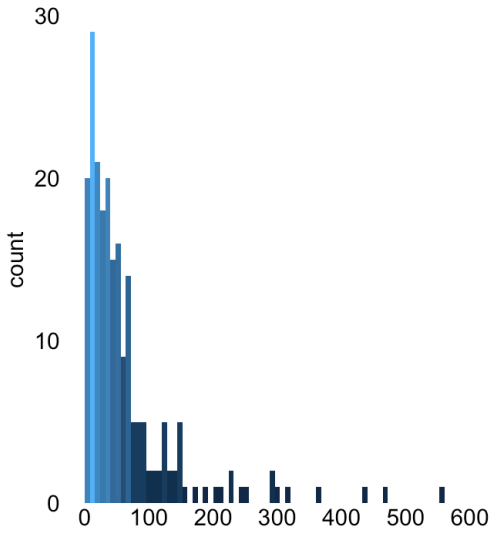
Figure 2.3: Distribution of geographical scales. Lighter colors indicate higher counts.



(a) Distance from Paris (km).



(b) Elevation above sea level (m).



(c) Terrain Ruggedness Index (m).

I test the effect of these geographical scales using the following estimation strategies. This paper employs ordinary least squares (OLS) estimators, with the variable on cumulative years under French rule to account for time, albeit imperfectly. Then I consider the robustness of the geographical hypotheses under various specifications and discuss the impact of geographical scale using different scenarios.

2.5.2 OLS regression results

First, I broadly discuss the effect of geographical measures. In all models, I focus on the quantities of interest and leave all estimation outputs in Appendix A. Figure 2.4 shows two OLS results with select predictors of theoretical interest for each of the two outcome variables, the saturated model on the left column, including all control, and and an interaction term added to it on the right. I present estimates in percentage changes as the dependent variable is coded 0–100 percent.

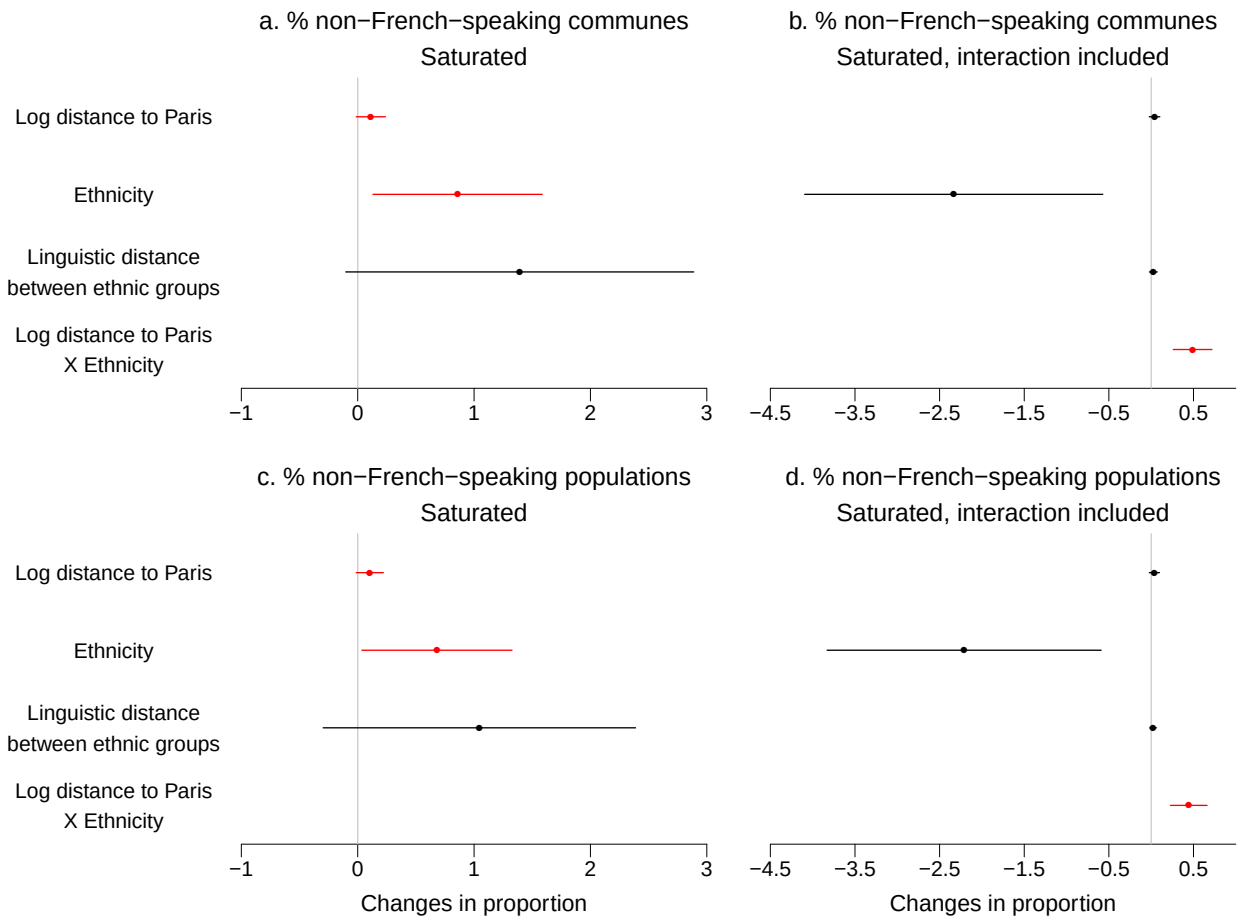
The top left panel indicates that one geographical measure, log distance to Paris, is positively and significantly associated with the percentage of communes lacking French speakers.[26] More concretely, 1 percent increase in the distance leads to an increase in proportion by approximately 0.0011.[27] Other geographical scales, elevation and terrain ruggedness, are not statistically significant and not shown here. At the same time, ethnicity is also positively correlated. So is linguistic distance between ethnic groups, although mostly not significant. Similar results obtain in the bottom left panel, where 1 percent increase in the distance from Paris leads to an about 0.001 increase in the proportion of the population speaking non-French tongues. Although the lower bound of the confidence interval crosses zero in terms of statistical significance, distance to Paris has substantive impacts as presented below.

---

[26] I log-transformed this variable before running regression, because of the non-linearity of the geographic distance variable. The effect of a one-unit change in short distances, say 10 to 20km, is likely to be greater than that of the same change in greater distances, say 500 to 510km. Since I use OLS, log-transformation allows for the linear interpretation of the effect of non-linear variables.

[27] The coefficient is 0.113 in Figure 2.4a. For a small value, the interpretation of an unlogged, one-unit change can be approximated simply by dividing the value by 100.

Figure 2.4: The OLS regression output on key predictors.

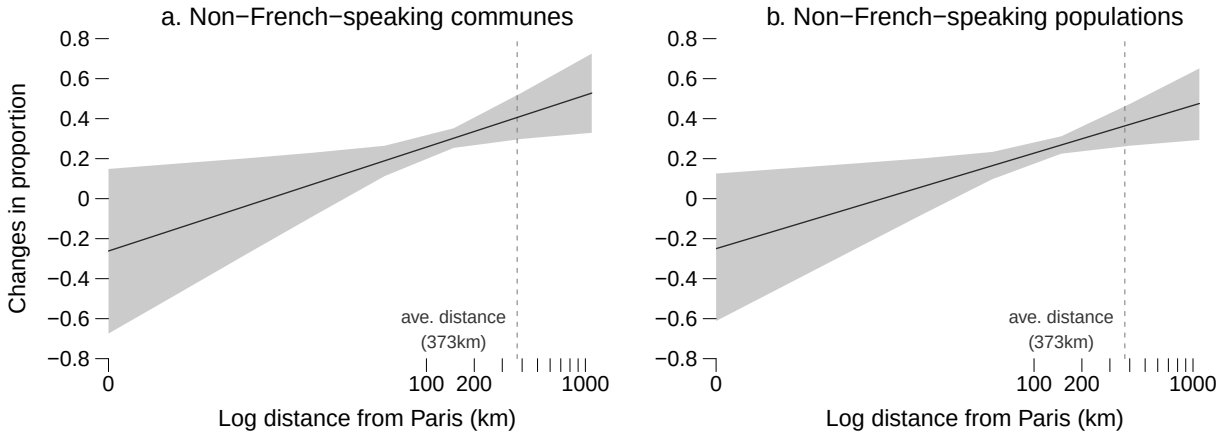


Note: Solid dots denote point estimates and horizontal bars denote 95 percent confidence interval.

The rationale for interaction models is to identify the effect of distance to Paris specifically for ethnic groups in the countryside. It is also to highlight the impact of ethnicity beyond its average effect when it is included as a predictor. The right column of Figure 2.4 reports that geographical distance has a *greater* magnitude, in which the farther ethnic groups are from Paris, the higher the proportion of non-French-speaking communes or population are expected to be by 0.49 or 0.44 percentage points (panels (b) and (d), respectively). Although the linguistic distance variable is statistically significant in the models without an interaction, it seems less robust when an interaction is added.

One can point out that the effects in one unit change in the distance-to-Paris variable may be interpreted as quite small. To show substantive effects, I estimate the marginal impact of a unit change in the distance variable on each of the outcome variables and report the results in Figure 2.5.

Figure 2.5: The substantive effect of distance from Paris (from OLS regression).



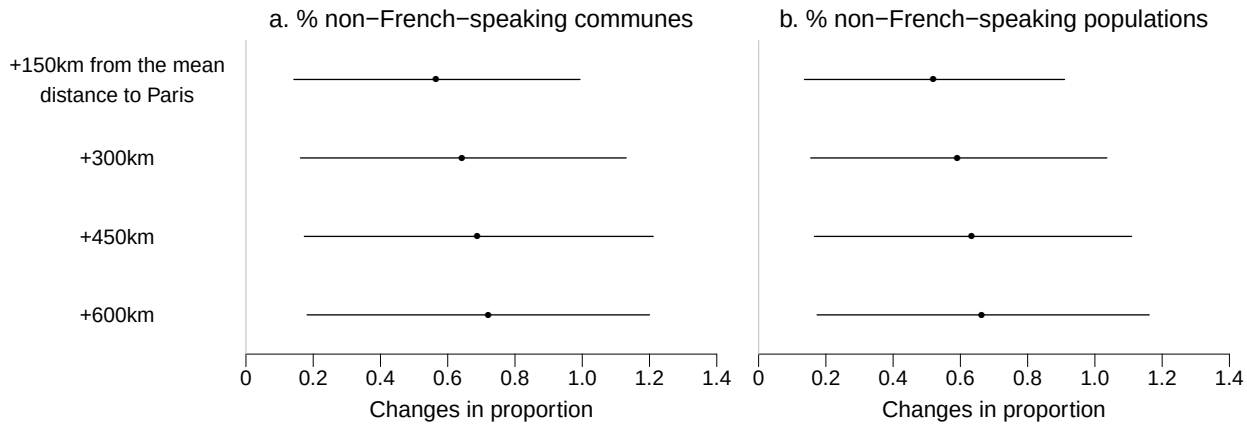
Note: The solid line denotes point estimates and the shaded region denotes 95 percent confidence interval.

Figure 2.5 suggests that distance does seem to matter in understanding ethnic groups' ability to retain autonomy. While ethnic autonomy is expected to be curtailed within 100 kilometers from Paris, it begins to grow as the location of communities gets farther. In each scenario, the proportion

of using non-French languages would increase by between 0.2 and 0.4 percentage points, with an average distance of 373km at the vertical dashed lines. This simulation supports the hypothesis that the distance parameter is an adequate proxy for the ruling state's capacity to project power over the periphery.

Another way to examine distance's substantive impact is to focus on the differences. In Figure 2.6, I simulate the first differences between the mean distance and several predefined values. It again shows that in both outcome variables, ethnic autonomy is expected to be greater as distance from Paris increases. See, for instance, the first scenario of 150km difference from the mean distance to Paris (i.e., about 520km away from Paris). The use of non-French languages is predicted to increase by around 0.56 percentage points for one standard deviation change (approximately 200km) from the mean, and the magnitude becomes greater as the distance increases. The results here suggest that land-based transportation seemed to be unreliable back then and would likely play a crucial role in enforcing rule from the distant center.

Figure 2.6: The simulated effect of varying distances from Paris (from OLS regression).

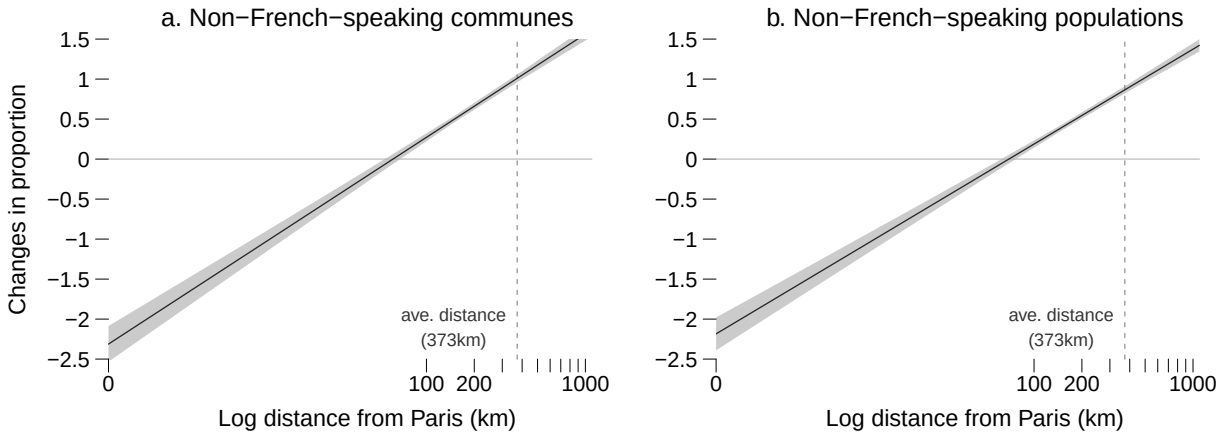


Note: Solid dots denote point estimates and horizontal bars denote the 95 percent confidence interval.

Next, I consider the effect of the distance variable in conjunction with ethnicity. The model is

designed to capture the role of ethnicity in terms of distance. Figure 2.7 presents substantive impacts of the interaction model between distance to Paris and ethnicity. For instance, at around 100km, where the two lines cross, location does not seem to matter (or is expected to have *negative* effects) in ethnic groups' ability to retain autonomy. At the mean distance, however, distance has a clear impact of nearly 0.8 percentage points (see the vertical dashed line) and the magnitude continues to grow larger. It gives strong support for the hypothesis that distance from the ruling state matters for ethnic groups residing in the periphery to retain autonomy. The model also suggests that the magnitude of distance gets more pronounced as the distance between the ruling center and ethnic groups in the countryside widens.

Figure 2.7: The effect of interactions between ethnicity and distance from Paris (from OLS regression).

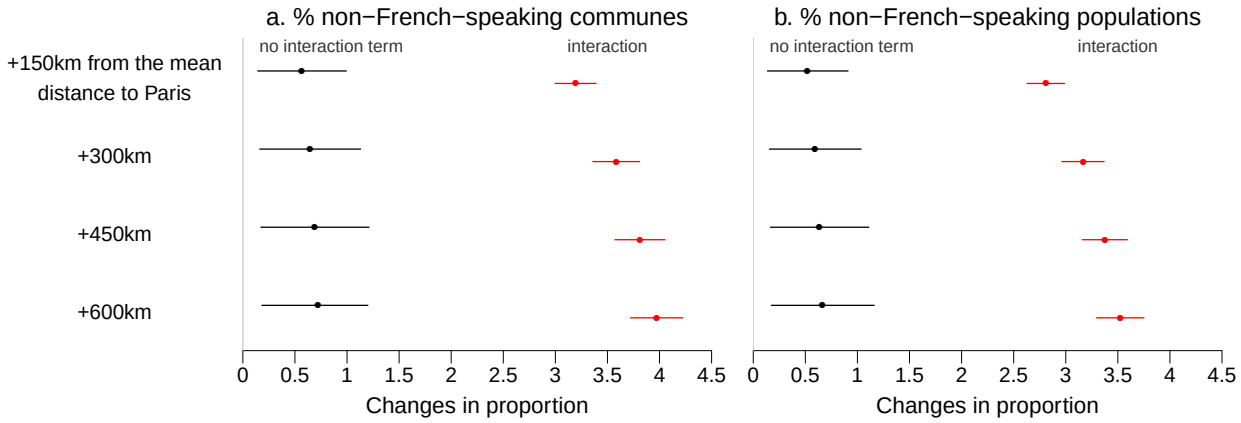


Note: The solid line denotes point estimates and the shaded region denotes the 95 percent confidence interval.

Figure 2.8 further investigates the interaction model. It shows the first differences effects of varying distances in four scenarios in a comparison between the interaction and non-interaction models (the latter adopted from Figure 2.6). The comparison provides additional support for the importance of distance to understand ethnic autonomy. Figure 2.8 documents clear differences—the magnitude of three to four times in point estimates—between the two models. These differ-

ences further support the hypothesis that distance plays a crucial role in preserving self-rule and autonomous cultural practices for ethnic groups without an independent state.

Figure 2.8: The effect of interactions between ethnicity and varying distances from Paris (from OLS regression).



Note: Solid dots denote point estimates and horizontal bars denote the 95 percent confidence interval.

In sum, an array of evidence presented in this section has shown that geographical scale, particularly the distance-to-Paris measure, plays a crucial role in understanding the development of ethnic autonomy. Although this variable has a positive, substantive effect, it becomes stronger when the impact is conditional on ethnic groups within a state. The analysis also demonstrates that the farther the homeland of these ethnic groups is located from the state center (i.e., Paris), the greater the impact of geographical scale is likely to be. Appendix A presents the analysis of further robustness checks, including an instrumental-variable model to account for unobserved heterogeneity and other confounding factors not measured in the OLS models.

2.6 Conclusion

This paper has discussed the evolution of political autonomy for European ethnic groups from the bottom-up perspectives. It points out that the literatures on sovereignty and the homeland focus on the state as the unit of analysis, leaving out of the research agenda how ethnic groups claimed and retained autonomy over time. I presented a simple theoretical framework, in which geographical conditions affect local institutional development and the the ability to retain autonomy against the ruling state for ethnic groups in the periphery, and offered two alternative mechanisms—regional assemblies and urbanization. Based on an original data set of 214 cities and seven historical ethnic groups in France, I examined these hypotheses in statistical analysis for the period 1400–1900 C.E. Evidence suggests that a geographical scale, especially distance from Paris, has a consistent and significant effect on one dimension of ethnic autonomy proxied by the proportion of users in non-French languages. Analysis also supports the claim that ethnicity plays a role in accepting or resisting France's language policy in the nineteenth century in that areas where ethnic groups claim their land of origin see greater proportions of non-French users.

Future research needs to widen the scope by testing other dimensions of political autonomy than language use and by extending the analysis to a greater population of ethnic groups in other areas. Rokkan (1999) argues that the path to state formation and related institution-building is distinct from the one France experienced. He hypothesizes that in the other route, state size is smaller, state capacity is weaker, and the political influence of cities with heavy trade activity is greater. If this hypothesis is put to test, one is expected to find a strong correlation between the short distance to the state capital and ethnic autonomy. Empirical data not only adds evidence to a theory of political autonomy regarding ethnic groups but offers a test for one of Rokkan's rich, complicated theories on ethnic diversity that are yet to be systematically examined.

Chapter 3

PUBLISHING NATIONS: TECHNOLOGY ACQUISITION AND LANGUAGE STANDARDIZATION FOR EUROPEAN ETHNIC GROUPS

3.1 Introduction

The causal connection between culture and economic growth has garnered increasing attention in social science. Although the broad linkage is traced back to Adam Smith, recent works in economic history focus on specific functions of culture. Some, for instance, argue that wealth may in part be driven by an "inclusive" culture in society, where no one is called heretical when challenging the conventional wisdom or where there is an institutionalized medium, such as journals, that guarantees the freedom of speech for good ideas (Mokyr 2002, 2016; McCloskey 2016). Empirical analyses also demonstrate that certain individual beliefs and preferences, which are geographically concentrated, are growth-enhancing (Guiso, Sapienza, and Zingales 2006; Tabellini 2008, 2010). Other works show that inclusivity matters but excessive diversity may have adverse effects by incurring high transaction costs (Ashraf and Galor 2013).

Among the myriad attributes of culture, language offers a tangible and testable causal channel to economic growth. Language, in theory, can be conceptualized as a medium to gain market access, information, public goods, and rents. Many societies use language as a political instrument to build barriers to entry and deny certain groups access to market transactions and economic resources (Laitin and Ramachandran 2016; Liu 2014). Language can also function as a focal point around which to build effective political coalitions (Fearon 1999). This is in part why speakers of the same language in highly diverse societies typically coalesce as an ethnic group in the competition for po-

litical power and economic opportunities.[1] Thus, the ability of ethnic groups to use language as an instrument of power can have profound consequences not only on the economy but for society as a whole.

However, few theoretical and empirical works systematically address why some ethnic groups gain the ability to use language more effectively than others. Although social science research has offered general hypotheses, the roots of the standardization of core cultural attributes for ethnic groups, especially language,[2] have not been widely investigated. It is one of the most difficult cultural attributes to be standardized, because, compared to flags and anthems that demand little learning, *codified* language requires literacy.[3] The work that comes closest to theorizing about language rationalization is that of Benedict Anderson (2006). His thesis connects the profit motive of printers and booksellers ("print-capitalism"), which first arose following the invention of the printing press in late fifteenth-century Europe, to "national consciousness." Yet, as I will show later, a broad description of this linkage leaves multiple causal mechanisms unaccounted for. Moreover, systematic evidence that tests Anderson's hypotheses about language standardization has not yet been offered.

This article fills these gaps by providing a simple conceptual framework and empirical analysis of how language standardization occurs for European ethnic groups. I distinguish two plausible channels that specify Anderson's (2006) hypothesis that the acquisition of the printing press could spur the greater use of the vernacular over Latin. The first is the selection process, in which access to print technology in early-modern times gave ethnic groups an impetus to make innovations, including vernacular codification, in later periods when such development is deemed necessary and viable. Accessibility is critical because fixing language was never a reason for the spread of presses. Ethnic groups with a sovereign state benefited from this channel. The second process hypothesizes

[1] For empirical works on the relationship between ethnic diversity and economic performance, see Alesina et al. (2003), Alesina and La Ferrara (2005), Desmet, Ortuño-Ortín, and Wacziarg (2012), Easterly and Levine (1997).

[2] For major works, see Gellner (2006) on the impact of industrialization, Anderson (2006) on the role of literacy and capitalism, Smith (1986) on the availability of ancient symbols.

[3] In this paper, I use "language standardization," "language rationalization," and "language codification" interchangeably.

the start of standardization by conscious choice on the part of early-modern printers who promoted vernacular publications to meet their interest in education, proselytization, and profit. Many ethnic groups that were able to codify their tongue in modern times without independent statehood took advantage of this coincidence. Multiple routes to the equifinality between print technology and language standardization reflect the wide-ranging effects of the printing press (Eisenstein 1979; Dittmar 2011).

Findings from a new data set support my argument. This article presents statistical evidence for 171 European ethnic groups in the period between 1400 and 2000 C.E. The evidence is consistent with the hypothesis that the acquisition of the printing press is positively associated with standardization of the vernacular and that the early adoption of the technology corresponds to greater chances of standardization. Using event history models, I find that ethnic groups that adopted the press have a 4 to 9 times increase in the chance that their vernacular gets standardized. Similarly, the evidence shows that early adopters are 6 percentage points more likely to rationalize their vernacular in the modern period than latecomers. The results hold whether or not ethnic groups owned an independent state. Main findings are based on Cox proportional hazard models, but I also test their robustness using logistic regression models. I address endogeneity concerns, first, by accounting for the possibility that the spread of print technology is driven by human capital and, second, by using the geographical distance to Mainz as a source of exogenous variation for the spread of presses. All models confirm that the adoption of print technology has a substantive and significant impact on language standardization.

In this article, I make two contributions. First, I seek to demonstrate that there are non-state routes to accounting for the variation in cultural consolidation among ethnic groups. Territorial sovereignty plays so integral a role in cultural preservation that it is commonly embraced in the literature on ethnicity and nationalism as a precondition for the construction of nations (Anderson 2006; Breuilly 1993; Gellner 2006; Hechter 2000). Yet recent research on the causes of ethnolinguistic diversity has empirically shown that the state is not the only determinant and that it is critical

to consider non-state determinants such as geography and technology (Ahlerup and Olsson 2012; Ashraf and Galor 2013; Laitin, Moortgat, and Robinson 2012; Michalopoulos 2012; Spolaore and Wacziarg 2009). This paper focuses on the impact of access to technology and the timing of technological access to address variation in a core attribute of ethnicity, namely language standardization. My second contribution is to show deep historical roots of contemporary politics that takes advantage of ethnicity and culture as political instruments. An emerging body of empirical work in economic and political history demonstrates that the impact of historical events can have centuries-long persistence on varying economic and political outcomes today (Comin, Easterly, and Gong 2010; Nunn 2014; Olsson and Paik 2016; Spolaore and Wacziarg 2013; Stasavage 2014). This paper joins this growing literature by providing evidence that contemporary ethnic groups' ability to use language for political participation and access to the market is determined many centuries ago.

This article is organized as follows. The next section briefly discusses key concepts of this study. The third section lays out a conceptual framework of my argument that yields testable hypotheses. The fourth section documents the empirical data to test these hypotheses. The fifth section documents the findings from statistical analysis. The concluding section summarizes key evidence from this study and discusses implications for comparative research beyond Europe.

3.2 Cultural Rationalization, Ethnicity, and the Nation

The unit of analysis in this paper is the ethnic group. Before laying out hypotheses, it is critical to define "ethnicity" and the related concept, the "nation," and address the relationship between the two. I define ethnicity as a category that delimits membership to a group, and ethnic group as a collectivity determined by a set of attributes shared among its members. Following many scholars, I use the myth of common descent as the most basic characteristic for the definition.[4] As explained below, many ethnic groups attempted to consolidate their cultural practices as an optimal way of political survival in a centuries-long competition over resources and territory, which was particularly intense

[4]Weber (1978) is oft-invoked. An extensive review of the definition of ethnicity and its implication for political analysis is found in Chandra (2006).

in Europe (McNeill 1982).[5] Any symbols reminiscent of common origin may be selected for group consolidation, such as a memorable landscape, historic buildings, clothing, and a flag. By contrast, the nation may be defined as a collectivity with standardized culture that commands loyalty among the community members. Standardization here is synonymous with institutionalization;[6] a nation has it a rule that designates certain cultural practices to be written as "official," as in an anthem, interpretations of its origins, and language use, which are consistently invoked to consolidate group membership. Successful standardization requires literacy among the members of the nation so that official culture is effectively promulgated and sustained from one generation to the next. Ethnicity, in short, is distinguished from the nation in terms of standardization.

When ethnicity and nation are understood in this way, a few theoretical and empirical implications emerge. First, the ethnic group can be an analytical building block to consider questions about the nation. This approach is useful not only for conceptual distinction but also for empirical analysis. Second, it renders the category—nation—more amenable to concrete observation, which opens a way for causal identification. One such strategy may be to identify the timing of standardization as one salient dimension of ethnic groups' cultural attributes and examine the mechanism by which standardization occurs.

Among the multitude of cultural practices, the standardization of the vernacular is the most important in specifying the mechanism through which ethnic groups become consolidated nations. The rationalization of non-linguistic dimensions, say an anthem, may precede that of linguistic ones, but enforcement would predictably be challenging in the absence of a shared means of communication. In contrast, once a vernacular is codified, this would drastically lower the cost of access to information, thereby making consolidation of the other dimensions of culture more efficient. In a classic study of cultural rationalization in Third-Republic France in the late nineteenth century, Eugen Weber (1976, 313) recounts the centrality of standardized language instruction at public school

[5] Tilly (1975, 1992) and Ertman (1997) are also associated with this interpretation, using the state as unit of analysis.

[6] That standardization is synonymous with institutionalization is consistent with the definition of standardization in linguistics. See Brown (2006, vol. 7, 121).

in the countryside. Weber vividly illustrates how a pupil who uttered words in an "unauthorized" tongue—*patois*—was chastised by having to hold a display showing a *faux pas* until the next student committed it. Further, a unitary language has a spillover effect in social organization. For instance, it makes the administration of military-related tasks such as recruitment, training, and the solidarity of personnel effective and therefore enhances warfighting capacity. Uniform language fulfills a variety of functions, not just as a chief repository of cultural knowledge but as a useful instrument for policy.

My focus on language is certainly not new. In fact, language plays a central analytic role in the literature on ethnicity and nationalism (Anderson 2006; Bell 2001; Gellner 2006; Kohn 2005). Convention holds that a nation may be observed when a government is conducted in the predominant tongue of its territory (Gellner 2006; Hechter 2000; Hobsbawm 1990; Tilly 1994). In this process, the literature indicates that fixing the vernacular is assumed to be critical to achieve what Wimmer calls "the rule of like over like" (2002, 213). Across Europe, language historically served as the most important cultural dimension with which to define membership to the given nation. The rise of modern citizenship, an institution in which language fluency is typically a requirement, is a good illustration.[7] Contemporary works continue to use language as the dependent variable to theorize about the variation in the salience of identity across time or within an ethnic group (Laitin 1998, 2007; Brubaker 2004) and the variation of "language regimes"—state policy on language instruction (Albaugh 2015; Cardinal and Sonntag 2015; Liu 2014; Safran and Laponce 2005).

However, a systematic investigation as to why language rationalization has become the default choice for ethnic groups seeking survival has not been undertaken. There is little doubt that the codification of the vernacular is, in theory, a highly costly project. It entails initially the classification of speech into main "trunks" and "branches" (as in the language trees), the transcription of speech into letters, the establishment of a grammar and related rules on usage, and finally the construction of an orthography, a set of rules about spelling. Once these rules are clearly written out and consistently

[7] Brubaker (1992) is most closely associated with this view.

used and taught among the speakers, a language may be said to have been "standardized." Publication of a dictionary may come at the very end of this lengthy process. Competent and devoted experts would be needed to bring the project to completion. It is easy to imagine that each step of the way is a labor-intensive, time-consuming, and financially demanding process. High fixed costs should make it clear that language standardization is not a "natural" choice. Yet, empirically, no ethnic groups that are well-known, such as the English, French, Italians, and Russians, fail to have their languages codified. Lesser-known ones, aspiring to survive, attempt to follow suit by consciously using their tongues in everyday communication and school instruction.[8] Two questions arise with respect to high fixed costs in labor and time on language rationalization. The first, theoretical question is: why do some ethnic groups make such a costly investment and fix their vernacular? The second is an empirical one: do hypotheses on the first question explain the variation of language standardization across ethnic groups? Addressing these questions fills an important lacuna in the study of ethnicity, nations, and ethnic diversity.

3.3 Print Technology and Language Standardization

3.3.1 The technological innovation

This paper hypothesizes the positive relationship between the printing press and language standardization. But it is crucial, first, to consider the technology's broad impact beyond language. The most general and profound effect that print technology brought to bear would be to reduce the cost of access to information (Mokyr 2006; Bernstein 2013). Metal movable typography, which was invented by Johannes Gutenberg among others circa 1450 in Mainz, Germany, greatly encouraged innovations in knowledge production on many scales—innovations occurred more cheaply, spread more widely and more quickly, and the incentives to produce grew more strongly.[9] The diffusion of print-

[8]Minahan (2000) offers an abundance of examples about the active use of the vernaculars by ethnic groups that constitute a demographic minority in their state of residence.

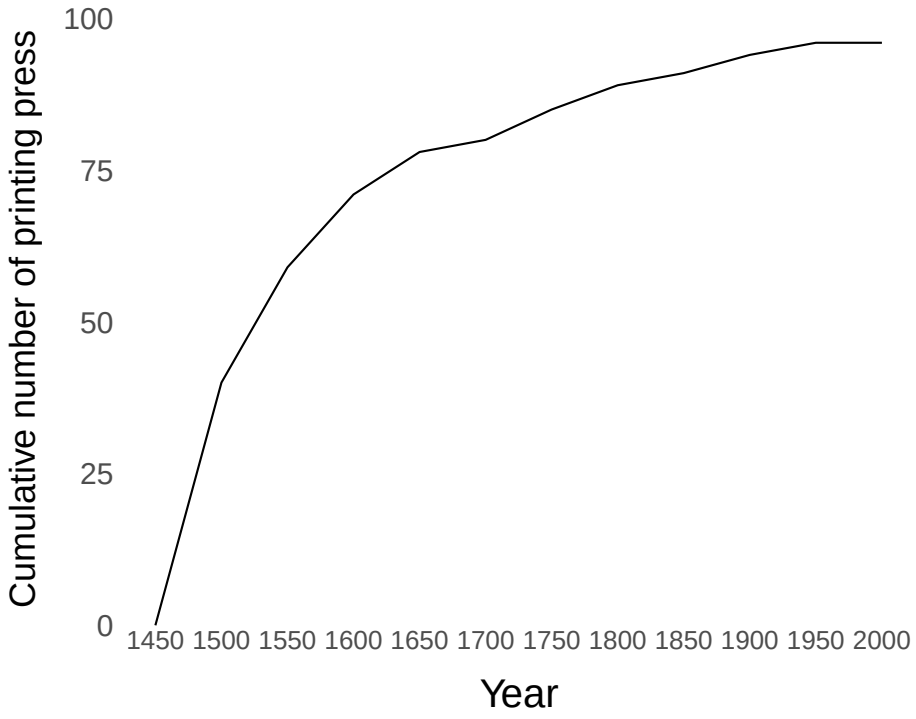
[9]Mokyr (2006, 1123–4) calls this phenomenon "propositional knowledge," the type of knowledge with which to describe, characterize, and theorize—that is, innovate. Examples are discussed in text. It is contrasted to "prescriptive

ing technology was remarkably swift by the fifteenth-century standard. More than 110 cities had a press established by 1480 (Febvre and Martin 1976, 182) and the number grew to over 240 by 1500 (Clair 1976). Thereafter book production surged dramatically. There were an estimated 5 million manuscripts produced during the fifteenth century in a dozen European countries, a 358-fold increase from the sixth century and the 82 percent increase from the fourteenth century (Buringh and van Zanden 2009, 416). Similarly, printed book production saw 6.3 times increase in the first half of the sixteenth century from twelve million books printed during the *incunabula* period of 1450–1500 (Buringh and van Zanden 2009, 417). In this period, the price of the book dropped by two-thirds (Dittmar 2011). Given that the price of books was, much like other commodities, determined by weight, lowered production cost was a remarkable improvement. Moreover, the technology significantly reduced person-hours to the extent that it pushed scribes out of business (Bernstein 2013). Figure 3.1 documents the rapid diffusion of printing technology among the European ethnic groups in my data set (described in detail in Section 3.4). Of the 96 groups that acquired the press, approximately half (41.7 percent) did so during the *incunabula* period.

---

knowledge," the type of knowledge to produce or invent. The movable-type printing press fits this category.

Figure 3.1: Cumulative number of printing press adoption among European ethnic groups.



Source: Clair (1976), Febvre and Martin (1976), among others.

3.3.2 The state's role in the spread of print and language standardization

The reduced cost of access to print media served as a catalyst for activities beyond book production.[10] The demand for the new technology came primarily from university instructors, clergy, lawyers, and wealthy merchants (Febvre and Martin 1976, 172–80). Yet the cost of operation remained high and left printers chronically short of capital (Febvre and Martin 1976, ch. 4). In addition, printers had to work in as many sectors as possible to uncover demand, because uncertainty about the emerging market on print media made it a highly risky venture (Pettegree 2010, ch. 3). For printers, finding

[10]Historical and empirical research shows that the printing press spurred commerce and economic growth (Dittmar 2011; McCusker 2005); literacy (Eisenstein 1979; Graff 1987; Buringh and van Zanden 2009); the Protestant Reformation (Rubin 2014); and democracy outside of Europe (Woodberry 2012).

a profitable market and sustaining it were equally critical.

States played little role in generating both supply and demand. In premodern Europe, industrial technology like the movable-type press spread primarily through skilled workers rather than states or related political channels (Cipolla 1972). Nevertheless, print technology had a ripple effect in politics by making the dissemination and enforcement of rules cheaper, because it proved useful for legal, legislative, and administrative purposes (Graff 1987, 109). There is some evidence that states took advantage of print media for two of their core functions: war and taxation. For instance, Maximilian I of the Holy Roman Empire issued propaganda broadsheets and pamphlets to raise manpower and revenue (Pettegree 2010, 132). At the same time, the state had the potential to induce supply of print by serving as a reliable patron for printers who always looked for one. Premodern European states chronically suffered from an unreliable flow of revenue due to a limited degree of institutional centralization (Dincecco 2015). All these rationales point to a favorable condition under which the state could employ printers as royal servants and create a win-win relationship. Such a relationship could be long-lasting, because revenue generation and law enforcement were permanent features of governance. Although printers called for monopoly rights of publishing for a more secure business environment, premodern European states remained unable to enforce regulations due to their limited institutional centralization (Pettegree 2010, 73). The evidence on the political use of the printing press by the state seems indirect at best in premodern times.

Equally important, independent statehood is not a prerequisite for language standardization. The development of the vernacular for the Bretons offers an illustrative case. Prior to French incorporation in 1532, there was no clear evidence of using Breton for official purposes during the five centuries of the independent kingdom (Price 1998, 36). However, beginning in the mid-fifteenth century the Breton language was substantially recorded in manuscripts, and the arrival of the printing press in Rennes in 1485 facilitated stability in language use. The reduced cost of printing led to the publication of the trilingual Breton-French-Latin dictionary in 1499 (the first printed book in Breton) and to the introduction of simplified spelling rules (orthography) in the mid-seventeenth

century (Price 1998, 36, 37). The greater supply of language use began in the early nineteenth century, when literates produced poetry, history, and novels in Breton, following the publication of grammar books and French-Breton dictionaries in the "purified" form by influential linguist Jean-François Le Gonidec (Minahan 2000, 131; Hardie 1948, 10). The separatist movement of the twentieth century, though failed, gave an impetus for further linguistic sophistication including spelling unification across the Breton-speaking region of France. Despite the legal exclusion of Breton use in the French school system for 1880–1951, Breton use became more widespread in journals and periodicals, culminating in the 1958 monolingual dictionary, *Geriadur istorel ar brezhoneg (Historical Dictionary of Breton)* (Dalby 1998, 64). The example of the Breton language demonstrates, first, that printing technology played a crucial role in stabilizing the vernacular and, second, that standardization is feasible without an independent polity.

Evidence from the literature on the cultural history of early-modern Europe indicates that the spread of the movable type and the development of print media relied less on state actors than private ones. Printers had to secure access to capital other than the public route, by creating a joint venture, locating private patrons, and relocating to high-demand cities like Wittenberg, the home of the Protestant Reformation (Pettegree 2010). Thus, the theoretical linkage between the printing press and language standardization is likely to be stronger outside the state.

3.3.3 Print and vernacular standardization

To begin, I discuss two ways in which the press gave rise to the greater use of the vernacular over Latin. First, printers themselves played a role in triggering the competition between vernaculars and Latin. As mentioned in the previous section, print was a high-risk business so that printers had to explore academic and private sectors to uncover high-demand publications and yield returns. As Anderson famously points out, after quickly filling the "thin" Latin market, these printers cultivated a "thicker" vernacular one in the lay public which could only understand the vernacular as no ethnic groups had Latin as a native language (Anderson 2006, 38). The growing popularity in the use of

the vernacular begot the "esotericization of Latin" (Anderson 2006, 42), which became increasingly certain after 1530, and the competition largely ended only eighty years following the invention of the Gutenberg press (Febvre and Martin 1976, 320).[11]

Second, access to print had an effect of stabilizing languages (McKitterick 1998, 296). This point is often overlooked or made in passing in the literature but is crucial for my argument. Once words are produced in printed form and widely circulated, they tend to acquire "staying power" in terms of consistency in usage. Furthermore, the technology of mass production enhanced scalability. One conspicuous dimension on which the printing press contributed to language rationalization is spelling. Prior to the invention of the press, spelling was primarily phonetic; thereafter, it became increasingly consistent as the transmission of words was based on not a human but a machine (Steinberg 1974, 125). Printers also played an influential role. When they processed text into publication, they simplified spelling at their own discretion to make their work more efficient (Eisenstein 1979, 87). The power of the press to fix language use attracted school teachers and priests who were concerned about what they deemed as the inconsistent or improper use of language and they subsequently became printers themselves.

What processes then account for the link between the greater use of the vernacular to codification? It is crucial to recognize that the invention of the technology did *not* intend to enable or facilitate the rise of the vernacular; standardization was an unintended consequence of it.

In the classic *Imagined Communities*, Benedict Anderson argued for the link between the profit motive of printers and booksellers ("print-capitalism") and national consciousness. He theorizes that "print-capitalism gave a new fixity to language, which ... helped build that image of antiquity so central to the subjective idea of the nation" (Anderson 2006, 44). This hypothesis describes the

[11] It is important not to attribute the decline of Latin solely to the printing press. As Ostler states, the press occasioned the "second death" of Latin. The "first death" occurred in the late eighth and ninth centuries during which Latin became less frequently used as the means of oral communication in favor of vernaculars (Ostler 2005, ch. 8). In addition, Latin was not completely displaced and continued to be used over the next few centuries. Febvre and Martin note that "the final blow was struck at Latin" around 1630 due chiefly to the decline in book trade but that Latin was not eliminated until the beginning of the eighteenth century (Febvre and Martin 1976, 330).

processes of language rationalization in general, but two points remain unaddressed. The first is to specify the causal mechanisms linking the press and language standardization. The second highlights the dimension of duration in the standardization process that is highly time-consuming and labor-intensive. The print-capitalism thesis holds that although the vernaculars won the competition with Latin relatively quickly, it does not consider the new race between vernaculars. The new rivalry lasts longer than that the old one. Latecomers in the game may be severely disadvantaged in that they miss the opportunity to standardize their tongue and are compelled to adopt similar ones that are better codified. These two points that are missing from the print-capitalism thesis merit further discussion.

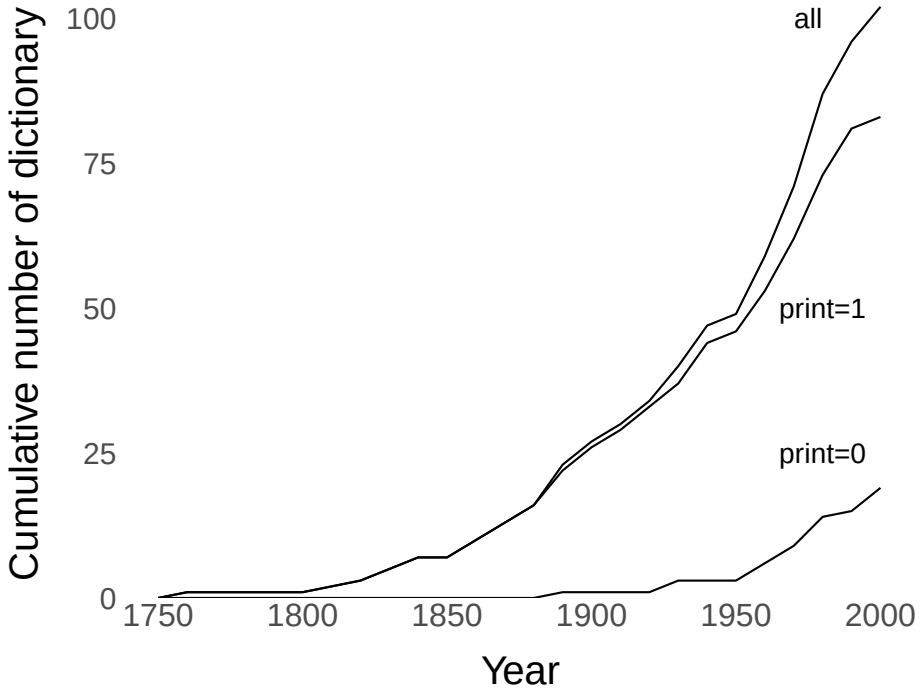
There are two plausible causal mechanisms from print technology to vernacular codification. The first is the selection process. If early modern printers established presses in homeland cities of some ethnic groups and started to produce books in their vernacular, the technology would likely prove to be useful later when language standardization became viable or necessary. Even though the original aim of print adoption was for printers' incentive to generate immediate cash, accessibility to the relevant technology gave a catalyst for other innovations in subsequent times. The state was clearly a beneficiary in this hypothesis. Although European states did not create strong demand for or supply of print, they benefited from the reduced cost of producing print media to facilitate the enforcement of rules and revenue generation. When states' investment in language rationalization is rewarded by the greater degree of fiscal and institutional centralization, other states may quickly follow suit to stay competitive. Similarly, the selection mechanism can also explain why demographically small ethnic groups have their tongues codified. As illustrated in the case of the Bretons, if there was a history of vernacular print, it could allow language entrepreneurs to revive or solidify their language to achieve national unity and increase the chances of group survival in modern times without an independent state. Thus, the selection process can account for the competitive and niche mechanisms.

The second process is by conscious choice. Actors may take advantage of the access to printing

technology using the vernaculars to advance their interests. Two paths merit attention. The first is the profit motive on the part of the printers. An increase in vernacular publications over those of Latin stirred new demand for translations. The translation business began to flourish in the early sixteenth century. Many print offices became workshops for translators busy churning out classical works in the vernaculars (Febvre and Martin 1976, 272). Ethnic entrepreneurs who want to advance learning in their native tongues, such as school teachers and university professors alike, would have an incentive to exploit the declining cost of becoming literate and producing books. Another path is Bible translation. Christian priests, both Catholic and Protestant, had an incentive to translate the Latin Bible into the local languages in the proselytization effort. The motive was particularly strong among Protestant reformers. They wrote books and translated Luther's catechism and other writings in many tongues. These were often the first printed vernacular works and proved to be a critical foundation for language standardization in subsequent centuries. The Slovene language fits this path. The first printed book in Slovene was a Protestant catechism, translated in 1551 by Primož Trubar, a Protestant preacher. It left the printed record of the language for writers in the nineteenth and twentieth centuries, who modernized Slovene by newspapers, books, and periodicals (Biggins and Crayne 2000). Given that literate ethnic Slovenes spoke German rather than Slovene which was a peasant tongue under centuries of Habsburg rule, the translation effort allowed the Slovene language to survive. Similar examples can be found in Estonian, Latvian, Livonian, Lithuanian, and Finnish (Steinberg 1974, 122). The conscious-choice process offers an alternative path to language standardization.

Figure 3.2 shows the cumulative publications of vernacular dictionaries between 1750 and 2000, with a subset data for those ethnic groups with a record of print adoption and those without it. It indicates that in either route, access to the printing press gives ethnic groups a strong advantage in consolidating the vernacular.

Figure 3.2: Cumulative number of vernacular dictionary publication among European ethnic groups.



Source: Burke (2004), Dalby (1998), Price (1998), among others.

3.3.4 The significance of the early adoption of the press

It is imperative to understand the timing of access to the printing press because language standardization is a slow process. It easily spans a few centuries, even under ideal circumstances in which an ethnic group had a state, resources, and institutions at its disposal specifically for developing its vernacular. Lexicography, in the monolingual edition, entails three steps: collecting words, "tokenizing" or making a tangible representation of words, and making an entry of these words (Brown 2006, vol. 7, 113). For each entry, lexicographers need to complete several tasks: (1) orthography (the rules about spelling); (2) guidance on pronunciation; (3) the classification of word-class (e.g., noun, verb); (4) definition making; (5) examples; (6) phraseology; (7) disputed points of usage; and (8) etymology and word histories (Brown 2006, vol. 7, 113).

This is an ideal level of progress for language standardization, because codified tongues play a central role in commanding authority, prestige, and legitimacy when ethnic entrepreneurs seek to mobilize the community and invoke a distinct, "national" identity. Language can have these functions as advocates for codification typically used the rationale to "purify" the vernacular, separated from the words which were characterized as "dirt" or "chaff," as their etymology could be traced to foreign tongues (Burke 2004, 144–50). Such mingling occurred because pre-standardized vernaculars were a mix of "indigenous" words which originated from the local language and "foreign" words which were imported from other locales due to trade or geographical proximity, or evolved from older languages like Latin. Standardization has this distinct attribute in comparison to other dimensions of language development such as grammar-making, which primarily concerns setting up the structural rules that govern the language.

The two processes I hypothesized earlier each have a distinct route to language standardization. First, the selection process stresses state capacity to promote vernacular codification. Ethnic groups with a sovereign state started in the late sixteenth century to establish a state-funded academy devoted to studying the vernacular, following the model of the Italian academy, Accademia della Crusca, founded in Florence in 1584. The state-centric approach is historically a phenomenon of continental Europe. Ethnic groups that adopted this model include the French, Spanish, Portuguese, Swedish, Russian, Danish, and others. Their centuries-long effort culminate in delivering monolingual dictionaries as the authoritative source of language use in the nineteenth and twentieth centuries (the Italian one published in 1861).

The second process focuses on actors' conscious choice to develop the vernacular. Given the high fixed costs of codification, ethnic groups have to secure access to patrons. Such an investment grants specialists time to concentrate on their project. The codification of the English language fits this model. Unable to secure royal patronage, Samuel Johnson managed to reach a contract with five booksellers, for the sum of 1,575 pounds, to cover the expenses (Brown 2006, vol. 6, 130).[12]

[12]Brown (2006, vol. 6, 130). Johnson opposed a state-funded English academy whose model was adopted in parts of continental Europe as contrary to "the spirit of English liberty" (Landau 2001, 54). The English founded an academy

Johnson's *Dictionary of the English Language*, published in 1755, has been widely regarded as the standard-bearer. This illustration points to a language standardization process alternative to the state-centric one.

My argument yields the following testable implications. I hypothesize that the variation on language standardization depends on the timing of the acquisition of the printing press for each ethnic group. Early adoption affords plenty of time to make literacy more accessible and develop a richer lexicography, by creating a specialized state agency or seeking patronage to cover the fixed cost of the codification effort. In contrast, late adopters of the press do not have this luxury of time. If they already possess a state and resources, they might not have as strong an incentive to codify their own vernacular as early adopters. Then latecomers could opt to accept the more developed language of one of their neighbors as a solution, which likely reduces the possibility to construct a distinct group of their own.

3.4 Empirical Strategy

To test my argument, I construct a new data set that contains information on the dimensions related to language standardization. It is unique in that I collect historical data on European ethnic groups as the unit of analysis. The data set is composed of 171 ethnic groups that exist today in Europe and has observations for the period between 1400 and 2000 C.E.[13] I draw primarily on Minahan (2000) for entries on these ethnic groups and add others referred to, but not entered, in Minahan's volume. I

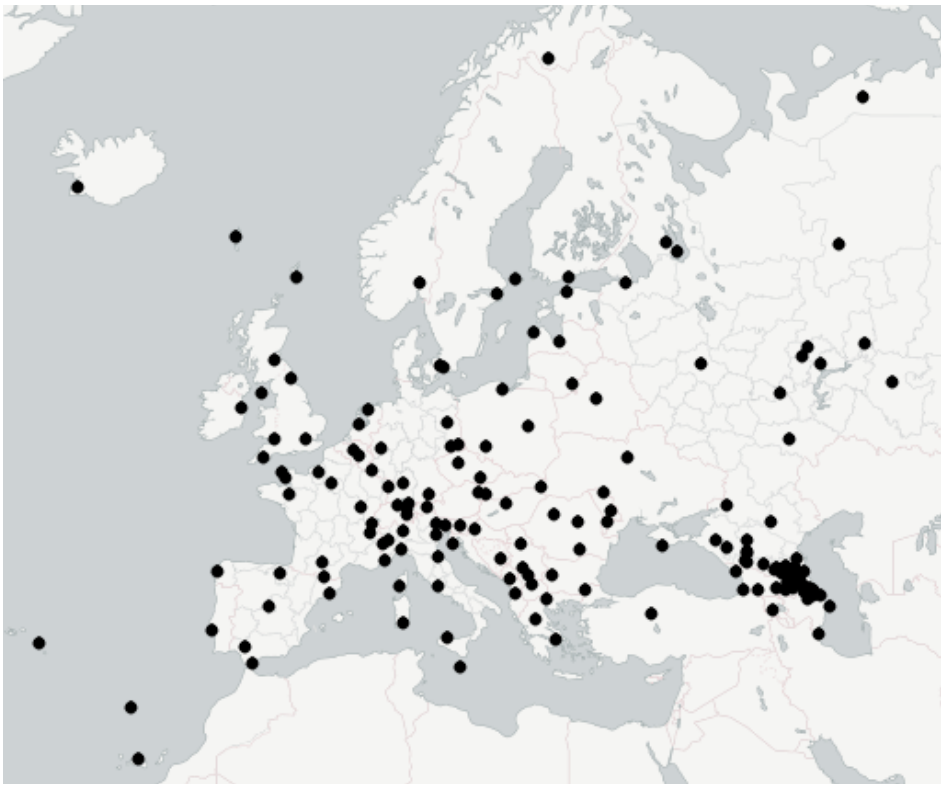
(the Royal Society of London for the Improvement of Natural Knowledge) in 1662 and did launch a committee that addressed the grammar and orthography. But the main mission of the academy remained scientific and the committee on language met only four times and disbanded (Martin 2008, 225).

[13]I follow Price (1998) who relies on the geographical, geological, and historical criteria to define "Europe." Europe geographically covers land from Iceland as the westernmost border to the Ural Mountains in Russia as the easternmost border. The southeastern limit is at the Caucasus Mountains, including Georgia, Armenia, and Azerbaijan but excluding Iran. Turkey is included. Two islands offshore Portugal, the Azores and Madeira, are included because Azores is geologically on the ridge of the European continent; while Madeira is geologically on the African continent, scholarly convention is to include it (for the cultural reason that Portuguese is the only language spoken there). Likewise, Malta is included for geological and practical reasons, although Cyprus is not, because geographically it is closer to Syria, a non-European country, than to Europe.

double-check them using *Ethnologue* compiled by Lewis, Simons, and Fennig (2013). The list of 171 ethnic groups, which provides the data on the homeland, the explanatory variable, and the outcome variable, is available in Appendix B. The starting point is the year 1400, as most explanatory variables in my data set appear after this date. I record data at the fifty-year interval, unless otherwise noted. To enable the observation of covariates at the ethnic-group level, I locate a "homeland" city for each group and find information on political, economic, social, and geographical dimensions for these cities. I regard these observations as specific attributes for ethnic groups.[14] To date, this is one of the most detailed data sets on the subject of European ethnicity and nationalism. Figure 3.3 displays the geographical location of the homeland cities for 171 European ethnic groups.

[14]This coding strategy is both theoretical and practical. Theoretically, virtually all ethnic groups identify a city as their unique "homeland" which is shared with no other groups. Ethnic groups typically base their national movement in the homeland to claim territorial autonomy (Smith 1986). Practically, treating ethnic groups to be discrete allows me to capture a geographical impact on them. I mainly rely on Minahan (2000) for the entry of the ethnic homeland.

Figure 3.3: Geographical distribution of the homeland cities for 171 European ethnic groups.



Source: Minahan (2000) and Lewis, Simons, and Fennig (2013).

The outcome variable is the timing of language codification. As I defined "nation" earlier by the timing of the standardization of cultural attributes for ethnic groups, this coding strategy allows me to capture the standardization process of a core attribute of ethnicity—language—and draw implications for how ethnic groups consolidate their cultural practices. I operationalize the timing of language standardization by the first publication year of a comprehensive vernacular dictionary for each ethnic group. The qualifier "comprehensive" refers to a "modern," monoglot dictionary that aims to cover most, if not all, words in the alphabet of the given language, and one judged as setting a linguistic standard by scholars in relevant fields such as lexicography, linguistics, and cultural history. As Landau (2001) points out, these comprehensive dictionaries are distinct in purpose,

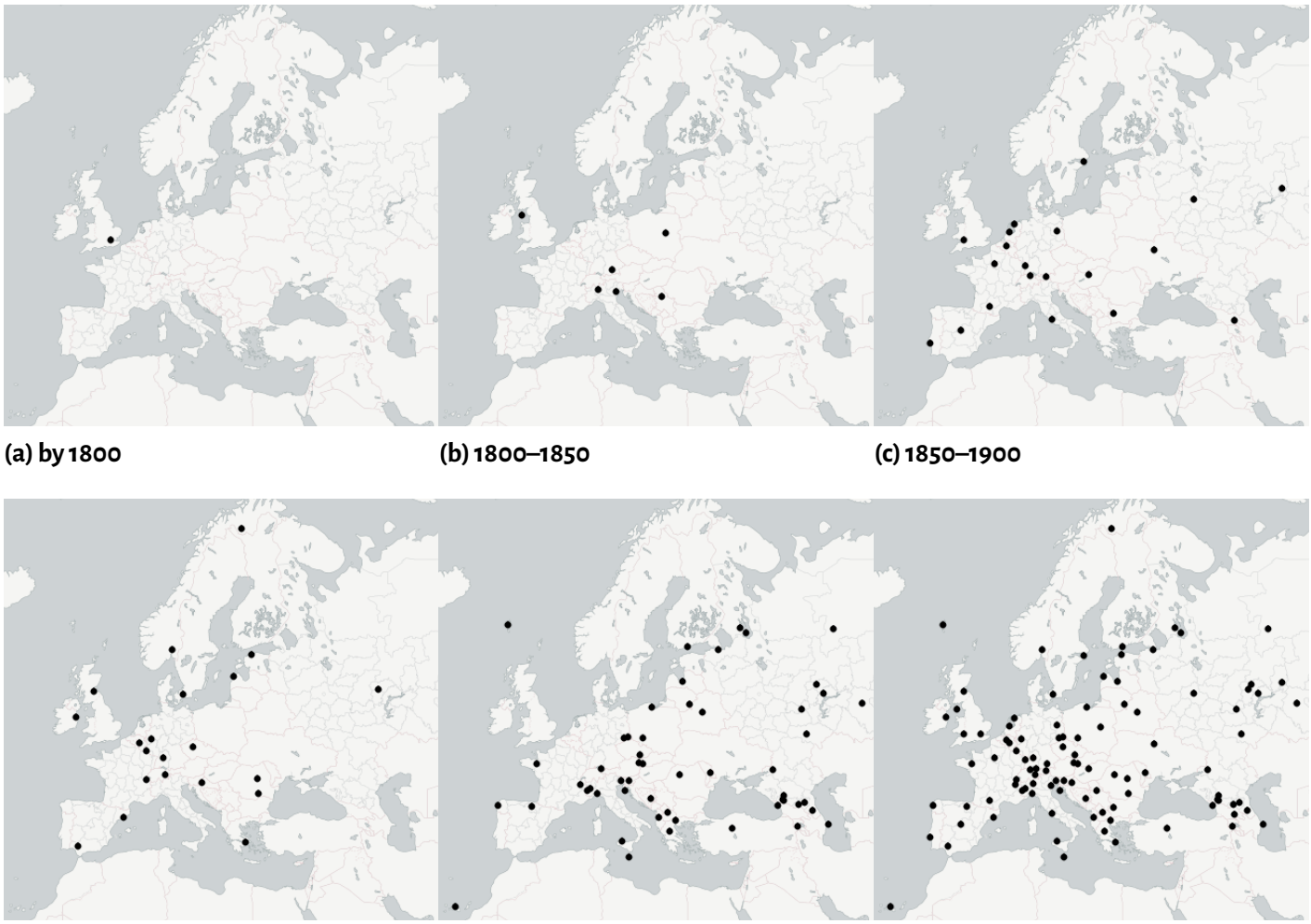
substance, and scope from glossaries and encyclopedia, although the latter may often bear the label "dictionary" in the title.[15] In a similar vein, this qualifier excludes polyglot dictionaries such as the premodern translation works between a vernacular and Latin. Wherever possible, I try to find a dictionary specifically complied for minority ethnic groups whose vernaculars today may be classified as "dialects." The dictionary data rely mainly on Burke (2004), Dalby (1998), and Price (1998). For some ethnic groups in the Caucasus, I use Nordhoff et al. (2013). As Table 3.1 shows, only a quarter of the observed data (n = 26) achieved standardization by the twentieth century, while the rest codified their tongue within the last century, especially the latter half. The data suggests that language standardization is a modern phenomenon. For many ethnic groups that are demographically small and stateless, it is still an ongoing project. Figure 3.4 exhibits the geographical distribution of vernacular dictionaries between 1800–2000.

[15] A glossary usually contains an alphabetical list of terms used in a specific topic. An encyclopedia is a collection of entries on a given branch of knowledge.

Table 3.1: Summary statistics.

	N	mean	median	sd	min	max
Outcome variable						
Year of vernacular dictionary	104	1935	1954	51	1755	2005
Explanatory variable						
Year of printing press adoption	96	1566	1526	115	1460	1930
Control variables						
Year of vernacular Bible	144	1789	1837	193	1380	2012
University	67	1672	1735	208	1150	1925
Bishop	99	1250	1400	608	100	2000
Elevation (m)	171	405.3	145	581	−24	2,823
Terrain ruggedness (m)	171	137.51	59.87	205	0	1,125
Urban potential	171	8.43	2.89	14.99	0.35	99.45
War frequency (in 50-year periods)	171	8.05	5	8.75	0	37
Distance from Wittenberg (km)	171	1,669	1,429	948	89	5,292
Distance from Zürich (km)	171	1,711	1,516	1,048	0	3,350
Distance from Mainz (km)	171	1,752	1,509	1,078.7	140	5,545
Hub Roman road	171	0.12	0	0.33	0	1
Roman road	171	0.32	0	0.46	0	1
Oceanic port	171	0.38	0	0.49	0	1
Island	171	0.1	0	0	0	1
Russian Empire/Muscovy	171	0.15	0	0.36	0	1
Ottoman Empire	171	0.105	0	0.31	0	1
Habsburg Empire	171	0.08	0	0.27	0	1
Orthodox	171	0.16	0	0.37	0	1
Catholic	171	0.35	0	0.47	0	1
Protestant	171	0.17	0	0.1	0	1
Islam	171	0.29	0	0.45	0	1
Western Europe	171	0.16	0	0.36	0	1
Northern Europe	171	0.14	0	0.35	0	1
Southern Europe	171	0.16	0	0.40	0	1
Eastern Europe	171	0.40	0	0.49	0	1
Western Asia	171	0.10	0	0.30	0	1

Figure 3.4: Geographical distribution of vernacular dictionaries 1800–2000.



(a) by 1800 **(b) 1800–1850** **(c) 1850–1900**

(d) 1900–1950 **(e) 1950–2000** **(f) the whole period**

Source: See Figure 3.2.

The main predictor is the timing of acquisition of the movable-type printing press. I operationalize it by recording the first date that the technology arrived in the homeland city for each ethnic group. In this process, I make a distinction between the year when the Gutenberg press was adopted in a homeland and the year when vernacular books were printed. These two sets of years

are not always identical and the latter typically occurs later. In this paper, I choose the former because one of the empirical goals is to assess the effect of accumulated time of technological access for a given ethnic group. The print variable comes from various sources but primarily from Clair (1976) and Febvre and Martin (1976), which cover Western, Southern, and Central Europe.[16] For the ethnic groups in the German-speaking area, I use Reske (2007).

I assess the relationship between the printing press and vernacular dictionary by taking the following steps. First, I evaluate the demand-side mechanism, in which social and economic developments shape demand for the printing press. As discussed earlier, printers were essentially capitalists who were willing to go to locales likely to yield returns on their investments. This mechanism captures whether such pre-Gutenberg press activities determine the acquisition of the press and subsequent language standardization. Drawing from recent research in economic and political history, I control for the following set of variables. The first is the university. Universities would benefit from the ability to mass-produce books and other printed material to promote literacy. I rely on Darby and Fullard (1970) and Rüegg (four volumes, 1992–2011) on the history of European universities to obtain the founding date. The second control is the bishoprics. Like universities, the printing press would make the proselytization effort much easier and efficient with the capacity to print pamphlets, posters, and booklets on a large scale. I use Chaney (2015) for the establishment year of a diocese or archdiocese. For the university and bishopric, I record the year of foundation. The third is the urban potential. It is originally constructed by de Vries (1984) as the sum of the populations in the given period divided by the geographical distance between a city and all others in my data set.[17] This variable gives a sense of whether a city is surrounded by competing urbanizing towns or located in a more sparsely-populated area. Higher values indicate greater potential for urbanization. The standard source on population size in preindustrial Europe is Bairoch, Batou, and Chévre (1988), which covers the period 800 through 1850 for hundreds of European cities. Bosker, Buringh, and

[16] I also use Burke (2004), Price (1998), and Steinberg (1974) as supplementary sources.

[17] More formally, this variable is obtained by $\sum_{j=1}^{N} \frac{\text{population}_j}{\text{distance}_{ij}}$.

van Zanden (2013) correct some of the data in Bairoch et al. and I follow their updates in compiling my data. For the period after 1850, I use several statistical handbooks including Mitchell (2003*a*,*b*). The population data are available at the hundred-year interval.

The second set of controls concerns the supply-side mechanism. These variables can accelerate or delay language rationalization, depending on how a particular variable works. A major covariate in this domain is war. Early-modern European history is characterized by recurring warfare, in which growing costs of fighting and preparing for war largely determined how to organize state entities most effectively (Bean 1973; Tilly 1975, 1985, 1992). A unitary language may emerge as desirable in this process. If tax collectors and subjects who are taxed communicate in a mutually intelligible tongue, this renders the administration of raising resources and manpower more efficient. War can serve as a catalyst that spurs the incentive for language rationalization. I draw on the database compiled by Brecke (1999) for war-related data. It records any conflict in the world with the minimum of 32 casualties in the period between 1400 and 2000. As Brecke uses states as the unit of analysis, I take care to localize the incidence of war at the ethnic-group level to the extent possible. For instance, if a war occurred in the Habsburg Empire, I regard that war as affecting all ethnic groups within the imperial domain. However, my motive here is to capture the institutional effect of war on taxation and governance. Therefore, if no mention is made about the place of a given war, I determine that war uniformly affects the residents in the country. If, for example, a war took place in Scotland prior to union with England, I regard it as having an impact on the Scots but not others in their neighbor (the English, Welsh). I construct war frequency, which measures how often an ethnic group experiences war in any fifty-year period, to capture war's impact.

Another supply-side covariate is overseas trade. If an ethnic group's homeland city is located on or near the coast, it enables oceanic trade that can promote language standardization for greater efficiency. In European history, geography played an important role in economic activity.[18] Because

[18]See Acemoglu, Johnson, and Robinson (2005) and Bosker, Buringh, and van Zanden (2013) for the importance of geography on long-term economic growth. See Stasavage (2010) for the role of geography on the frequency of parliamentary sessions.

roads for overland trade were not well-paved and thus proved unreliable, access to ports provided a critical precondition for economic growth. This access, in turn, may provide incentives for a unitary language. I use Nüssli (2011) to produce an indicator taking the value of 1 if an ethnic group's homeland lies on an oceanic coast.[19] While access to trade is favorable to growth, other geographical conditions may have the opposite effect. In particular, "bad terrain" may prove prohibitively costly for undertaking vernacular codification. To assess this impact, I include a series of time-invariant measures. One is a set of variables such as terrain ruggedness and land elevation above the sea level for each ethnic group. I obtain these observations from the GLOBE (Global Land One-kilometer Base Elevation project) database (GLOBE Task Team and others 1999).[20] Related is a measure on island, an indicator taking the value of 1 if an ethnic group's homeland is on an island. All these geographical variables are intended to capture different dimensions of geography's impact.

In addition, I include indicators on the Russian, Ottoman, and Habsburg Empires based on Nüssli (2011). They take the value of 1 if an ethnic locale was under any of these polities at the beginning of the century. These are designed to capture the institutional impact on the acquisition of the printing press. Historians indicate that Russians and Ottomans in particular had centralized control over private, vernacular print until the eighteenth century.[21] Thus it is expected that the ethnic groups under Russian or Ottoman rule would be late in adopting the printing press and consequently in standardizing their vernacular.

Finally, I use two indicators to capture potential long-term institutional effects. One is a set of variables on the Roman Empire's influence. Romans built roads across the cities they reached, which gave these cities an opportunity to develop institutional capacity in economic activity and political organization. I use Talbert (2000), Hammond (1981), Åhlfeldt (2015), and *Pleiades* (2015) to collect

[19] I exclude a similar indicator on the river, as virtually all ethnic groups founded their homeland near the river.

[20] Values for terrain ruggedness in the GLOBE data set come from the Terrain Ruggedness Index, originally devised by Riley, DeGloria, and Elliot (1999).

[21] An exception includes some cities in the west of Istanbul, where Jewish printers were allowed to set up non-Turkish and non-Arabic presses, and some ethnic groups in the Balkans, where Ottoman rule was relatively weaker and religious legitimacy was less effective than in the Anatolian heartland (Coşgel, Miceli, and Rubin 2012).

information about the Roman legacy. More specifically, I create an indicator taking the value of 1 if a city had major or minor Roman roads.[22] Another institutional effect I control for is the impact of the Protestant Reformation. The data set includes two measures. One is the the measure on geographical distance to Wittenberg or Zürich, two crucial cities in understanding the movement. Following Pfaff and Corcoran (2012), I construct this variable by taking the nearest distance between the given ethnic homeland and either of the Protestant centers. It is important to note that this variable captures not the impact of the confessional movement but the magnitude of the religious reform on institutions. One of the Reformation's consequences is that cities and other localities were forced to enact institutional reforms to reduce tension after being forced to choose a side. One example is education reform for the upper-class and university education (Gorski 2003, 19). This implies that the Reformation could promote vernacular literacy and demand for a linguistic standard. Proximate location to Wittenberg or Zürich indicates a greater impact of such institutional reform. The second measure on the Reformation is a fixed effect on the predominant religious preference (Catholicism, Protestantism, Orthodoxy, or Islam) for each group.[23] Geographical indicators for the region are also included in the data set.

3.5 Estimation Results

3.5.1 Baseline regression

I begin by documenting the baseline correlation between the printing press (the explanatory variable) and vernacular dictionaries (the outcome variable). First, the bivariate relationship is shown in Table 3.2. Of the 171 observations, approximately a half (48.5 percent) acquired the printing press and achieved vernacular codification. For those ethnic groups that have the recorded date of print adoption, 86 percent standardized the vernacular by 2000 (the cutoff year for this study). By

[22] Following Bosker, Buringh, and van Zanden (2013), those cities with two or more major roads are coded "Roman hub" and those with one major road or one or more minor roads are coded cities with "Roman roads."

[23] For those ethnic groups that adopted Protestantism after the Reformation (before 1500), I code them to be associated with Catholicism.

contrast, of those observations without access to the technology, 72 percent did not achieve standardization by 2000. It is important to highlight that language standardization is an ongoing project. Although 67 groups did not publish a monolingual dictionaries and are defined as "unstandardized," they may be able to codify their vernacular in the future.

Table 3.2: Two-by-two table on the relationship between print technology and vernacular dictionary (n=171)

		Language Standardization	
		1	0
Printing Press	1	83 (48.5%)	13 (7.6%)
	0	21 (12.3%)	54 (31.6%)

Second, I discuss the timing of printing press adoption. Of the 83 ethnic groups that got the press and language standardization, elapsed time between the two is, on average, 363 years (the median is 396 years). There is a huge lag because 75 percent of those groups that acquired the press did so by the early seventeenth century, while 75 percent of those that standardized language did so in the twentieth century. These two pieces of descriptive evidence provide preliminary support for my hypotheses that the adoption of the printing press is positively linked to language standardization and that, given this lag, the early adoption of presses is crucial to it.

To test my argument more systematically, this paper employs the following estimation strategies. First, I use the Cox proportional hazard model to examine the effect of time-varying covariates on vernacular codification. The key advantages of the Cox model include the assumption that the baseline hazard rate does not follow a particular distribution. Instead, the duration times are parameterized in terms of a given set of covariates (Box-Steffensmeier and Jones 2004, 49). The positive coefficients are interpreted as increasing hazards for an event of interest as a function of the covari-

ates, namely increasing the chances of language standardization for ethnic groups. By contrast, the negative coefficients are interpreted as decreasing hazards, meaning that the chances of language rationalization decrease. More specifically, I estimate the following reduced-form model:

$$h_i(t) = h_0(t)\ \exp(\beta_1 \text{Press}_i + \gamma_1 X_i + \delta_i).$$

Here $h_i(t)$ is the codification of vernacular dictionaries for ethnic group i at time t. This event of interest is parameterized by whether ethnic groups acquired the printing press (β's), a vector of covariates γX_i, and a set of fixed effects δ for empires, geography, and religion. $h_0(t)$ is the baseline hazard but dropped in estimation. In addition to the Cox model, I use logistic regression and estimate the same specifications as robustness checks.[24]

Table 3.3 documents regression estimates for the Cox and logistic models, each in four sets of specifications: (1) the bivariate model; (2) the model with demand-side covariates; (3) the model with supply-side covariates (the demand-side ones included); and (4) the fully-specified model that includes all the controls and fixed effects.[25] The bivariate model shows that if ethnic groups acquire a printing press in the homeland, the chances that they standardize the vernacular is 8.66 times greater than those without access to the technology (the value of the hazard ratio obtained by exp(2.159) in Model 1). The result holds when the demand-side covariates are introduced. The magnitude of the hazard ratio gets attenuated for the press (to 4.95 in Model 3), as the founding of a university is also significantly positively correlated with language standardization. Other specifications exhibit a similar pattern. The supply-side scenario indicates that rough terrain and war are expected to have a significantly negative impact on language standardization, although the substantive effect of the press remains approximately the same at 4.29 (Model 5). Finally, the magnitude of the press stays constant when a host of fixed effects on imperial rule, religion, geographical region. In all specifications, the adoption of print technology is significantly positively correlated with vernacular codification and the result is robust to the inclusion of a host of covariates and fixed effects.

[24]The logistic models include the time trend and its squared function as suggested by Carter and Signorino (2010).

[25]The complete estimates discussed in this paper all are reported in Appendix 3.

Table 3.3: Regression outputs of the printing press' impact on language standardization.

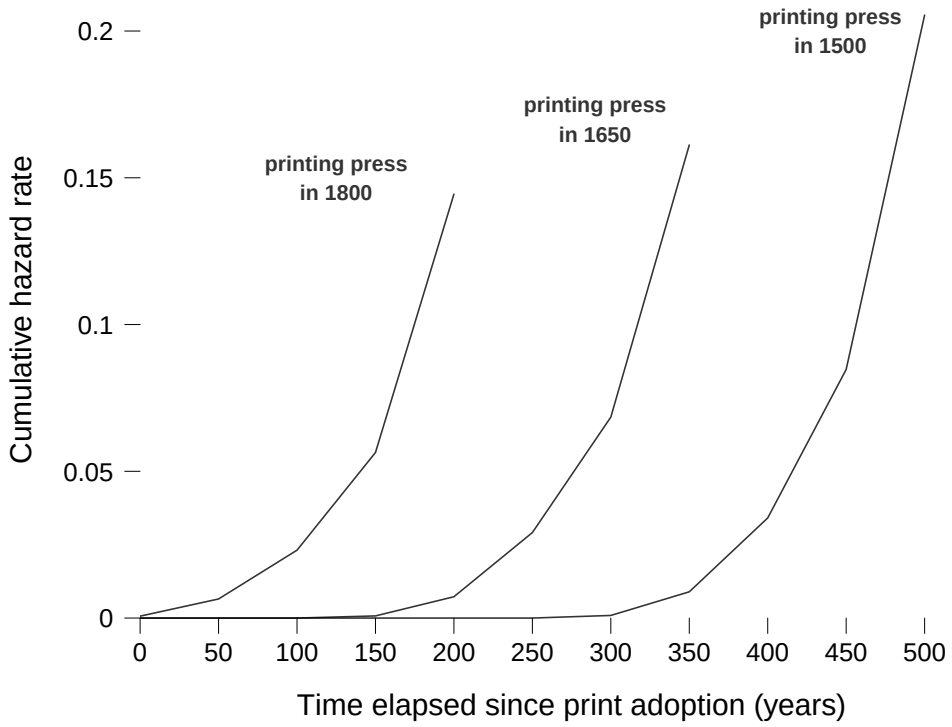
	bivariate		demand side		supply side		fully specified	
model	Cox	logit	Cox	logit	Cox	logit	Cox	logit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Printing press	2.159***	2.933***	1.600***	2.178***	1.455***	1.932***	1.450***	2.083***
	(0.269)	(0.342)	(0.356)	(0.456)	(0.354)	(0.503)	(0.339)	(0.556)
University			0.0004***	0.001***	0.0004***	0.001***	0.0004***	0.001***
			(0.0001)	(0.0002)	(0.0001)	(0.0003)	(0.0001)	(0.0002)
Bishop			−0.000002	−0.0001	0.00003	0.00004	−0.00002	−0.00002
			(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.0003)
Log urban potential			0.114	0.098	−0.002	−0.307	0.186	−0.276
			(0.226)	(0.365)	(0.607)	(0.793)	(0.636)	(0.821)
Log distance to Wittenberg/Zürich					0.0001	−0.0001	0.0004	0.0002
					(0.0003)	(0.001)	(0.0003)	(0.001)
War frequency					−0.031**	−0.043***	−0.030**	−0.048***
					(0.014)	(0.016)	(0.013)	(0.018)
Elevation					−0.0001	−0.00002	0.0001	0.0005
					(0.0004)	(0.001)	(0.0004)	(0.001)
Terrain ruggedness					−0.003**	−0.004**	−0.002**	−0.004**
					(0.001)	(0.002)	(0.001)	(0.002)
Hub Roman road					0.156	0.494	0.129	0.431
					(0.296)	(0.634)	(0.337)	(0.7)
Roman road					−0.009	−0.023	0.081	0.151
					(0.247)	(0.467)	(0.304)	(0.603)
Oceanic port					−0.415**	−0.712*	−0.477*	−0.721
					(0.212)	(0.411)	(0.26)	(0.539)
Island					−0.057	0.185	−0.038	0.2
					(0.475)	(0.85)	(0.52)	(0.965)
Empire fixed effects	No	No	No	No	No	No	Yes	Yes
Religion fixed effects	No	No	No	No	No	No	Yes	Yes
Geographic region fixed effects	No	No	No	No	No	No	Yes	Yes
Time dummies		Yes		Yes		Yes		Yes
Observations	2,052	2,052	2,052	2,052	2,052	2,052	2,052	2,052

Note: Robust standard errors clustered by ethnic groups for all models. Intercept is not reported. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects so they are omitted. Full results are reported in Appendix B. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

A second hypothesis from my argument is that the early acquisition of the printing press is crucial for the early codification of the vernacular. To test this proposition, I create counterfactual scenarios where ethnic groups acquired the press at the different timing and estimate the cumula-

tive hazard for language standardization. I set three scenarios in the 150-year interval, 1500, 1650, 1800 C.E., and report in Figure 3.5 the estimated cumulative hazard from the saturated Cox model. It shows that the advantage of early adoption become discernible over time. The cumulative hazard rate reaches 0.2 for early adopters (presses by 1500), 0.16 for "mid"-adopters (presses by 1650), and 0.14 for late adopters (presses by 1800). Though the differences in magnitude are not large, the simulation offers two insights. First, early technology adoption can trigger knowledge accumulation. Vernacular printing leaves a written record of cultural practices for ethnic groups. The stock of knowledge in a unique language grows over time through access to concrete, verifiable information in printed material (Mokyr 2002, 2006). Printing technology can facilitate cultural innovation by lowering the costs of writing novels, mythologies, and history, which in turn consolidates language. Second, this simulation suggests that increasing efficiency in technology may not permit quick catch-up in cultural rationalization for late adopters. Although the costs associated with printing were expected to have reduced over this period, the challenge of assembling knowledge on culture does not appear to be quickly overcome. Estimates from these scenarios indicate that the timing of acquiring the enabling technology matters for understanding language standardization.

Figure 3.5: Differences in the timing of printing press adoption on language standardization.



3.5.2 Robustness checks: The role of the state

Regression estimates so far support my hypotheses that the adoption of printing technology is positively correlated with language standardization and that the timing of adoption is crucial. However, there are two endogeneity concerns to my causal claims. One is the impact of independent statehood. It is clearly of the selection process, in which some ethnic groups might have been disproportionately endowed with resources or capacity that are related to building their own state and are unobserved by my data set. This process could have begun before 1400 C.E., the year when my analysis time starts, and print adoption is endogenous to this process. For instance, although the literature suggests that the spread of the press depended primarily on skilled workers (Cipolla 1972), an early start on building state capacity in taxation or rule enforcement can determine the

demand for the printing press, because the technology can make governance more efficient. It can thus be imagined that the acquisition of print technology is a function of having an independent state. To address this concern, I employ the following empirical strategy. I begin by constructing an alternative data set with the state as the unit of analysis (n = 47). Within this sample, I recode the printing variable by identifying its first adoption year for any city in each state. I then reestimate the impact of printing technology on vernacular codification. For states with multiple official languages, I use the first publication date of a vernacular dictionary for each tongue and use it as an alternative outcome (my list has up to four official languages). Appendix 3 offers the list of the new sample with key variables. Table 3.4 presents the Cox regression estimates for the state sample.

Table 3.4: Cox proportional hazards model outputs of the printing press' impact on language standardization.

sample	state				stateless
model	List 1	List 2	List 3	List 4	
	(1)	(2)	(3)	(4)	(5)
Printing press	14.711***	14.984***	14.725***	14.814***	4.915**
	(0.98)	(0.92)	(1.1)	(0.97)	(2.293)
All controls	Yes	Yes	Yes	Yes	Yes
Observations	564	564	564	564	1,488

Note: Robust standard errors clustered by country for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects and are omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

All models include the covariates from the fully-specified model. The impact of the first printing press on language standardization remains significantly positive. The magnitude of coefficients is stable across alternative lists of official languages and much greater than that in the ethnic-group sample. This is unsurprising, because the vast majority of 47 states have the recorded date of print

adoption and have standardized the vernacular. At the same time, if unobserved state-related forces determine language standardization, the impact of those ethnic groups without access to the state is expected to be statistically indistinguishable from zero. Column 5 of Table 3.4 reports estimates of Cox regression for these groups (n = 124). Although the substantive effect of print technology is much smaller than in the state-only sample, the result remains the same. This simulation suggests that while the selection process in which the adoption of the press is a function of having a state may be at work, the unobserved state capacity is unlikely to determine the chances of language standardization.

3.5.3 Robustness checks: Human capital's impact

The second endogeneity concern is the influence of human capital on the development of the vernacular. While the urbanization potential is controlled for as a proxy for economic growth, a more specific channel through human capital, or individual-level literacy, may account for language standardization. In the literature on economic history and on modern-era growth, strong cognitive skills can enhance economic well-being.[26] Increase in literacy is hypothesized to have a spillover effect on an investment in the vernacular for greater efficiency in communication, writing, and business transactions within ethnic groups members. In the context of early-modern Europe, a major driver for human capital is the Protestant Reformation (Becker and Woessmann 2009, 2011; Boppart, Falkinger, and Grossmann 2014; Cantoni, Dittmar, and Yuchtman 2016).[27] The religious movement prodded lay followers to read the Bible in the vernacular. Literate reformers also issued church ordinances to reinforce lay literacy (Dittmar and Meisenzahl 2016). Recent research demonstrates that an (early) access to the printing press played a critical role, because campaigners of the religious reform would take advantage of the technology's capacity to mass-produce, disseminate

[26] For general works on the role of human capital in growth, see, for instance, Easterlin (1981), Goldin (2001), Lindert (2003), Glaeser and Saiz (2004), Algan and Cahuc (2014), Hanushek and Woessmann (2008).

[27] Becker, Pfaff, and Rubin (2016) overviews major theories and empirical works. For a negative association between the religious reform and growth, see Cantoni (2015).

information, and canvass support through the vernacular Bible, ordinances, and broadsheets (Rubin 2014). If the human capital hypothesis is correct, ethnic groups that observe Protestantism as the primary religion are expected to exhibit high literacy, which in turn is positively correlated with language standardization. The adoption of the printing press may be endogenous to this process.

This paper has already addressed the impact of the Protestant Reformation on print and vernacular codification by including the religion fixed effects and taking the shorter distance from an ethnic group homeland to either Wittenberg or Zürich. Yet these measures may be too broad to capture the human capital channel precisely.

To account for this channel, I construct the vernacular Bible indicator. It is an appropriate proxy for a source of human capital in early-modern Europe, because it captures the Protestant advocacy of lay literacy built on the access to printing technology. This variable takes the first publication date (year) of the vernacular translation of the Latin Bible drawn from *Ethnologue* which is compiled by Lewis, Simons, and Fennig (2013). The data is also supplemented by Price (1998). As with the dictionary publication, the year 2000 is used as the cutoff. My strategy is, first, to estimate the impact of the vernacular Bible in the fully-specified Cox and logistic models *minus* the print variable (i.e., Columns 7 and 8 in Table 3.3, respectively). Second, I reintroduce the print variable with the Bible variable included. If the human capital channel determines the chances of language standardization, the vernacular Bible should be positively associated. Moreover, the substantive impact of the printing press is expected to be statistically indistinguishable from zero, while the Bible's effect should be retained.

Table 3.5: Regression outputs of the impact of human capital and printing technology on language standardization.

model	Cox	logit	Cox	logit
	(1)	(2)	(3)	(4)
Printing press			1.394***	2.023***
			(0.342)	(0.558)
Vernacular Bible	0.531**	0.745**	0.461*	0.676*
	(0.251)	(0.356)	(0.257)	(0.365)
All controls	Yes	Yes	Yes	Yes
Observations	2,052	2,052	2,052	2,052

Note: Robust standard errors clustered by ethnic groups for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects and are omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 3.5 documents the role of human capital in language standardization. As before, all models are fully-specified ones including the fixed effects. Table 3.5 shows that the vernacular Bible is significantly positively correlated with vernacular codification in all models. In Model 1, ethnic groups with the Bible have 1.7 times greater chances of language standardization than those without it. When the press is included in Model 3, the magnitude gets attenuated (with the chances of the event by 1.59 times) but stays positive. Yet printing technology exhibits the much greater magnitude in the same model, in which the press increases the chances of language standardization by 4 times. Even when compared to fully-specified Model 7 of Table 3.3, the hazard ratio for print drops only slightly by 0.26. Although the human capital channel may account for why ethnic groups standardize the vernacular, the printing press hypothesis remains robust and exhibits a greater impact.

To disentangle the regression results from Table 3.5, it is useful to revisit the sequence of historical events. The invention of the movable type *preceded* not just the Protestant Reformation but also Bible translation for virtually all cases in my observations (with the exception of the Czech and

English translation of 1380 and 1382, respectively). Many European ethnic groups enjoyed the vernacular print earlier than the translation of the Bible promoted by Protestants. Model 3 of Table 3.5 shows that the changes in the hazard ratio of the press is 2.53 points greater than those of the Bible, pointing to the importance of *earlier* adoption of the press for European ethnic groups. The chronological order matters when considering the differences in the magnitude of these different channels.

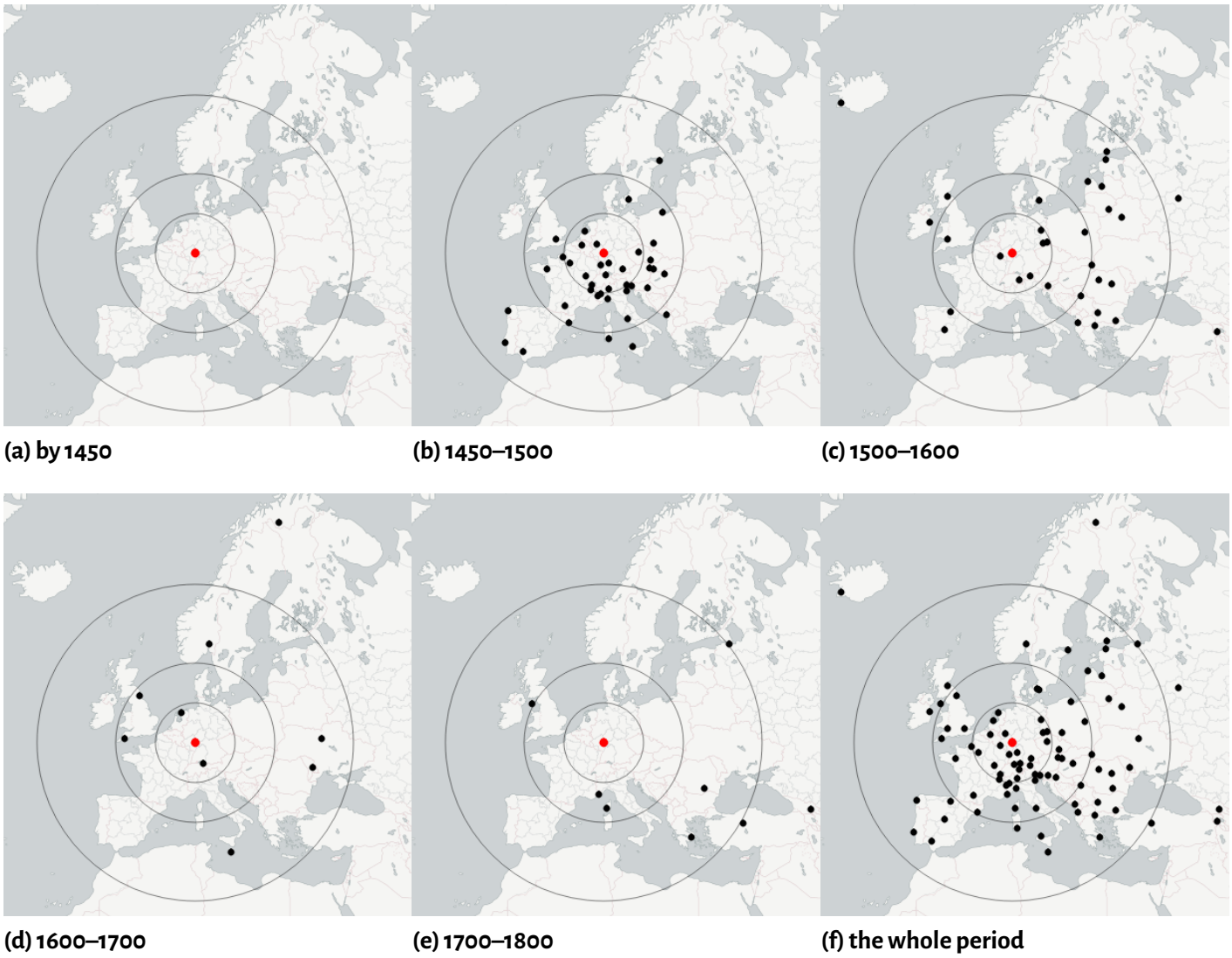
3.5.4 Distance to Mainz as an instrumental variable

The regression analysis thus far documents the positive association between the printing press and language standardization, which is robust to the inclusion of a host of covariates under various scenarios. However, as previously mentioned, the distribution of the printing press is not random. Unobserved (pre-press) characteristics may drive the technology's spread or jointly determine press adoption and language standardization. To account for this endogeneity concern, I follow Dittmar (2011) and by Rubin (2014) to exploit the exogenous variation of the geographical distance to Mainz, Germany, as an instrumental variable for the homelands of European ethnic groups.

Distance to Mainz is an ideal instrument for my argument. Such an instrument should be correlated with the printing press but not with vernacular dictionaries. At the same time, it should affect the outcome variable only through the proposed causal mechanism. The distance-to-Mainz variable satisfies these criteria. To begin, it has been shown that the Gutenberg press diffused through Europe in roughly a concentric-circle fashion. This describes not only the spread of the press but also the patterns of human interactions in early-modern Europe more generally. Dittmar (2011) points out that Gutenberg and his collaborators jealously shielded the proprietary knowledge of the technology from erosion. Only nearly a century after invention was the earliest known manual on the metal type was published (Dittmar 2011, 1140). Geographical proximity thus offers greater chances of accessing this information. More broadly, premodern times were characterized by the "small world," in which the distribution of technology or information occurred in the concentric-

circle manner. Using a mathematical model, Marvel et al. (2013) show that the bubonic plague, which triggered the Black Death and decimated approximately the two-thirds of the European population in the mid-fourteenth century, spread in this manner. Hence, the closer an ethnic group's homeland is to Mainz, the more likely it is to adopt the printing press. Figure 3.6, albeit not perfect, broadly supports this statement.

Figure 3.6: Geographical distribution of printing press 1450–1800.



(a) by 1450 **(b) 1450–1500** **(c) 1500–1600**

(d) 1600–1700 **(e) 1700–1800** **(f) the whole period**

Note: The red dot denotes Mainz (the center in each map) and black dots denote the homeland city that adopted the press in the given period. The circles indicate, from small to big, the geodesic distance from Mainz at 500km, 1,000km, and 2,000km, respectively. *Source:* See Figure 3.1.

The second rationale for using this instrument is the absence of a theoretical connection be-

tween Mainz and language standardization. No known ethnic groups have identified Mainz as their homeland city or chosen it by its geographical proximity to Mainz; nor did the city play a role in the diffusion of the vernacular use. Similarly, Rubin (2014) demonstrates that Mainz was not a political, economic, or religious center before the printing press. That Mainz was an ordinary town also implies that it is unlikely to be connected to war or economic growth. The invention may have fueled the pace of growth but the location of Mainz is unlikely to predict it. Recent empirical works in economic history also exploit the city's exogeneity in their estimation methods for outcomes such as the Protestant Reformation and economic growth in early-modern Europe (Dittmar 2011; Rubin 2014).

To estimate the effect of the distance-to-Mainz variable as the instrument, I estimate the IV probit model with the following system of equations:

$$\text{Press}_i = \alpha_2 + \beta_2 \text{Distance to Mainz}_i + \gamma_2 X_i + \delta_i + \epsilon_i,$$

$$\Pr\left(\text{ethnic group } i \text{ having dictionary in Year } t = 1 | X_i, \widehat{\text{Press}_i}\right)$$

$$= \Phi\left(\alpha_2 + \beta_3 \widehat{\text{Press}_i} + \gamma_3 X_i + \delta_i + \epsilon_i\right).$$

I use the OLS to estimate the first stage where the outcome variable is the adoption of the printing press. In the second stage, I use probit regression, where Φ denotes the normal cdf, by including the predicted values of the first stage model for the press. The model includes all controls from the fully-specified model in Table 3.3, but for the first stage I recalculated the urban potential and war frequency variables to confine their effects to the fifteenth century. The reason is to see if Mainz and its distance to the ethnic homelands were correlated with these variables *prior to* the Gutenberg invention. In addition, I omit the distance to Wittenberg or Zürich and replaced the Protestant with Catholic fixed effects for the first stage to follow the chronological sequence of events.[28] This calibration also allows me to assess whether Mainz was an economic center or underwent conflict

[28]See Febvre and Martin (1976), Bernstein (2013), Eisenstein (1979) for theoretical propositions and Rubin (2014) for empirical examination.

before or around the time when the printing press was invented. In the second stage, I use the predicted values of the first-stage regression to estimate language standardization with the same set of covariates, including the distance to Wittenberg or Zürich and the Protestant dummy reintroduced. To see the impact of the printing press on early standardization, I subset the data set in accordance with vernacular codification (1) by 1850, (2) by 1900, (3) by 1950, and (4) by 2000.

Table 3.6: IV probit regression outputs of the printing press' impact on language standardization.

	First stage	Second stage			
		language standardization			
dependent variable	printing press	by 1850	by 1900	by 1950	by 2000
	(1)	(2)	(3)	(4)	(5)
Log distance to Mainz	-0.159^{***}				
	(0.060)				
Printing press		11.822	4.639	4.741^{**}	4.018^{**}
		(37.038)	(3.051)	(2.284)	(1.828)
All controls	Yes	Yes	Yes	Yes	Yes
N cumulative dictionary		7	27	48	104
Observations	2,052	1,539	1,710	1,881	2,052
p-value, Wald exogeneity test		0.75	0.12	0.038^{*}	0.028^{*}
F-statistic on weak instrument	24.99				

Note: Robust standard errors clustered by ethnic groups. IV probit estimation: first stage is OLS, second stage is probit, regressed on predicted values from the first stage. In the first stage, *Log urban potential* and *war frequency* are for the fourteenth centuries. *Western Europe* is used as the reference category for region fixed effects and thus omitted. In the second stage, *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 3.6 documents the results from IV probit regression. It indicates that the impact of the printing press on language standardization is positive and largely significant, when instrumented for the distance to Mainz. In the first stage, the distance-to-Mainz variable is negative and significant.

The sign of the coefficient means that the closer ethnic groups' homeland is located to Mainz, the more likely they are to acquire the print technology. This result is consistent with the empirical literature that uses the same IV. The *F*-statistic on the weak instrument is approximately 25, which is above the conventional threshold of 10. In the second stage, the magnitude of the printing press is consistently in the expected direction. For the first periods of 1850 and 1900, the printing press is positive but not significant. This is largely because of the lack of variation in the outcome variable: by 1850 there were only seven vernacular dictionaries and by 1900, twenty-seven. But standard errors shrink as data in the outcome variable for each period gets richer. 75 percent of vernacular codification took place during the twentieth century. The coefficients of the print variable after 1900 become significant in the second-stage regression, reflecting this underlying data. Notably, the magnitude remains quite large: access to printing technology increases the probability of producing vernacular dictionaries by 400–470 percentage points in the twentieth century. Despite the lack of data for early periods, the instrumental-variables approach provides additional support for my overall hypothesis that the printing press predicts language standardization for European ethnic groups.

3.6 Conclusion

This article has systematically investigated the association between the printing press and the standardization of the vernacular for European ethnic groups from a long-term historical perspective. I argued that the Gutenberg press substantially reduced the cost of access to information, thus enabling vernaculars to be more popularly used and eventually win over Latin as the primary vehicle of written communication for political, economic, and social transactions. Moreover, I hypothesized that since language codification takes a long time in the order of centuries, the early acquisition of the press should give ethnic groups a head-start to develop their tongue. Using a new data set I constructed, this article has demonstrated that the time between the press and vernacular dictionaries takes 360 years on average. Statistical analysis confirms my hypotheses that print technology

is positively and significantly associated with language standardization. It also supports my broader arguments that variation in cultural consolidation for ethnic groups is not solely attributed to territorial sovereignty and that historical events have a persistent impact on contemporary outcomes.

To what extent does my argument carry outside Europe? Although such an analysis is beyond the scope of this study, it seems that there are a few Europe-specific attributes relevant to language standardization. African experience, for instance, provides a useful comparative perspective. One similarity is that consistent with the European experience, recent empirical research documents that (early) access to technologies that enable human-capital development is positively associated with vernacular codification and other outcomes such as democracy in the long-run (Cagé and Rueda 2016; Woodberry 2012). However, unlike in Europe, access to printing technology in sub-Saharan Africa beginning in the nineteenth century was largely limited by proximity to the Protestant missionaries; in addition, there was no indigenous capacity to produce the movable-type press (Cagé and Rueda 2016, 73, 74). A reliance on imports suggests that the cost of access to information would remain high, leaving smaller room for literacy in general and the development of vernacular culture more specifically. European colonialism likely reinforced this trend. By comparison, early-modern Europe had an environment in which printers moved across Europe to spread the technology, and there were multiple routes to gain access to it. This paper has sought to demonstrate that such conditions were critical to standardization for many minority tongues in Europe that have survived despite institutional centralization in modern times.

Chapter 4

DOES CULTURAL CONSOLIDATION LEAD TO PUBLIC GOODS PROVISION? EVIDENCE FROM THE FAILURE OF THE OCCITAN NATION IN FRANCE

4.1 Introduction

Does cultural consolidation lead to public goods provision? The convention is that ethnic heterogeneity stirs competition for scarce resources especially in developing societies, in which winning groups disproportionately provide public goods to the members of their own groups over those of others (Alesina, Baqir, and Easterly 1999; Habyarimana et al. 2007, 2009; Miguel 2004; Miguel and Gugerty 2005). The hypothesized negative relationship between ethnic diversity and public goods provision has informed the large empirical scholarship in political science and economics on a range of outcomes, including economic growth (Alesina et al. 2003; Alesina and La Ferrara 2005; Desmet, Ortuño-Ortín, and Wacziarg 2012; Easterly and Levine 1997),[1] redistribution (Desmet, Ortuño-Ortín, and Weber 2009), and the quality of government (La Porta et al. 1999).[2] Put more broadly, the degree of cultural homogeneity in society can affect the development of institutions, which, in turn, has an impact on society's political and economic performance in the long-run.

One implication of this stylized fact is that within-group homogeneity, or cultural consolidation within group, should lead to the strong provision of public goods. The logic is that high degrees of consolidation, as reflected in a well-codified language, make communications among the members of the group more efficient and reduce administrative costs of taxation, the enforcement of rules,

[1] For recent works on cultural practices as a cause of economic growth, see, for instance, Mokyr (2002), Mokyr (2016), and McCloskey (2016).

[2] For a recent review and the direction for future research on this relationship, see Singh and vom Hau (2016) (Volume 49, Issue 10) and the subsequent September 2016 issue of *Comparative Political Studies* (Volume 49, Issue 11).

and other transaction costs. Public goods provision within group is expected, because this investment strengthens the group's ability to compete with others in a country for limited state resources. In developing countries, ethnicity often provides an effective building bloc to form a minimum-winning coalition (Bates 1983; Fearon 1999). In this competitive environment, investing in cultural consolidation is a strategic move and has profound consequences to a group's future capacity to use its cultural attributes in politics. This decision also matters because it can eventually decrease ethnic diversity in society by attracting less-developed groups to adopt the more developed group's cultural practices. This scenario is most likely to happen when a country has a majority ethnic group with consolidated cultural attributes, and the majority group stamps out the cultural practices of minority groups. This process depicts broadly how nation-states come about.

However, the hypothesis that within-group homogeneity yields public goods provision has yet to be tested systematically. The literature on ethnicity and nations typically considers "positive" cases, in which ethnic groups consolidate their culture, build strong institutions, and invest in public goods. It tends to leave out "negative" cases, where none of these outcomes happen or overlook the possibility that ethnic groups achieve cultural consolidation but fail to provide public goods for their own members.[3] However, a quick survey on empirical evidence demonstrates that these outcomes do occur. For instance, there are an estimated 7,000 ethnolinguistic groups in the world, but only about 200 successfully acquired strong institutions. More specifically, there is not a hypothesis that explains why the Timorese, a minority ethnic group in Indonesia, invested in public goods and eventually achieved independence, while the Acehnese, another minority in Indonesia, did not. Investigating failed cases like this offers useful insights for more sophisticated theories of ethnicity and nations.

In this paper, I use quantitative and qualitative methods to consider the failed case of the Occitans to generate new hypotheses about the relationship between cultural consolidation and public goods provision. An in-depth investigation of one ethnic group is crucial, because there are few

[3]The fourth possibility is that ethnic groups do not consolidate their culture but invest in public goods. It is highly unlikely unless public goods are given by a third actor such as a colonial master.

studies that account for the failure of nations. It provides building blocks for failure as a theoretically crucial outcome. I first offer a brief literature review on theoretical and empirical research on failure and discuss the importance of addressing failure. I select two cases with top scores—the French and Occitans based on propensity-scores matching methods. Since there are virtually no investigations about the causes of failure, I turn to process-tracing methods to examine the success of the French nation, which is recounted briefly and serves as the reference category, and the failure of an Occitan nation, whose causes are investigated. My main finding is that although the Occitans standardized their language early and established an active, growing literary society with successful publications, the society's senior leadership had little motive to pursue the politically more ambitious goal of federalism or self-determination. They remained unable to resist the the effort of the French counterpart on language and cultural unification primarily through free, compulsory primary education across the country. Occitans confined their organization's activity largely to language preservation. This inaction proved fateful, as the movement of language revival never regained another momentum, and the political status of the distinct vernacular is being under threat today.

This paper makes two contributions. First, it offers a new hypothesis about the failure of nations. I trace how the absence of leadership in the literary society of the Occitans missed the opportunity to mobilize vernacular speakers politically and led to the decline of the organization's activity. This outcome is counterintuitive to the convention in the literature on ethnicity and nations which assumes that once vernacular codification is achieved, ethnic groups would naturally invest in their public goods like schooling to further consolidate their cultural practices and maximize the chances of survival. The Occitans partially fit this stylized fact—codification but no subsequent institution-building. My process-tracing analysis uncovers a hitherto overlooked hypothesis about a group's leadership to document this "near-miss" case, in which the Occitans could have taken advantage of their early language development but did not make further progress. I propose to examine more closely the attributes of ethnic groups' elite organizations to better understand variation in the success and failure of the nation.

My second contribution is that within-group cultural consolidation does not necessarily lead to public goods provision. The political economy scholarship on the relationship between ethnic diversity and public goods provision implies that ethnic solidarity is positively associated with high degrees of provision. My analysis, however, shows that this is not always the case. Occitan elites had the capacity and resources to provide public goods for schooling and expanding their literary society for political mobilization, but they opted for a more modest goal and not to compete openly with the French. My finding joins emerging research on this relationship that highlights the importance of paying closer attention to the politics and history of a given context to understand variation in the provision of public goods more fully (Singh and vom Hau 2016).

This paper proceeds as follows. The next section provides a theoretical discussion about the failure of nations. In the third section, I discuss my empirical strategy about case selection using matching methods. The fourth section offers a comparative case study to explore the causes of failure, in which I compare a brief, canonical case of the French to an investigation of the failed case of the Occitans. The fifth section concludes with a discussion about future research.

4.2 Theorizing the failure of nations

4.2.1 Why study failure?

In the literature on ethnicity and nationalism, there are few systematic investigations on the failed cases of ethnicity or nations. Extant work typically begins with the assumption that groups under study have, or would eventually have, a sovereign territorial state and seeks to uncover patterns or important factors that account for the consolidation of groups' cultural attributes (Breuilly 1993; Hobsbawm 1990; Roeder 2007; Tilly 1994; Wimmer 2002). The focus on these "ideal types" yields a hypothesis that a centralized state is virtually a prerequisite; ethnic groups aspiring to survive or institutionalize their cultural practices would need to acquire a state in the first place. This precondition also implies that "positive" cases that *succeed* in consolidating cultural practices would be disproportionately studied. Other studies in the literature that do not explicitly adopt the statist

assumption follow the ideal types. For example, early scholars of nationalism choose cases, which are typically European ones, to examine variation in citizenship rules, education policy, or festivals in national holidays that constitute the public display of "national consciousness" (Deutsch 1966; Kohn 2005).[4] Although these works do not specifically discuss the capacity to make and enforce rules by the central government, the statist assumption and the ideal-type method lie in their analysis.[5]

However, the focus on these "positive" types necessarily excludes "negative" ones which are far greater in number. Ernest Gellner recognizes this point in his notes, where he points out that of an estimated 8,000 ethnolinguistic groups in the world, only a fraction acquired an independent state while many nationalisms "failed to bark" (Gellner 2006, 42–3).[6] Gellner does not elaborate on them further in his work. In addition, on a practical matter, it is difficult to classify studies on "national consciousness." Should researchers count them as "positive" when expressions about group identity are made in public? Do such public displays require state sponsorship? I argue that one difficulty in this empirical strategy stems from the lack of attention to negative cases, because their absence provides no point of comparison in understanding why only few groups succeed at consolidating. Research on positive cases strengthens the scholarship by enriching the variation of examples within a single class of cases but is not informative of understanding *other* classes of cases whose populations are greater.

Two reasons merit discussion for the imperative of studying failed cases. The first is evidence. It adds greater depth in the empirics to the universe of ethnic groups that succeed at standardizing culture or fail to do so. In the literature on ethnicity and nations, more evidence is particularly im-

---

[4]Anderson (2006) is a prominent example with illustrations on the chapter on "Census, Map, Museum." Subsequent scholars produce detailed studies based on this theme. See, for instance, Kertzer and Arel (2002) on census, Thongchai (1994) on maps, and Rydell and Gwinn (1994) on museums. See also Hobsbawm and Ranger (1983) on government policy can "invent" and impart a particular mass identity.

[5]Brubaker (1992) runs the middle ground in that he offers a compelling account of variation in citizenship policy between German and France by focusing on the impact of the history of state-building in the respective countries.

[6]Language is typically used to define ethnicity in the literature.

portant as there are far more ethnic groups without an independent state or consolidated culture. These data show that most groups do not invest in cultivating a unique "brand" or identifier that distinguishes their own community from others. Since one group's cultural attributes can sometimes overlap those of another (e.g., language or religious denomination), self-branding is critical for survival, especially for those with small population size. The fact that many groups choose not to make such an investment is puzzling from evolutionary perspectives. Thus, richer evidence can not only expand the pool of observations but also hone existing hypotheses about which causal factors are stronger in explaining empirical variation in success and failure. Second, new evidence would lead to greater theoretical sophistication. In particular, data on failures help specify the conditions under which an ethnic group successfully consolidates its cultural practices.[7] Only when there is strong evidence on negative cases can the analyst argue compellingly that independent statehood increases the chances of ethnic group survival or that there are other, non-state routes are viable. Studying failure thus offers greater empirical variation in the success and failure of nations and thrusts the scholarship for more systematic analysis.

4.2.2 Defining failure

Defining the "failure of the nation" entails two components—"failure" and "nation." The literature on nations and nationalism typically holds that one can observe the development of a "nation" when the sovereign state is present (Gellner 2006). The argument suggests that the category of nations without states is excluded from analysis.[8] The precondition of the state makes the definition of the nation inseparable from it, which, in turn, creates difficulty understanding the failure of na-

[7] To give an example in political science, the scholarship of democratization followed this path. It started with the pathbreaking work by Pzreworski and Limongi (1997) and Przeworski et al. (2000), whose evidence challenged the modernization hypothesis of Lipset (1959). The subsequent literature provided new evidence and led to the honing of democratization theories in such works as Boix and Stokes (2003), Boix (2003), Acemoglu and Robinson (2006), etc.

[8] Another possible outcome is one of "a state without a (single) nation." It typically refers to a state with no discrete, cohesive cultural group (like Bosnia and Kosovo) or to a state with multiple competing groups (like Switzerland and Sri Lanka). Either case, however, presupposes the presence of the central state.

tions, because in this conceptualization the failure of nations can be equated as the absence of the sovereign state.[9] Therefore, it is important to define the nation first.

In this paper, I define the *nation* as a collectivity with standardized cultural attributes. By contrast, I define the *ethnic group* as a collectivity without standardized cultural attributes. My definition is useful for understanding the "success" and "failure" of nations. First, it breaks the nation into a set of discrete, clearly identifiable observations. These include language, flag, anthem, school curricula (about traditions), public goods, and other cultural attributes that comprise the group. In my definition, access to the state is just one dimension, although it is a critical one. Second, my definition removes "state" as a precondition from "nation." This conceptual distinction is important for two reasons. One is that it is now possible to theorize about the success and failure of nations without first assuming the state. In addition, my definition allows for investigating a number of ethnic or national groups that do not have an independent state, which greatly expands a pool of observations for empirical analysis. By focusing on concrete and observable dimensions of cultural attributes, nations can be distinguished from ethnic groups that lack the standardization of cultural attributes.

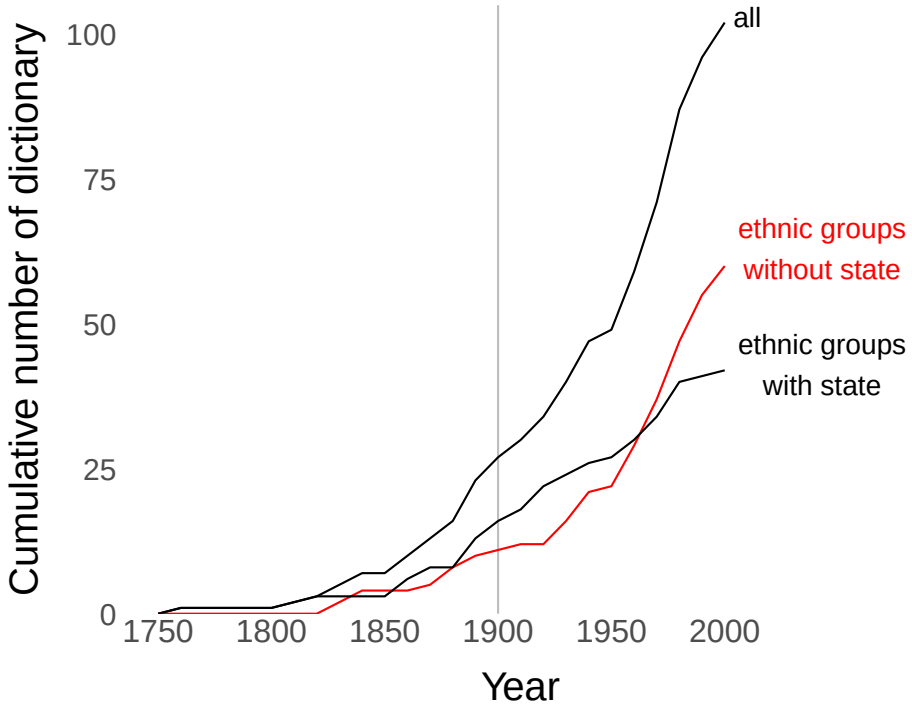
Table 4.1: "Success" and "failure" of nations.

ethnic group	language	public goods	flag	school curriculum	.	.
A	Yes	Yes	Yes	No	.	.
B	Yes	No	No	Yes	.	.
C	No	Yes	Yes	Yes	.	.
D	No	No	No	No	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

[9] I suggest that the theoretical indivisibility between state and nation has partly created the confusion in the social sciences about defining and capturing the nation as a standalone concept and as a concrete observation. See, for instance, Chapter 4 of Connor (1994), "Terminological Chaos ("A Nation Is a Nation, Is a State, Is an Ethnic Group, Is a ...)," for a discussion.

My definition of the nation enables a systematic understanding about the successes and failures of nations as outcomes of interest. Table 4.1 visualizes these outcomes in a hypothetical data set. It highlights two issues. First, this approach permits the theorizing about the relative importance of one cultural dimension over another. I argue that among the cultural attributes, language is the most important dimension to standardize. Once the vernacular is codified, an ethnic group has an easier time communicating within group, which makes the standardization of other cultural dimensions more effective. Thus, ethnic groups are expected to invest first in standardizing their vernacular. However, language standardization is difficult because it requires literacy in contrast to other cultural attributes like flag and anthem, which do not. Moreover, codification requires high fixed costs. Languages evolve through importing and exporting words from one another, and codification typically entails the "purification" of a language by studying the etymology of words and removing foreign ones to demonstrate that the language serves as an untainted symbol of identity as a group (Burke 2004).

Figure 4.1: Cumulative number of vernacular dictionary publication among European ethnic groups.



Source: Burke (2004), Dalby (1998), Price (1998), among others.

Figure 4.1 documents the cumulative output of language standardization for 171 European ethnic groups 1750–2000. I proxy language standardization by taking the first publication year of vernacular dictionaries for each ethnic group ($n = 104$). This operationalization demonstrates the difficulty and high fixed costs of vernacular codification. As for the difficulty, Figure 4.1 indicates that not all ethnic groups achieve vernacular codification, even though it is an essential cultural dimension to standardize. Languages standardization is never a "natural" outcome and requires an investment. With respect to fixed costs, many groups, such as the Italians, Swedes, French, and Russians, founded a specialized, government-sponsored academy to cover the expenses. However, Figure 4.1 shows that the state is not a prerequisite for language standardization. When early standardization is defined as the publication of vernacular dictionaries by 1900, where it occurred among only 27

out of 171 ethnic groups (or 16 percent) in my data set, there is little difference in the outcome by this time between the ethnic groups with an independent state and those without one (highlighted in red). This discussion about the historical sequence between language and state is but one example. As Table 4.1 suggests, an expansive list of dimensions and greater data enable more systematic theorizing about the causes of success and failure.

The second issue is empirical. My definition of the nation permits the flexible but consistent operationalization of the cultural attributes of interest in empirical analysis. For example, I use the first publication date of vernacular dictionary in Figure 4.1, but this is not the only way to capture vernacular codification. One can create an indicator of language standardization (coded 0 or 1) by examining if ethnic groups pass a pre-defined threshold, such as standardization by 1900. Another way to operationalize language is to calculate time elapsed since standardization up to a certain date. This quantity can proxy how widespread the language is within group and is an appropriate variable if a success is captured as how long a language is sustained.

In this paper, I define the failure of nations as their failure to provide public goods. More concretely, the "public goods" that are relevant to cultural consolidation include schooling, newspapers, or even political parties that facilitate the integration of the fellow members of the community and highlight the differences in cultural attributes between one's own ethnic group and another. Extant scholarship provides ample evidence which shows that ethnic groups that own a sovereign territorial state typically resort to the power of state institutions, and often invest in infrastructure to expand state capacity, to promote one language over all others within the state and command loyalty through institutional innovations like national citizenship and conscription.[10] Ethnic groups without an independent state can adopt a similar approach, for instance by establishing language classes at a local university or by launching an ethnic political party, to mobilize co-ethnics for greater degrees of self-rule, create an opportunity to resist integration by state-owning ethnic groups, and maximize the chances of group survival.

[10]For *loci classici*, see Colley (2005), Weber (1976). For recent contributions, see, *inter alia*, Aktürk (2012), Martin (2001), Mynolas (2013), Sahlins (2004).

Currently, supply of studies on "failures" in the literature is, at best, minimal. Charles King's (2010) short essays appears to be the only one that addresses failure as an outcome of interest. It is unclear if similar works exist, because it cites no work specifically on failure. Although King provides no scrutiny about this lacuna, he suggests that scholarly path dependence plays a role, where early scholars who founded the field select cases that exist and reconstruct their trajectories of success (King 2010, 38–40).

In his essay, King (2010) offers a preliminary theoretical discussion about failure. He discusses three core attributes shared among all successful cases, i.e., nation-states: mass appeal, state-building, and international recognition. More specifically, successful ethnic groups employ effective strategies to construct and sustain the unique, discrete group identity that co-ethnics internalize as their primary group identity. These groups almost invariably have an independent state to begin with or some degree of state capacity which enables elites to mobilize the mass for independence. These mass-appeal strategies are so salient as to win international recognition. In King's essay, each of these three variables is treated dichotomously, which yields eight ($2 \times 2 \times 2 = 8$) types of outcomes. His argument is supported by a set of short illustrations about the Southern Whites in the United States, the Scots in Britain, and the Circassians in the Caucasus region of Russia. This pioneering essay is crucial for bringing attention to a rarely-discussed subject in the literature. One way to advance the understanding of the failure of nations is to take King's model as a stylized fact and expand on it in a comparative study or detailed case study. Yet before such a test, more groundwork is required to identify a plausible causal mechanism of failure. This includes a discussion of how to specify causal variables or processes and case selection procedure to consider which cases are more likely to provide a generalizable mechanism. The remainder of the paper undertakes this exercise.

4.3 Empirical strategy

4.3.1 Process-tracing methods

I use process-tracing methods to investigate the failure of nations. More specifically, my goal is to uncover a causal process in which ethnic groups with early language standardization fail to build on this momentum and invest in public goods. Process-tracing is a method to identify causal processes between an explanatory variable and the outcome variable (George and Bennett 2005, 206). The main rationale is to establish what scholars of this methodological tradition call "causal-process observations" (CPO), which refer to "pieces of data that provide information about context, process, or mechanism," and one that enables analysts to draw causal inference (Brady and Collier 2010, 318). This is an appropriate method for the purpose of this paper, as few investigations address nation failure as an outcome of interest. No hypothesis or stylized fact has yet to be offered to establish causality. The researcher can use process-tracing to both test and develop hypotheses (Bennett and Checkel 2014; Gerring 2007). It is an ideal tool to develop a conceptual framework or specify a set of key variables that should be included in the framework (Mahoney 2003, 2010).

Although scholars who use process-tracing methods have developed numerous strategies and standards to appraise the claims of causal connections, implementing these standards can be challenging (Dunning 2012, 2014). Case-studies approaches may prove to be most useful when validating the statistical correlation between an explanatory variable and the outcome variable or when seeking to identify alternative hypotheses or omitted variables. However, it seems unclear *which* cases would be appropriate to make causal inferences and *how* they can be drawn. Case selection remains a challenge in case-based causal analysis because of the need to control for background conditions before proceeding to comparison. In traditional strategies, e.g., the methods of agreement and of difference, the researcher conditions on the value of a main independent variable or the dependent variable. But the probabilities that these values occur or the conditioning of con-

trol variables attract less attention (Sekhon 2004).[11] In my case whose goal is to establish causes of failure, it is important not only to find similar cases in terms of the chances of language standardization (the main explanatory variable) and failure (the outcome variable) but also cases with similar background conditions.

4.3.2 Matching methods for case selection

I use matching methods for case selection and causal inference. In general, matching is a nonparametric preprocessing approach to adjust observational data before the parametric analysis (Ho et al. 2007; Stuart 2010). Empirical research in the social sciences often relies on observational data when an experimental method is infeasible as in historical analysis or when an experiment is impractical for cost and other reasons. Unlike data from a clean experiment, observational data almost always lack balance or are not randomly distributed across the sample. Imbalance creates difficulty for parametric analyses because it violates the assumption of ignorability or the absence of selection on observables (Ho et al. 2007, 206–7). If this violation is suspect, one needs to address the concern, among others, that unobserved factors might determine the outcome variable and/or explanatory variables. Parametric analyses without satisfying this assumption may yield spurious causal inference.

Matching methods develop to adjust for imbalance in data. To begin, the researcher takes data preprocessing steps by finding cases where the values of the covariates (exactly) match except for an indicator variable. This variable serves as a tool to assign matched observations into the "treatment" group and the "control" group. Then the analyst uses parametric models to compare the two groups to estimate causal effects of the treatment variable. Matching is thus a technique, which has extended from the experimental literature, to make observational studies approximate an experimental setting through covariate adjustment to facilitate causal inference (Gerring 2007; Sekhon

[11] For recent innovations that systematically leverage a control case to draw inference from comparative case studies, see Glynn and Ichino (Forthcoming), Glynn and Ichino (2015).

2009).

The primary goal of matching is to improve balance in the distribution of covariates (Ho et al. 2007, 215; Stuart 2010, 1). This strategy offers two major advantages. First, preprocessing data reduces dependence on one's choices of statistical modeling and specifications (Ho et al. 2007). By taking an extra step to achieve balance in the distribution of covariate data between the treatment group and the control group, the link between the treatment variable and the control variables is eliminated or diminished (Ho et al. 2007, 212). Subsequent statistical analysis, if the researcher chooses to do so, is less dependent on the assumptions of their proposed models and specifications. Second, matching methods focus on the distribution of covariates where there is not much overlap between the treatment and control groups when causal effects are estimated. Statistical models tend to perform poorly when overlap is insufficient, but researchers become more aware of data limitations in their claims about causal estimates when using smaller data through matching methods (Stuart 2010, 2).

4.3.3 Propensity score matching

In this paper, I use a propensity-score matching strategy. It is a popular strategy for case selection and causal inference and is widely used not just in the social sciences but also in medicine and health (Stuart 2010, 1). The propensity score is an index of the likelihood that a key explanatory (or "treatment") variable occurs based on observational data. A study that employs this method typically proceeds in two steps. First, the researcher estimates a model, where the matching variable is the outcome, to calculate propensity scores of the matching variable. These scores are used to assign observations into the treatment group (the matching variable that occurred) and the control group (the matching variable that did not). In the second step, the analyst estimates the effect of the propensity as the approximate average treatment effect (ATE) on the outcome of interest.

In my case, the outcome is whether an ethnic group provide public goods after early language standardization. I define *early* as language standardization by 1900, because only 16 percent (27

out of 171 ethnic groups in my data set) had published a vernacular dictionary by the turn of the twentieth century (26.5 percent when NAs are omitted) (or see Figure 4.1).[12] I use this indicator T_i as the treatment variable to assign ethnic groups in my data set:

$$T_i = \begin{cases} 1 \text{ if ethnic group standardized language by 1900} \\ 0 \text{ if ethnic group did not standardize language by 1900} \end{cases}$$

$T_i = 1$ refers to a subset of my data set where ethnic group i codified their vernacular by 1900 and $T_i = 0$ indicates the rest of the ethnic groups that did not do so, given the pre-treatment characteristics of the homeland city for each ethnic group. More formally, this case selection strategy aims to capture the effect of T_i on the outcome, Y_i, which is whether ethnic groups invest in public goods provision:

$$Y_i \equiv Y_i(T_i) = (T_i)Y_i(1) + (1 - T_i)Y_i(0)$$

To assign ethnic groups in the treatment and control groups, I first compute propensity scores that denote the probability of achieving vernacular codification by 1900, conditional on a vector of ethnic group attributes X_i:

$$P(X_i) = \Pr(T_i = 1|X = X_i) = \mathbb{E}[T_i|X = X_i]$$

This data preprocessing allows me to capture the effect of early language standardization on whether to start institution-building by conditioning on potential confounding variables for standardization.

It is important to note that matching based on propensity scores is not without problems. One is that causal estimates from matching models can suffer from model dependence. It is the problem in which the researcher makes inferences from fitted models but extrapolations are not supported by the existing data or these extrapolations are outside the bounds of underlying data (King and Zeng

[12] I discuss the sources of these 171 ethnic groups in Appendix C.

2007). Propensity score matching does not, in theory, reduce model dependence (Iacus, King, and Porro 2012). But the rationale of using this matching strategy in my paper is to select cases for hypothesis-generation rather than to estimate causal effects of the models. More specifically, my goal is not to extrapolate the impact of early language standardization on state-building but to find cases that bear empirical similarities and differences as a preparatory step for in-depth investigation.

I estimate propensity scores using a logistic regression model, where the outcome variable is a binary indicator of whether or not 171 European ethnic groups standardized their language by 1900. Following Sasaki (Forthcoming), I proxy language standardization by taking the first publication date of vernacular dictionaries. I include all covariates used in the estimation in his work but I made two major changes. One is that I truncate the data to pre-1900 to capture the pre-treatment effects of the variation in language standardization. Second, I transformed the original time-series and cross-sectional data set to enable cross-sectional estimation, because the original data set contained information from 1400 through 2000.[13] More specifically, I recode the following time-varying covariates: First, I obtain the total time (in years) that ethnic groups have the printing press by calculating the time elapsed since print adoption. I also create the total time for the university and bishoprics for each group using the same method. Second, I calculate the mean for the war frequency and urban potential variables between 1400 and 1900. War frequency represents the average occurrence of war during this period. Urban potential is a measure of economic growth used in economic history by capturing the extent to which population growth of a given homeland city is affected by nearby competing cities. Higher values indicate that cities are surrounded by other competing (i.e., growing) cities and lower values mean that growth occurs in a sparsely-populated area. Third, I recode the fixed effects variables for the Ottoman, Russian, and Habsburg Empires by calculating time under each of these polities. Some stateless ethnic groups, especially those in Eastern and Central Europe, underwent multiple state "owners" over the course of the five centuries.

[13]The sources of these variables and the summary statistics are in Appendix C.

Fixed effects do not represent this change, so I measure total years of each of these imperial rulers. The rest of the variables is time-invariant and needs no transformation.

Table 4.2: Regression outputs of language standardization for ethnic groups by 1900.

model	logit	OLS
	(1)	(2)
Time since print adoption	0.006*	0.0005*
	(0.003)	(0.0003)
Time since university founded	0.003*	0.001**
	(0.002)	(0.0002)
Time since bishopric erected	−0.001	−0.0001
	(0.001)	(0.0001)
Mean urban potential	0.074	0.001
	(0.349)	(0.041)
Log distance to Wittenberg/Zürich	−0.0002	−0.00003
	(0.001)	(0.0001)
Mean war frequency	0.071	0.009
	(0.080)	(0.010)
Elevation	−0.0002	0.00002
	(0.002)	(0.0001)
Terrain ruggedness	−0.002	−0.0001
	(0.005)	(0.0002)
Hub Roman road	0.344	0.039
	(0.861)	(0.107)
Roman road	0.345	0.051
	(0.736)	(0.081)
Oceanic port	−0.395	−0.048
	(0.909)	(0.081)
Island	0.941	0.100
	(1.120)	(0.111)

continued on next page

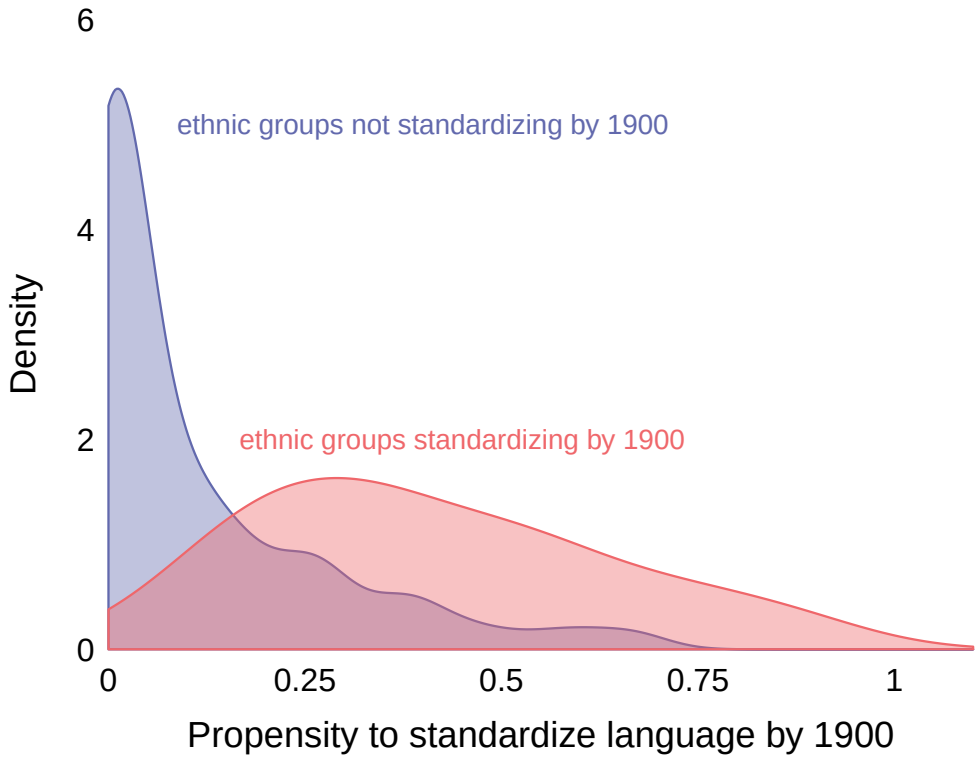
continued from previous page

model	logit (1)	OLS (2)
Time under Ottoman Empire	−0.003	−0.0004
	(0.004)	(0.0004)
Time under Russian Empire	−0.001	−0.0002
	(0.004)	(0.0004)
Time under Habsburg Empire	−0.003	−0.0005*
	(0.003)	(0.0003)
Religion fixed effects	Yes	Yes
Geographic region fixed effects	Yes	Yes
Observations	171	171
F-statistic		2.49***
Wald Chi square	3.58*	
R^2	0.342	0.162

Note: Western Europe is used as the reference category for region fixed effects and thus omitted. *Catholic* is used as the reference category for religion fixed effects and thus omitted. The Wald Chi Square is the effect of time since print adoption. R^2 for logistic regression is pseudo-R^2. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 4.2 documents logistic and OLS (ordinary least square) regressions on language standardization by 1900 (complete results available in Appendix C). Consistent with the results in the work by Sasaki (Forthcoming), the variable on the printing press is significantly positively associated with language standardization. More concretely, elapsed time since the adoption of print technology increases the chances of vernacular codification by 0.6 percentage points. Similarly, accumulated time time the founding of a university has a significant positive impact, suggesting that higher education generates demand for learning and teaching in a local tongue. Similar results obtain when OLS is used. It is important to note that a main reason that magnitudes are relatively small for many covariates in both models is that only 27 of 171 ethnic groups achieved standardization by 1900. The vast majority of cases occurred in the twentieth century.

Figure 4.2: Distribution of propensity scores for two sets of ethnic groups on language standardization by 1900.



I use the parameter estimates of the logistic model to calculate propensity scores. Figure 4.2 displays the densities of the scores for two groups: those that standardized language by 1900 (denoted in red) and those that did not (denoted in blue). It shows that ethnic groups with standardization have largely higher scores, while those without it are typically concentrated in low propensity scores, especially in the scores between 0 and 0.25. Still, the distribution of both subsets is similar in that the scores are largely less than 0.5.

Table 4.3: Ethnic groups with propensity score above 0.5.

ethnic group	score	dictionary by 1900	sovereign state	country
French	0.888	Yes	Yes	France
Occitans	**0.838**	**Yes**	**No**	
Swiss-Germans	0.726	Yes	Yes	Switzerland
Russians	0.722	Yes	Yes	Russia
Rhinelanders	0.680	No	No	
Flemish	**0.667**	**Yes**	**No**	
Romands	0.663	No	No	
Burgundians	0.620	No	No	
Belarussians	**0.587**	**No**	**Yes**	**Belarus**
Alsatians	**0.574**	**Yes**	**No**	
Armenians	**0.565**	**No**	**Yes**	**Armenia**
Bulgarians	0.541	Yes	Yes	Bulgaria
English	0.539	Yes	Yes	Britain
Germans	0.536	Yes	Yes	Germany
Swabians	0.516	No	No	
Dutch	0.508	Yes	Yes	Netherlands

Note: Theoretically counterintuitive outcomes are highlighted in bold.

Table 4.3 further breaks down the distribution for sixteen groups with the propensity score of 0.5 or above. It adds two columns of information: dictionary publication by 1900 and territorial sovereign today. Two points merit discussion. First, high propensity scores reflect the high likelihood of early vernacular codification. Of the sixteen groups, ten achieved codification by 1900. Second, nine groups acquire an independent state. Observations are more equally distributed between those with a state and stateless units.

This variation provides purchase for case selection. It raises an empirical puzzle as to why early language standardization does not always lead to public goods provision. This is at odds with the hypothesis discussed earlier, in which ethnic groups, once standardization is achieved, would have an incentive to build institutions to increase the chances of survival. The decision *not* to build on this momentum appears counterintuitive. In the European context, it is a highly risky choice to

the extent that it is near self-destruction because of recurring warfare and territorial conquest in the early modern period through modern times (McNeill 1982; Tilly 1975, 1992; Wimmer 2013). The inability to raise revenue and manpower for defense would leave ethnic groups vulnerable to political incorporation by others with access to the state. However, descriptive evidence from Table 4.3 shows that this empirical puzzle offers a case-selection strategy based on the counterfactual framework.

Table 4.4: Two-by-two table on the relationship between language standardization and state-capacity building.

		Public goods provision	
		Yes	No
Early language standardization	Yes	French, Dutch (expected outcome)	Occitans, Alsatians (counterfactual outcome)
	No	Belarusians, Armenians (counterfactual outcome)	Burgundians, Swabians (expected outcome)

Table 4.4 maps the relationship between language standardization and public goods provision, where the ethnic groups in the table are examples drawn from the previous table. It is unsurprising that ethnic groups with early language standardization would invest in state institutions as in the case of the French and others in the top left cell. Similarly, ethnic groups like the Burgundians in France that did not codify the vernacular have a weak incentive to provide public goods, because doing so is not cost-effective (the bottom right cell).

But Table 4.4 also presents counterfactual cases—realizations in the opposite direction of the theoretical expectations. First, there are relatively late language standardizers which eventually acquired an independent state (the bottom left cell). Groups like the Belarusians and Armenians were long under the yoke of stronger powers, like the Russian and Ottoman Empires, and only in the

twentieth century did they achieve independence. Imperial rule suggests indirect governance, in which subjects tend to enjoy day-to-day autonomy in exchange for consent to pay taxes. Despite such autonomy, ethnic subjects are unlikely to develop their vernacular either because the imperial language is the means of political administration and economic transactions or because the fixed cost of codification is perceived to be too large, even though the vernacular is the preferred medium among the local populations. These groups would quickly invest in vernacular codification following political independence. They are willing to pay greater costs on language through schooling and centralized institutions such as the state-sponsored academy, as these policies help consolidate governance in the long-run.

Second, another set of counterfactual cases refers to those early language standardizers without public goods provision (the top right cell). This category represents groups like the Occitans that miss the critical opportunity to strengthen public goods provision. This is a counterintuitive outcome, because the provision can eventually lead to independent statehood and other institutions that maximize the chances of survival for ethnic groups. As Gellner (2006) hypothesizes, in modern times when stateless groups are surrounded by sovereign states, the struggle for independence begins with mobilization in the form of a political party or an activist group, typically led by intellectuals whose upward mobility is blocked due to their ethnicity. Their claims for self-determination, if successful, would have to be recognized by other independent states, because this recognition is a constituting factor of sovereignty. Even when the first attempt at independence fails, the institutions and popular support amassed through the movement can be further developed and refashioned in the next opportunity. Investment in institutions is expected, but the ethnic groups in this category deviate from it. These "near miss" observations merit investigation to discern the causes of inaction.

This paper specifically chooses the French and Occitans. First, the case-selection procedure described above suggests that the two ethnic groups are comparable on several dimensions. Both are early standardizers of the respective vernacular: the French published the first monoglot dictionary in 1858 and the Occitans in 1875 (Dalby 1998). They rank the two highest propensity scores as

reported in Table 4.3. In addition, this comparison controls for attributes like historical and institutional developments that are not observed in the regression models. Second, there are important differences between the two groups. One difference is that the French case followed the expected outcome but the Occitan case did not. Although the French took advantage of language standardization and further advanced linguistic and cultural unity throughout the country beginning in the late nineteenth century, the Occitans failed to do so. The deviation of the Occitans raises questions about the theoretical link between ethnic groups' investment in their cultural attributes and investment in public goods. Substantively, this pair controls for the historical background factors that are unobserved in the parametric analysis in Table 4.2, because the French and Occitans are in the same country. An investigation into the causal mechanisms through this comparison can uncover "omitted variables" that are previously accounted for and allows for specifying the causes of lack of public goods provision despite early language standardization. Thus, this is a hypothesis-generation case study and makes the theoretical literature on nations and states richer (Gerring 2007). Finally, the French and Occitans make a useful comparison, because one group—the French—already constitutes the "reference category" as a control in regression analysis. This brings a sharper analytical focus on the case of the Occitans as a theoretically crucial study, where the number of observations is small but the analysis has broader external validity.

4.4 Searching for the causes of failure

4.4.1 The French

This paper provides a comparative study between the French and Occitans with a disproportionate focus on the latter. The growth of public goods provision in France and its impact on the consolidation of the French culture are well-documented in the literature, where the rise of the unitary "national," i.e., countrywide, identity is regarded as a prototypical case. For the purpose of this paper, therefore, a brief recount will suffice. In France, only after the 1789 did Revolutionary elites come to grapple with the absence of cultural unity across the country. For example, Henri Grégoire,

a Catholic priest and a revolutionary leader, realized, to his dismay, that an estimated six to eight million out of 28 million populations living in France, or 21–29 percent, count not speak French in 1792 (Bell 2001, 177–8). He was also concerned that this linguistic diversity would tax Grégoire and his fellow revolutionaries to persuade the populace to subscribe to the Republican ideals (and, perhaps for Grégoire personally, to proselytize) (Bell 2001, 185, 187–94). Great ethnic and linguistic heterogeneity persisted until the mid-nineteenth century when 22 percent of the population spoke no French (Weber 1976, 67); languages spoken in France at the time included Breton, Occitan, Flemish (a Dutch variant), Alsatian (a German variant), Catalan, and Basque (Bell 2001, 16).

The French took advantage of the industrializing economy to provide two types of public goods across the country, both of which occurred after the publication of monolingual vernacular dictionaries. The first concerns roads and railways.[14] Railways started to expand in the 1840s, while Paris began to build roads, especially in rural areas, in the 1840s as "branches" linked to main railway lines (Weber 1976, 196, 203, 205). Railway density grew over time. In 1870, it was 3 kilometers (km) per every square kilometers, which was less than the railway density of Ireland. By 1910, however, the density more than doubled to 6.9km per every km^2, which ranked the third most dense coverage in industrial Europe (Caruana-Galizia and Martí-Henneberg 2013, 180). This massive construction project was significant across the country not just economically but culturally. It helped integrate the de facto auturkic economy in the non-French-speaking countryside into the market economy by paving and improving roads that peasants regarded as too shabby to be of use (Weber 1976, 199–202, 206). Culturally, a more reliable means of transportation gave peasants an incentive to join the cultural community of the French through market participation (Weber 1976, 209, 217). The enterprise of road- and railway-building thus attracted non-French-speaking peasants to *choose* to identify as French.

The second type of public goods that helped expand the French nation was public schooling. Free, mandatory primary education arose in the early 1880s due to a series of legislative reforms

[14]On the impact of railways on economic growth in Europe, see O'Brien (1977, 1983).

spearheaded by minister of public education Jules Ferry, and its success, at least in part, lay in the attendant improvements on the roads which made regular attendance easier (Weber 1976, 303). At the same time, non-French regional languages were excluded from the school system by law (Price 1998, 39). Initially, the new institution met with strong resistance in the countryside, where suspicions toward education was long prevalent. The teacher quality was low, and parents of peasant families remained unwilling to send their children to school, because the children were considered workforce and education was considered a luxury (Weber 1976, 314, 318–26). Yet once basic literacy and numeracy were shown as a path to upward mobility for the educated children and poverty alleviation for the family, public support for schooling expanded. The effect on language was clear: an increasing proportion of the population used French, the only language of instruction, over other tongues (Weber 1976, 330–31). Evidence supports this observation: in 1850, an estimated 3.3 million pupils attended primary school; the number grew to 5 million in 1880 (a more than 50-percent increase) and to 5.6 million in 1910 (12-percent increase from 1880) (Mitchell 2003*b*).

In short, the French undertook concrete policies that consolidated their culture and appealed to non-French speakers within France by providing public goods on many fronts. The expansion of railways and public schooling in France, combined with other developments like conscription, constitutes an ideal type that describes how cultural consolidation occurs and how the culture of a majority ethnic group becomes the standard within a country over that of demographic minorities. Therefore, French case provides the reference category for this paper's comparative analysis.

4.4.2 The Occitans

Overview of language standardization

In comparison to the textbook case of the French on the success of nations, few investigate the failure of the Occitans. This paper first outlines the brief history of the political and economic development in southern France which Occitans see as their homeland and then discusses the process through which little public goods provision occurred within the group.

The geographical area of Occitania, the territory that Occitans claim as their homeland, broadly refers to the southern half of today's France (the "Midi"). The hilly region that cuts across the middle of France served as the natural borders for demarcation. The south is climatically and culturally different from the north. Prior to French incorporation in the early thirteenth century, those who lived in Occitania had closer cultural ties with Spanish and Italian speakers than with French speakers (Paterson 1993). The most important region is Languedoc, a southernmost area that faces the Mediterranean, whose political and economic centers comprise the cities of Toulouse, Carcassonne, and Narbonne (see Figure 4.3). This region began to be a target of political incorporation by the Capetian monarchs in 1209,[15] and the French effort to assert authority lasted through much of the thirteenth century (Given 1990, 39). The Capetians installed what may be described as "indirect rule," in which local institutions were retained in preference to the transplantation of the imperial institutions and local aristocrats managed to preserve and exercise limited authority over legal and economic affairs (Given 1990, 57–63). This is a typical governance strategy in Europe from the Middle Ages until around 1750, because the political centers like Paris were too weak to project authority over their nominal domain (Elliott 1992; Koenigsberger 1987). It is an inexpensive yet practical compromise that ruling monarchs were willing to make so long as the autonomous local elites pledged loyalty in terms of taxation.[16] Although struggle for control over this territory with the English monarchs continued intermittently through the mid-fifteenth century, Occitania remained primarily under French rule (Beik 1985, 37).

Beginning in the early fifteenth century, a new institution of indirect rule was established. Languedoc, along with other provinces in the countryside, was accorded with the status of a *pays d'états*, one of the two types of administrative jurisdiction (provincial estates) under the Ancien Régime.[17]

---

[15]France's participation was in response to the Pope's call to stamp out the Cathar heretics in southern France and was part of the Albigensian Crusade (1209–29).

[16]See Dincecco (2015) for a discussion of the problems of reliably and consistently raising revenue due to the fragmentation of authority.

[17]The other type is the *pays d'élection*, where royal bureaucrats came to levy taxes directly. There is a third category, *pays d'imposition*, which was applied to later-acquired territories.

It is a provincial-level financial district for the purposes of taxation. Residents of this status possessed the right to hold representative assemblies, where elites could negotiate tax burdens with the monarch and collect taxes on their own (Kwass 2000, 70). This institution thus allowed the residents to retain a degree of self-government in terms of rights and privileges. Languedoc, whose provincial capital was located in Toulouse, convened the assemblies roughly once a year from 1418 until 1789 when all the Ancien Régime institutions were abolished.[18] In addition to the matters of taxation, attendees introduced legislation and debated budgetary allocation on public goods (Swann 2012). The assemblies of Languedoc enjoyed greater discretion over the regional affairs than those of other regions, such as Brittany, Burgundy, and Artois, and as a result, the residents were able to preserve strong regional identities (Beik 1985, 38).[19]

[18]The data on the frequency of regional bodies came from Kiser and Linton (2002) and Blockmans (1976).

[19]This interpretation of the early-modern era, especially during the "absolutist" period in which the monarch depended much more on the provincial elites for stability than previous research showed, is based on the emerging historiography on absolutism as "social collaboration." See Beik (2005) and Campbell (2012) for reviews.

Figure 4.3: Region of Occitan speakers at the time of the French Revolution.

Note: The Occitan-speaking region is denoted by gray and Languedoc, a province within the Occitan region, is denoted by blue. The white dot within Languedoc represents Toulouse.

A degree of self-rule implies that Occitans would enjoy the freedom to use and develop the vernacular as the primary means of communication. It seems that the vast majority of the Occitans knew little French and, especially in Languedoc, spoke virtually no French throughout the early modern period. Although direct evidence is scant, this linguistic situation can be inferred from the 1863 survey which shows few Occitans throughout the region spoke French words well into the late

nineteenth century (Weber 1976, 68). At the same time, a small group of literate Occitans learned French (Cohen 2000). They had a stronger incentive to invest in French over developing Occitan, because the ability to use the language of the political center allowed them to negotiate directly with the monarch over taxation at the provincial assemblies. These literates could take advantage of the language market to gain access to the information that was available only in French and advocate the tax regime to their own benefit (Laitin 1988). Moreover, from the sixteenth century on these local elites formed a marriage of convenience with Parisian bureaucrats at the expense of the peasantry so that local elites could retain their privilege and the central government could raise revenue (Bates 1988; Root 1987).

The province of Béarn showcases this logic. Located along the westernmost borders between Spain and France, Béarn was united with France at the turn of the sixteenth century. As with other areas of Occitania, Paris installed indirect rule. Although Occitan was the *lingua franca*, local notables accepted French as the language of the court in the 1620 royal decree so they could negotiate the terms of local authorities directly with their new rulers in Paris (Cohen 2003, 176).[20] As this example suggests, there appeared to be no organized effort to codify the vernacular or standardize language use. The Occitan language faced inconsistent spelling and language use depended on authors' idiosyncrasies (Price 1998, 344). Meanwhile, the French founded in 1635 the *Académie française*, or the national academy, which covered the large fixed costs of studying and codifying the French language, but the Occitans built no comparable institution.

The nineteenth century saw sort of a Renaissance in intellectuals' attitudes toward local history and culture. Influenced by German Romanticism, scholarly interest in the Occitan language and the history of southern France grew. Both French and Occitan authors took up this effort. For instance, Jean-Bernard Mary-Lafon, a historian, dramatist, and linguist, published a series of influential works on the history of the Occitan region that recounts not just politics but also religion and

[20]Despite French became the language of the court, Occitan was used to keep the records of the deliberations of the provincial estate up until the French Revolution (Cohen 2003, 181).

literature (Roza 2003, 59). In addition, there were French works on the Troubadours, the poets and singers during the "golden age" of the Occitan culture developed in the thirteenth century, and on grammar books on the vernacular. Occitan literates, on their part, also composed poems in Occitan. These poems described the lives of the working-class people in the region and gained popularity among regular Occitans, as Occitan poets, who also belonged to the working class as bakers, hairdressers, and wig makers, traveled across the region (Roza 2003, 61). This growing intellectual effort to understand the history, language, and culture of the Occitans culminated in 1875 in the publication of the first comprehensive vernacular dictionary, *Dictionnaire analogique et étymologique des idiomes méridionaux* (Explanatory and etymological dictionary of the southern dialects).[21] This 1,344-page volume was edited by Louis Boucoiran, a little-known historian originally from Nîme. To be clear, the publication was not tantamount to the unification of various Occitan dialects. Although the growing investment in and the popularity of the Occitan language was a reaction to the parallel—and larger—effort on the standardization of the French language by the Third Republic, it seems that the Occitans had no political entrepreneur who took the effort to unify the Occitan tongues (Smith 1973). Vernacular codification took place, but there was little institution-building or public goods provision to consolidate Occitan culture and launch a political movement.

Effort on public goods provision after standardization

In the mid-nineteenth century, Occitans did mobilize for cultural revivalism. They founded the Félibrige, a literary society, in 1852 primarily to promote the greater use of the Occitan language throughout the Midi or the Occitan homeland. The most prominent member was Frédéric Mistral, a poet and one of the seven founding members, who wrote not only a series of poetry but also edited two-volume dictionaries in the late nineteenth century. His literary work was recognized with the

[21] The full title is *Dictionnaire analogique et étymologique des idiomes méridionaux qui sont parlés depuis Nice jusqu'à Bayonne et depuis les Pyrénées jusqu'au centre de la france, comprenant tous les termes vulgaires de la flore et de la faune méridionale, un grand nombre de citations prises dans les meilleurs auteurs, ainsi qu'une collection de proverbes locaux tirés de nos moralistes populaires.*

award of the Nobel prize for literature in 1904. The appellation, Félibrige, was derived from *felibre*, an Occitan word for a doctor which, in turn, denotes "teacher" and "sage" (Roza 2003, 74). The motive of the organization is summarized in Article 1 of the mission statement:

> The Félibrige is established to forever preserve for Provence its language, its color, freedom for its grace, its national honor and its great ranks of intelligence, for it is Provence that agrees with us. By Provence, we mean the entire Midi of France (Roza 2003, 96).

Roza (2003, 65) discerns two concrete goals from the record of the organization's activities. The first is to remove the social stigma attached to using the Occitan tongue; the second is to secure survival as a literary language while Occitan was widely spoken within southern France.[22] Contributing members sought to achieve these goals through publication on a regular basis and participation in social events.

The most important publication that the Félibrige put together was the *Armana Prouvençau*, an almanac. First released in 1855, this annual publication contained a collection of news, poems, and essays in Occitan. Literates decided on the almanac as an ideal outlet, because it would not just convey their thoughts in the vernacular but also chronicle events that were important specifically to the Occitans (Roza 2003, 76). The almanac's first edition was printed for around 500 copies, but it quickly became popular so that it sold up to 3,000 copies by 1872 and 10,000 copies annually in the decade between the 1880s and the 1890s (Roza 2003, 83). With the *Armana Prouvençau* Mistral sought to encourage lay readers to use Occitan in everyday communication without fear of being chastised not to use French. He argued:

> From a habit of imitation and bourgeois prejudice—which unfortunately is spreading more than ever—people had become accustomed to avoiding as "vulgar" the words most ingrained in Provençal speech, and ... commonly and quite uncritically used the corrupt, bastardized,

[22]Roza also makes an interesting observation that despite strong interest in language preservation, French was the working language among these literates of Occitan. See Footnote 21 in Roza (2003, 65).

> Frenchified forms that were spoken in the streets. Once we had made up our minds to write in the language of the people, [the Félibre] had to bring out and set off to advantage its characteristic bluntness, vigor, and richness of expression, and we agreed to write the language purely as it as it was spoken in places free of outside influence (Mistral 1986, 81).

When the Félibrige worked to establish its foothold, the political and social environment in France was an auspicious one. In 1868, Napoleon III lifted the imperial decree of 1852 on the press, which gave an impetus to publish on politically sensitive issues. More important, the mid- to late nineteenth century was a time for great uncertainties in France. There was a constant struggle for the constitutional system between monarchism and republicanism. The uncertainties were especially acute in 1870 following the sudden, unexpected defeat in the Franco-Prussian War. These created room for new ideologies, which some political entrepreneurs exploited to champion their causes. In the new Third Republic that began in 1871, several parties with new ideologies fought for supremacy to fill this ideological space (Hanson 2010, ch. 4). Ambitious leaders of the Savoyards and Bretons within France stoke nationalist sentiments among their respective populations for self-determination. By contrast, the political uncertainties gave other emerging political actors a pause. For example, Catholic elites were considering to form a Christian party and mobilize lay followers to secure their influence in the government. They feared that if "secular" Republicans came to power for a prolonged period, this would threaten the political power that the Church enjoyed under centuries of monarchism. The Catholics eventually chose not to form a party as they expected the republicans in power not to last long (Kalyvas 1996, ch. 3). The uncertain environment thus offered an historic opportunity for new ideologies to arise. This period was also prior to the introduction of free, mandatory primary education. Thus, in this context the Félibrige could have taken advantage of the instability in the political center to mobilize further lay Occitans and invest in public goods provision.

The uncertainties about *what type* of the constitutional structure would emerge and *how long* it would last would afforded a historic opportunity to demand a change in governance arrangements.

More specifically, demographic minorities seeking greater self-rule may pursue autonomy or some form of a federal system that can guarantee self-rule. In France, indirect rule gave way to more direct rule through the Revolution. In the initial stage of the process, Jacobin leaders improvised an institutional arrangement that granted commercial centers like Marseille and Lyon with regional autonomy. However, as the revolutionary government took shape, it instead installed a centralizing system of direct rule on key dimensions of governance like taxation, policing, and public goods across the country (Tilly 1992, 109–10).[23]

The transition period between the Second Empire and the Third Republic over the 1860s through early 1870s appeared another window of opportunity for institutional transformation. Extant scholarship hypothesizes that a federal system is likely to emerge when the gap in military capacity between the core unit and the peripheral units is small (Riker 1964). Alternatively, a country may adopt federation if subordinate units had the state capacity to enforce rules and administer policy within them and can raise revenue for the political center (Ziblatt 2004, 2006). Given that the political and military turmoil during this period left the state capacity of Paris relatively underdeveloped, it could give minority groups like Occitans an incentive to invest in public goods and state institutions to strengthen bargaining leverage and draw concessions from the Parisian authorities (Gerring et al. 2011, 380).[24]

Moreover, a group of French intellectuals supported federation as a constitutional foundation in the event of the collapse of monarchism toward the end of the Second Empire. Federalism emerged as one of the three major ideological strands among the Parisian intelligentsia.[25] French federalists used the commune as the primary political unit and highlighted unity based on liberty, equality, and loyalty to the state (patriotism) (Hazareesingh 1998, 284). At the same time, this ideology

[23]Historical works show that France transplanted direct rule—under banner of *mission civilisatrice* (civilizing mission)—in its colonial possessions. See, for instance, Conklin (1997).

[24]The recent literature shows that the incentives of politicians play a central role in understanding the institutional shapes of and the effects of federalism. See Rodden (2008) and Beramendi (2009) for review.

[25]The other two forces include Jacobinism, which believed in centralization as a remedy to the constitutional struggle, and municipalism, a compromise between federalism and Jacobinism.

advocated democracy as it espoused the consensus and self-governance of the communes. As Auguste Vermorel, a social republican, argues: "[Unity] should result from the free will of groups and their harmonious adhesion to common principles, which are the guarantee of the greatness, independence, and prosperity of the nation. It is the communes ... that should direct it" (Hazareesingh 1998, 284). Decentralization was deemed preferable to state centralization, another ideological current, because advocates believed that support for republicanism still remained weak in the countryside and that federation would compensate for this reality (Hazareesingh 1998, 285). Although federalists eventually gave way to Jacobinists who promoted centralization under the Third Republic (Hanson 2010, 101–2), federalism was a viable ideological force in the mid-nineteenth century.

In this political climate, the Félibrige continued to expand in the first decade of the Third Republic. It established the new branch in Paris, through which to interact with French intellectuals on a regular basis and make political and social networks in the capital. The new office was also used to raise public awareness about the linguistic situations in the south and garner support for the Occitan language. This strategy initially proved effective in securing the right to use the vernacular in higher education. In 1875, Occitan literates placed the petition in the National Assembly to restore the teaching and study of Occitan at the Universities of Aix, Toulouse, and Montpellier, and it was approved at Montpellier in three years (Roza 2003, 242). This initial investment in the public goods provision on the vernacular could have expanded to appeal to a wider circle of Occitan speakers beyond the literates. However, this was the apogee of institution-building by the Félibrige, and it seems that the hope of the greater use of the language through university education petered out in the mid-1870s.

One reason was certainly practical. In a time when the French government expanded primary education only in French while discouraging the use of non-French tongues, advocates were unable to find teachers who could or were willing to teach in Occitan in defiance of the government policy (Roza 2003, 243). Félibrige members thus had to watch new generations of school pupils learn and speak in French, which began to break the chain of Occitan-speaking generations in southern

France.

However, a more important reason is the absence of leadership. There is little evidence that suggests that the Félibrige sought to transform the literary society to launch a political platform. It appears that it was a deliberate decision as an organization not to seek regional autonomy or independence. Through the interactions in Paris, Occitan literates concluded that overt political activism for formal language rights would make too high risks to bear in an industrializing France where proficiency in French was not only a major political goal for cultural consolidation but also a highly desirable skill for productivity and upward mobility. In this environment that was adverse to cultural diversity, the Félibrige highlighted the importance of using Occitan, or non-French languages more broadly "to improve students' knowledge of French," because French was not a native tongue for most pupils in the south (Roza 2003, 248). This defensive strategy reflected the Félibrige's belief that French policymakers would never accept bilingual education in Occitan and French, let alone monolingualism in the Occitan-speaking region.

After the summer of 1886, the Félibrige abandoned its goal to press for language rights and promote the greater use of Occitan in France in favor of a more modest objective of language preservation. In April 1887, Alphonse Roque-Ferrier, a prominent member, contributed an article in the Montpellier daily, *Messager du Midi*:

> The goal of the Félibrige is then the maintenance of this language that we do not have to learn in prep school, and the desire to see it spoken with the same purity as the language of Corneille, the first serving the teaching of the second, and the second serving the teaching of the first (Roza 2003, 250).

This trajectory suggests that the lack of institution-building among the Occitan elites lies primarily in the absence of leadership to pursue formal rights to use the Occitan language at school or political activism for greater autonomy.

The absence of mobilization by the Félibrige can be further attributed to two factors. The first

is generational divide. The Félibrige was never a monolithic group; by contrast, there were significant differences in political goals across generations. On the one hand, young members embraced the greater ambition of a federal France, because Paris's initiative on institutional centralization was perceived as promoting homogeneity and thus a grave threat to cultural diversity within the country. For instance, Paul Meriéton, a young Félibre and one of the most vocal members calling for federation, argued that centralization would lead to "vaingloriousness" and the "depravation and death" of the Occitan culture that his organization was seeking to revive (Roche 1954, 27). In February 1892, he put forward his political goal in the "Déclaration des Félibres Fédéralistes," addressed to the president of the literary society:

> Yes, we want an Assembly at Bordeaux, at Toulouse, at Montpellier; we want one at Marseilles or at Aix. And these Assemblies must control the administration, the courts, the schools, and universities, as well as public works (Roche 1954, 64).

The statements by him and other young members who shared the view were soon printed in newspapers in both Occitan and French and widely circulated. They raised the suspicion from Paris that the Félibrige was harboring the intention to pursue not just federation but ultimately separatism, the charge that older members had carefully avoided (Wright 2003, 46–7).

Older generations, on the other hand, remained modest about the political ambitions for the literary society. As research on intellectual history finds, Frédéric Mistral, a founding member and the inaugural president, always equivocated the questions about the group's political goals or the politicization of its activity. While he seemed sympathetic to younger members' call for a political turn, he neither condoned nor reproved such demand and left virtually no writings on politics (Wright 2003, 20, 47). It seems that the primary goal of the older generations was a renaissance of what they perceived as the distinct civilization that had achieved its apex prior to Capetian incorporation in the wake of the Albigensian Crusade. Evidence from older generations' remarks suggests that the victory of the Crusaders constituted the "sources of spiritual life dried up." As Mistral noted: "because, what really subdued, note it well, was less the South materially speaking, than the spirit

of the South" (Roche 1954, 52). The Félibrige's leadership consistently reminded that its purpose was cultural preservation, and Mistral maintained that only through loyalty to *la grande patrie* (the big, native land of France) could Occitans' language and culture achieve "communal glory" (Roche 1954, 65).

The second reason for the lack of political mobilization is the internal discord within the literary society. Julian Wright (2003, 48) points out that the generational divide was found not just over the organization's political goals but also over the politics within the organization. Elections, titles, and statutes were main sources of the fight. The leadership positions and titles were scarce goods, and office- or title-holders could exercise disproportionate influence over the group's political future. For example, these members can ultimately decide whether to pursue greater language use in school instruction or whether to form a political party. Given that the Félibrige was the leading literary society of the Occitans, access to these goods would be highly desirable, or even essential, if some members espoused the ambition for greater autonomy. The presence of bureaucratic politics suggests that founding members occupied senior positions while restricting younger members access to them who were dissatisfied with their leaders' apolitical attitudes toward group activity. The 1892 declaration of the demand for federalism could be characterized as an outburst of the internal discord.

The generational divide eventually led to the split of the organization. Meriéton and other like-minded young members left the Félibrige to join an existing political party or found a new one. The breakup was the harbinger of the more formal—yet, again, unsuccessful—political movement of regionalism, most notably the FRF (Fédération régionaliste française), that arose in the twentieth century. Occitan speakers led the way by inheriting their ideological inspirations from the federalist strand of the constitutional debate that French intellectuals had toward the end of the Second Empire (Beer 1980; Wright 2003).

For the Félibrige, the turn of the century was the period of sharp decline as a literary society. More and more school-age children went to school and studied only in French in preference of Occ-

itan (or other non-French languages spoken at home) (Weber 1976). Vocal young members quit the organization. Senior members, by contrast, had a modest success at securing the right to teach Occitan at one university but largely abandoned the strategy of persuading the Parisian counterparts to recognize bilingual education in favor of a more moderate, practical goal of language preservation. As founding members aged and began to die, the organization's activity shrank in terms of the volume of publications and the number of pubic events (Roza 2003, 274). Although the Félibrige has still been in existence today, their role has remained the same—language and cultural preservation.

This case study examines how the Occitans failed to provide public goods and build institutions for a more consolidated nation. It shows that although they never became independent again after the thirteenth century or achieved *de jure* self-rule, this outcome is not predetermined. Rather, they could have succeeded at developing a nation: language standardization was achieved by 1900, an active literary society was founded in the mid-nineteenth century to advocate the greater awareness of the vernacular, the vernacular publication project was a success, and the organization expanded. In the latter half of the nineteenth century, the political environment also seemed favorable to non-French-speaking minorities with political ambitions like Occitan intellectuals. In France, the constitutional foundations were unstable and unpredictable in the early to mid-nineteenth century, when monarchism and republicanism alternated with relative frequency. Félibrige leaders could have exploited this turmoil to mobilize lay Occitans and perhaps organize a political party that would call for regional autonomy. However, they chose not to politicize the issues of the Occitan language and kept the salience of Occitan ethnicity low. There was no organized resistance to the French government's top-down effort for language unification, which blew the prospects that the Occitans develop into a consolidated nation. It was a missed opportunity due to the absence of leadership.

4.5 Conclusion

This paper accounts for the failure of building a nation based on a case study of the Occitan ethnic group. The Occitans achieved early language standardization but did not take advantage of this

development for further cultural consolidation and political authority. This case study is among the first attempts to theorize about failure. It offers a new hypothesis on this outcome in which the absence of leadership in an organization of ethnic group elites can undermine the opportunity to become a more coherent group. This hypothesis is overlooked in extant scholarship, mainly because most if not all ethnic groups under study invest in building institutions and provide public goods to ensure survival.

In addition, my study has broader implications for understanding the relationship between ethnic diversity and public goods provision. It shows that cultural consolidation does not necessarily beget public goods provision or institution-building. This outcome is contrary to the stylized fact in which high diversity is likely to result in the inadequate provision of public goods while homogeneity is expected to yield greater provision. My study highlights the importance of considering within-group behavior to understand variation in public goods provision among ethnic groups.

This paper offers lessons for future research on the success and failure of nations. More specifically, it would be useful to examine emerging organizations, such as literary societies, among ethnic groups. Even demographically small groups have some cultural advocacy organization. Some resist the institutions imposed by the political center in society and launch a political campaign for autonomy or independence, while others, like the Félibrige, pursue a more apolitical path for cultural preservation. But this empirical question is not yet systematically investigated. This paper suggests that examining these cultural organizations may offer a clue. For instance, the researcher can collect attributes of these organizations on the dimensions of endowment, political ambitions, geography, etc. This approach opens a new avenue for addressing variation not only in the development of nations but also in public goods provision across ethnically diverse societies.

BIBLIOGRAPHY

Abramson, Scott F. 2017. "The Economic Origins of the Territorial State." *International Organization* 71(1): 97–130.

Acemoglu, Daron, and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2005. "The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth." *American Economic Review* 95(2): 546–79.

Ager, Dennis. 1999. *Identity, Insecurity and Image: France and Language*. Clevedon: Multilingual Matters.

Ahlerup, Pelle, and Ola Olsson. 2012. "The Roots of Ethnic Diversity." *Journal of Economic Growth* 17(2): 71–102.

Åhlfeldt, Johan. 2015. "Digital Atlas of the Roman Empire.".
URL: `http://dare.ht.lu.se/`

Aktürk, Şener. 2012. *Regimes of Ethnicity and Nationhood in Germany, Russia, and Turkey*. New York: Cambridge University Press.

Albaugh, Ericka A. 2015. *State-Building and Multilingual Education in Africa*. New York: Cambridge University Press.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8(2): 155–94.

Alesina, Alberto, and Eliana La Ferrara. 2005. "Ethnic Diversity and Economic Performance." *Journal of Economic Literature* 43(3): 762–800.

Alesina, Alberto, Reza Baqir, and William Easterly. 1999. "Public Goods and Ethnic Divisions." *Quarterly Journal of Economics* 114(4): 1243–84.

Algan, Yann, and Pierre Cahuc. 2014. "Trust, Growth, and Well-Being: New Evidence and Policy Implications". In *Handbook of Economic Growth*, ed. Philippe Aghion and Steven N. Durlauf. Vol. 2A New York: Elsevier, pp. 49–120.

Amano, Tatsuya, Brody Sandel, Heidi Eager, Edouard Bulteau, Jens-Christian Svenning, Bo Dalsgaard, Carsten Rahbek, Richard G. Davies, and William J. Sutherland. 2014. "Global Distribution and Drivers of Language Extinction Risk." *Proceedings of the Royal Society B: Biological Sciences* 281(1793): 20141574.

Anderson, Benedict. 2006. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Rev. ed. London: Verso.

Ashraf, Quamrul, and Oded Galor. 2013. "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development." *American Economic Review* 103(1): 1–46.

Austin, Peter K., and Julia Sallabank. 2011. "Introduction". In *Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank. New York: Cambridge University Press, pp. 1–24.

Bairoch, Paul, Jean Batou, and Pierre Chévre. 1988. *La Population des Villes Européennes de 800 à 1850: Banque de Données et Analyse Sommaire des Résultats (The Population of European Cities from 800 to 1850: Data Bank and Short Summary of Results)*. Genéve: Librairie Droz.

Barth, Fredrik. 1969. "Introduction". In *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*, ed. Fredrik Barth. Oslo: Johansen and Nielsen Boktrykkeri, pp. 9–38.

Bates, Robert H. 1983. "Modernization, Ethnic Competition, and the Rationality of Politics in Contemporary Africa". In *State versus Ethnic Claims: African Policy Dilemmas*, ed. Donald Rothchild and Victor A. Olorunsola. Boulder, CO: Westview Press, pp. 152–71.

Bates, Robert H. 1988. "Lessons from History, or the Perfidy of English Exceptionalism and the Significance of Historical France." *World Politics* 40(4): 499–516.

Bean, Richard. 1973. "War and the Birth of the Nation State." *The Journal of Economic History* 33(1): 203–21.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124(2): 531–96.

Becker, Sascha O., and Ludger Woessmann. 2011. "The Effects of the Protestant Reformation on Human Capital". In *The Oxford Handbook of the Economics of Religion*, ed. Rachel M. McCleary. New York: Oxford University Press, pp. 93–110.

Becker, Sascha O., Steven Pfaff, and Jared Rubin. 2016. "Causes of Consequences of the Protestant Reformation." *Explorations in Economic History* 62: 1–25.

Beer, William R. 1980. *The Unexpected Rebellion: Ethnic Activism in Contemporary France*. New York and London: New York University Press.

Beik, William. 1985. *Absolutism and Society in Seventeenth-century France: State Power and Provincial Aristocracy in Languedoc*. Cambridge: Cambridge University Press.

Beik, William. 2005. "The Absolutism of Louis XIV as Social Collaboration." *Past and Present* 188: 195–224.

Bell, David A. 2001. *The Cult of the Nation in France: Inventing Nationalism, 1680–1800*. Cambridge, MA: Harvard University Press.

Bellwood, Peter. 2001. "Early Agriculturalist Population Diasporas? Farming, Languages, and Genes." *Annual Review of Anthropology* 30: 181–207.

Bennett, Andrew, and Jeffrey T. Checkel. 2014. "Process Tracing: From Philosophical Roots to Best Practices". In *Process Tracing: From Metaphor to Analytic Tool*, ed. Andrew Bennett and Jeffrey T. Checkel. New York: Cambridge University Press, pp. 3–37.

Beramendi, Pablo. 2009. "Federalism". In *Oxford Handbook of Comparative Politics*, ed. Carles Boix and Susan C. Stokes. New York: Oxford University Press, pp. 752–81.

Bernstein, William J. 2013. *Masters of the Word: How Media Shaped History from the Alphabet to the Internet*. New York: Grove Press.

Biggins, Michael, and Janet Crayne, eds. 2000. *Publishing in Yugoslavia's Successor States*. New York: Haworth Information Press.

Biggs, Michael. 1999. "Putting the State on the Map: Cartography, Territory, and European State Formation." *Comparative Studies in Society and History* 42(2): 374–405.

Blaydes, Lisa, and Eric Chaney. 2013. "The Feudal Revolution and Europe's Rise: Political Divergence of the Christian West and the Muslim World before 1500 CE." *American Political Science Review* 107(1): 16–34.

Blockmans, Wim P. 1976. "Le Régime Reprèsentatif en Flandre dans le Cadre Européen au bas Moyen Age avec un Projet d'Application des Ordinateurs". In *Album Elémer Màlyusz: Studies Presented to the International Commission for the History of Representative and Parliamentary Institutions. (Études Présentées à la Commission Internationale pour l'Histoire des Assemblees d'États.)*. Brussels: LVI, pp. 211–45.

Blockmans, Wim P. 1994. "Voracious States and Obstructing Cities: An Aspect of State Formation

in Preindustrial Europe". In *Cities and the Rise of States in Europe, A.D. 1000 to 1800*, ed. Charles Tilly and Wim P. Blockmans. New York: Cambridge University Press, pp. 218–20.

Boix, Carles. 2003. *Democracy and Redistribution*. New York: Cambridge University Press.

Boix, Carles, Bruno Codenotti, and Giovanni Resta. 2011. "War, Wealth and Formation of States". In *Political Economy of Institutions, Democracy and Voting*, ed. Norman Schofield and Gonzalo Caballero. New York: Springer, pp. 45–74.

Boix, Carles, and Susan C. Stokes. 2003. "Endogenous Democratization." *World Politics* 55(4): 517–49.

Boppart, Timo, Josef Falkinger, and Volker Grossmann. 2014. "Protestantism and Education: Reading (the Bible) and Other Skills." *Economic Inquiry* 52(2): 874–95.

Bosker, Maarten, Eltjo Buringh, and Jan Luiten van Zanden. 2013. "From Baghdad to London: Unraveling Urban Development in Europe, the Middle East, and North Africa, 800–1800." *Review of Economics and Statistics* 95(4): 1418–37.

Bouchot, Henri. 1890. *The Book: Its Printers, Illustrators, and Binders, from Gutenberg to the Present Time*. London: H. Grevel & Co.

Box-Steffensmeier, Janet M., and Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press.

Brady, Henry E., and David Collier, eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 2nd ed. Lanham, MD: Rowman and Littlefield.

Branch, Jordan. 2014. *The Cartographic State: Maps, Territory, and the Origins of Sovereignty*. New York: Cambridge University Press.

Brecke, Peter. 1999. Violent Conflicts 1400 A.D. to the Present in Different Regions of the World. Paper presented at the 1999 Meeting of the Peace Science Society.

Breuilly, John. 1993. *Nationalism and the State*. 2nd ed. Chicago: University of Chicago Press.

Brown, Keith, ed. 2006. *Encyclopedia of Language and Linguistics*. Vol. 1–6. 2nd ed. London: Elsevier.

Brubaker, Rogers. 1992. *Citizenship and Nationhood in France and Germany*. Cambridge, MA: Harvard University Press.

Brubaker, Rogers. 1996. *Nationalism Reframed: Nationhood and the National Question in the New Europe*. New York: Cambridge University Press.

Brubaker, Rogers. 2004. *Ethnicity without Groups*. Cambridge, MA: Harvard University Press.

Brubaker, Rogers, and David D. Laitin. 1998. "Ethnic and Nationalist Violence." *Annual Review of Sociology* 24: 423–52.

Buringh, Eltjo, and Jan Luiten van Zanden. 2009. "Charting the 'Rise of the West': Manuscripts and Printed Books in Europe, A Long-Term Perspective from Sixth through Eighteenth Centuries." *The Journal of Economic History* 69(2): 409–45.

Burke, Peter. 2004. *Languages and Communities in Early Modern Europe*. Cambridge: Cambridge University Press.

Cagé, Julia, and Valeria Rueda. 2016. "The Long-Term Effects of the Printing Press in sub-Saharan Africa." *American Economic Journal: Applied Economics* 8(3): 69–99.

Campbell, Peter R. 2012. "Absolute Monarchy". In *The Oxford Handbook of the Ancien Régime*, ed. William Doyle. New York: Oxford University Press, pp. 11–38.

Cantoni, Davide. 2015. "The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands." *Journal of European Economic Association* 13(4): 561–68.

Cantoni, Davide, Jeremiah Dittmar, and Noam Yuchtman. 2016. "Reformation and Reallocation: Religious and Secular Economic Activity in Early Modern Germany." Mimeo.

Cantoni, Davide, and Noam Yuchtman. 2014. "Medieval Universities, Legal Institutions, and the Commercial Revolution." *Quarterly Journal of Economics* 129(2): 823–87.

Cardinal, Linda, and Selma K. Sonntag, eds. 2015. *State Traditions and Language Regimes*. Montréal: McGill-Queen's University Press.

Carter, David B., and Curtis S. Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18(3): 271–92.

Caruana-Galizia, Paul, and Jordi Martí-Henneberg. 2013. "European Regional Railways and Real Income, 1870–1910: A Preliminary Report." *Scandinavian Economic History Review* 61(2): 167–96.

Chanda, Areendam, and Louis Putterman. 2007. "Early Starts, Reversals and Catch-up in the Process of Economic Development." *Scandinavian Journal of Economics* 109(2): 387–413.

Chandra, Kanchan. 2004. *Why Ethnic Parties Succeed: Patronage and Ethnic Head Counts in India*. New York: Cambridge University Press.

Chandra, Kanchan. 2006. "What Is Ethnic Identity and Does It Matter?" *Annual Review of Political Science* 9: 397–414.

Chaney, David M. 2015. "The Hierarchy of the Catholic Church: Current and Historical Information about Its Bishops and Dioceses.".
URL: `www.catholic-hierarchy.org`

Cipolla, Carlo M. 1972. "The Diffusion of Innovations in Early Modern Europe." *Comparative Studies in Society and History* 14(1): 46–52.

Clair, Colin. 1976. *A History of European Printing*. New York: Academic Press.

Coşgel, Metin M., Thomas J. Miceli, and Jared Rubin. 2012. "The Political Economy of Mass Printing: Legitimacy, Revolt, and Technological Change in the Ottoman Empire." *Journal of Comparative Economics* 40(3): 357–71.

Cohen, Paul. 2000. "Of Linguistic Jacobininsm and Cultural Balkanization: Contemporary French Linguistic Politics in Historical Context." *French Politics, Culture, and Society* 18(3): 21–48.

Cohen, Paul. 2003. "Linguistic Politics on the Periphery: Louis XIII, Béarn, and the Making of French as an Official Language in Early Modern France". In *When Languages Collide: Perspectives on Language Conflict, Language Competition, and Language Coexistence*, ed. Brian D. Joseph, Johanna DeStefano, Neil G. Jacobs and Ilse Lehiste. Columbus: The Ohio State University Press, pp. 165–200.

Colley, Linda. 2005. *Britons: Forging the Nation, 1707–1837*. 2nd ed. New Haven: Yale University Press.

Comin, Diego, William Easterly, and Erick Gong. 2010. "Was the Wealth of Nations Determined in 1000 BC?" *American Economic Journal: Macroeconomics* 2(3): 65–97.

Conklin, Alice L. 1997. *A Mission to Civilize: The Republican Idea of Empire in France and West Africa, 1895–1930*. Stanford: Stanford University Press.

Conner, Philip. 2001. "A Provincial Perspective: Protestant Print Culture in Southern France". In *The Sixteenth-Century French Religious Book*, ed. Andrew Pettegree, Paul Nelles, and Philip Conner. Aldershot: Ashgate, pp. 286–302.

Connor, Walker. 1994. *Ethnonationalism: The Quest for Understanding*. Princeton: Princeton University Press.

Dalby, Andrew. 1998. *A Guide to World Language Dictionaries*. London: Library Association Publishing.

Darby, H.C., and Harold Fullard, eds. 1970. *The New Cambridge Modern History*. Vol. 14: *Atlas*. Cambridge: Cambridge University Press.

de Vries, Jan. 1984. *European Urbanization, 1500–1800*. Cambridge, MA: Harvard University Press.

DeLong, J. Bradford, and Andrei Shleifer. 1993. "Princes and Merchants: European City Growth before the Industrial Revolution." *Journal of Law and Economics* 36(2): 671–702.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. 2012. "The Political Economy of Linguistic Cleavages." *Journal of Development Economics* 97(2): 322–38.

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

Deutsch, Karl W. 1966. *Nationalism and Social Communication: An Inquiry into the Foundations of Nationality*. 2nd ed. Cambridge, MA: MIT Press.

Dincecco, Mark. 2011. *Political Transformations and Public Finances: Europe, 1650–1913*. New York: Cambridge University Press.

Dincecco, Mark. 2015. "The Rise of Effective States in Europe." *The Journal of Economic History* 75(3): 901–18.

Dittmar, Jeremiah E. 2011. "Information Technology and Economic Change: The Impact of Printing Press." *Quarterly Journal of Economics* 126(3): 1133–75.

Dittmar, Jeremiah E., and Ralf R. Meisenzahl. 2016. "State Capacity and Public Goods: Institutional Change, Human Capital, and Growth in Early Modern Germany." Mimeo.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-based Approach*. New York: Cambridge University Press.

Dunning, Thad. 2014. "Improving Process Tracing: The Case of Multi-Method Research". In *Process Tracing: From Metaphor to Analytic Tool*, ed. Andrew Bennett and Jeffrey T. Checkel. New York: Cambridge University Press, pp. 211–36.

Easterlin, Richard A. 1981. "Why Isn't the Whole World Developed?" *The Journal of Economic History* 41(1): 1–19.

Easterly, William, and Ross Levine. 1997. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *Quarterly Journal of Economics* 112(4): 1203–50.

Eisenstein, Elizabeth L. 1979. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-modern Europe*. Cambridge: Cambridge University Press.

Elliott, J.H. 1992. "A Europe of Composite Monarchies." *Past and Present* 137: 48–71.

Ertman, Thomas. 1997. *Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe*. New York: Cambridge University Press.

Fearon, James D. 1999. "Why Ethnic Politics and 'Pork' Tend to Go Together." Mimeo.

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. 2000. "Violence and the Social Construction of Ethnic Identity." *International Organization* 54(4): 825–44.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97(1): 75–95.

Febvre, Lucien. 1973. "*Frontiére*: The Word and the Concept". In *A New Kind of History: From the Writings of Febvre*, ed. Peter Burke. New York, Evanston, and San Francisco: Harper and Row, pp. 208–18.

Febvre, Lucien, and Henri-Jean Martin. 1976. *The Coming of the Book: The Impact of Printing, 1450–1800*. 3rd ed. London: Verso.

Friedman, David. 1977. "A Theory of the Size and Shape of Nations." *Journal of Political Economy* 85(1): 59–77.

Frijhoff, Willem. 1996. "Patterns". In *A History of the University in Europe*, ed. Hilde de Ridder-Symoens. Vol. 2: *Universities in Early Modern Europe*. Cambridge: Cambridge University Press, pp. 43–110.

Gellner, Ernest. 2006. *Nations and Nationalism*. 2nd ed. Malden, MA: Blackwell.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerring, John. 2007. *Case Study Research: Principles and Practices*. New York: Cambridge University Press.

Gerring, John, Daniel Ziblatt, Johan van Gorp, and Julián Arévalo. 2011. "An Institutional Theory of Direct and Indirect Rule." *World Politics* 63(3): 377–433.

Ginsburgh, Victor, and Shlomo Weber. 2011. *How Many Languages Do We Need? The Economics of Linguistic Diversity*. Princeton: Princeton University Press.

Given, James. 1990. *State and Society in Medieval Europe: Gwynedd and Languedoc under Outside Rule*. Ithaca: Cornell University Press.

Glaeser, Edward L., and Albert Saiz. 2004. "The Rise of the Skilled City." *Brookings-Wharton Papers on Urban Affairs* pp. 47–105.

GLOBE Task Team and others. 1999. "The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model.".
URL: http://www.ngdc.noaa.gov/mgg/topo/globe.html

Glynn, Adam N., and Nahomi Ichino. 2015. "Using Qualitative Information to Improve Causal Inference." *American Journal of Political Science* 59(4): 1055–71.

Glynn, Adam N., and Nahomi Ichino. Forthcoming. "Increasing Inferential Leverage in the Comparative Method: Placebo Tests in Small-*n* Research." *Sociological Methods and Research* pp. 1–32.
URL: http://smr.sagepub.com/content/early/2014/05/01/0049124114528879

Goldin, Claudia. 2001. "The Human-Capital Century and American Leadership: Virtues of the Past." *The Journal of Economic History* 61(2): 263–92.

Gorski, Philip S. 2003. *The Disciplinary Revolution: Calvinism and the Rise of the State in Early Modern Europe*. Chicago: University of Chicago Press.

Graff, Harvey J. 1987. *The Legacies of Literacy: Continuities and Contradictions in Western Culture and Society*. Bloomington: Indiana University Press.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2006. "Does Culture Affect Economic Outcomes?" *Journal of Economic Perspectives* 20(2): 23–48.

Gurr, Ted Robert. 1971. *Why Men Rebel*. Princeton: Princeton University Press.

Habyarimana, James, Macartan Humphreys, Daniel N. Posner, and Jeremy M. Weinstein. 2007. "Why Does Ethnic Diversity Undermine Public Goods Provision?" *American Political Science Review* 101(4): 709–25.

Habyarimana, James, Macartan Humphreys, Daniel N. Posner, and Jeremy M. Weinstein. 2009. *Coethnicity: Diversity and the Dilemmas of Collective Action*. New York: Russell Sage Foundation.

Hale, John. 2007. "Warfare and Cartography, ca. 1450 to ca. 1640". In *The History of Cartography*, ed. David Woodward. Vol. 3: *Cartography in the European Renaissance*. Part I. Chicago and London: University of Chicago Press, pp. 719–37.

Hammond, Nicholas G.L., ed. 1981. *Atlas of the Greek and Roman World in Antiquity*. Park Ridge, NJ: Noyes Press.

Hansen, Jason D. 2015. *Mapping the Germans: Statistical Science, Cartography, and the Visualization of the German Nation, 1848–1914*. New York: Oxford University Press.

Hanson, Stephen E. 2010. *Post-Imperial Democracies: Ideology and Party Formation in Third Republic France, Weimar Germany, and Post-Soviet Russia*. New York: Cambridge University Press.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3): 607–68.

Hardie, D. W. F. 1948. *A Handbook of Modern Breton (Armorican)*. Cardiff: University of Wales Press.

Hariri, Jacob Gerner. 2012. "The Autocratic Legacy of Early Statehood." *American Political Science Review* 106(3): 471–94.

Hazareesingh, Sudhir. 1998. *From Subject to Citizen: The Second Empire and the Emergence of Modern French Democracy*. Princeton: Princeton University Press.

Hechter, Michael. 2000. *Containing Nationalism*. New York: Oxford University Press.

Hirsch, Francine. 2005. *Empire of Nations: Ethnographic Knowledge and the Making of the Soviet Union*. Ithaca: Cornell University Press.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

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

Hobsbawm, Eric, and Terence Ranger, eds. 1983. *The Invention of Tradition*. Cambridge: Cambridge University Press.

Horowitz, Donald L. 1985. *Ethnic Groups in Conflict*. Berkeley and Los Angeles: University of California Press.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20(1): 1–24.

Kagan, Richard L., and Benjamin Schmidt. 2007. "Maps and the Early Modern State: Official Cartography". In *The History of Cartography*, ed. David Woodward. Vol. 3: *Cartography in the European Renaissance*. Part I. Chicago and London: University of Chicago Press, pp. 661–79.

Kalyvas, Stathis N. 1996. *The Rise of Christian Democracy in Europe*. Ithaca: Cornell University Press.

Kasfir, Nelson. 1979. "Explaining Ethnic Political Participation." *World Politics* 31(3): 365–88.

Kertzer, David I., and Dominique Arel, eds. 2002. *Census and Identity: The Politics of Race, Ethnicity, and Language in National Censuses*. New York: Cambridge University Press.

King, Charles. 2010. *Extreme Politics: Nationalism, Violence, and the End of Eastern Europe*. New York: Oxford University Press.

King, Gary, and Langche Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* 51(1): 183–210.

Kiser, Edgar, and April Linton. 2002. "The Hinges of History: State-Making and Revolt in Early Modern France." *American Sociological Review* 67(6): 889–910.

Koenigsberger, H.G. 1987. *Early Modern Europe, 1500–1789*. London: Longman.

Kohn, Hans. 2005. *The Idea of Nationalism: A Study in Its Origins and Background*. New Brunswick, NJ: Transaction.

Krasner, Stephen D. 1999. *Sovereignty: Organized Hypocrisy*. Princeton: Princeton University Press.

Krasner, Stephen D. 2001. "Rethinking the Sovereign State Model." *Review of International Studies* 27(5): 17–42.

Kwass, Michael. 2000. *Privilege and the Politics of Taxation in Eighteenth-Century France: Liberté, Égalité, Fiscalité*. Cambridge: Cambridge University Press.

Kymlicka, Will. 2009. *Multicultural Odysseys: Navigating the New International Politics of Diversity*. New York: Oxford University Press.

La Porta, Rafael, Florencio Lopez de Silanes, Andrei Shleifer, and Robert Vishny. 1999. "The Quality of Government." *Journal of Law, Economics, and Organization* 15(1): 222–79.

Laitin, David D. 1988. "Language Games." *Comparative Politics* 20(3): 289–302.

Laitin, David D. 1998. *Identity in Formation: The Russian-Speaking Population in the Near Abroad*. Ithaca: Cornell University Press.

Laitin, David D. 2007. *Nations, States, and Violence*. New York: Oxford University Press.

Laitin, David D., Joachim Moortgat, and Amanda Lea Robinson. 2012. "Geographical Axes and the Persistence of Cultural Diversity." *Proceedings of the National Academy of Sciences* 109(26): 10263–68.

Laitin, David D., and Rajesh Ramachandran. 2016. "Language Policy and Human Development." *American Political Science Review* 110(3): 457–80.

Landau, Sidney I. 2001. *Dictionaries: The Art and Craft of Lexicography*. 2nd ed. New York: Cambridge University Press.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, eds. 2013. *Ethnologue: Languages of the World*. 17th ed. Dallas, TX: SIL International.
URL: `www.ethnologue.com/`

Lindert, Peter H. 2003. "Voice and Growth: Was Churchill Right?" *The Journal of Economic History* 63(2): 315–50.

Lindgren, Uta. 2007. "Land Surveys, Instruments, and Practitioners in the Renaissance". In *The History of Cartography*, ed. David Woodward. Vol. 3: *Cartography in the European Renaissance*. Part I. Chicago and London: University of Chicago Press, pp. 477–508.

Lipset, Seymour Martin. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53(1): 69–105.

Liu, Amy H. 2014. *Standardizing Diversity: The Political Economy of Language Regimes*. Philadelphia: University of Pennsylvania Press.

Mahoney, James. 2003. "Strategies of Causal Assessment in Comparative Historical Analysis". In *Comparative Historical Analysis in the Social Sciences*, ed. James Mahoney and Dietrich Rueschemeyer. New York: Cambridge University Press, pp. 337–72.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62(1): 120–47.

Martin, Peter. 2008. *Samuel Johnson: A Biography*. Cambridge, MA: Harvard University Press.

Martin, Terry. 2001. *The Affirmative Action Empire: Nations and Nationalism in the Soviet Union, 1923–1939*. Ithaca: Cornell University Press.

Marvel, Seth A., Travis Martin, Charles R.Doering, David Lusseau, and M.E.J. Newman. 2013. "The Small-World Effect Is a Modern Phenomenon." Mimeo.

Marx, Anthony. 2003. *Faith in Nation: Exclusionary Origins of Nationalism*. New York: Oxford University Press.

McCloskey, Deirdre N. 2016. *Bourgeois Equality: How Ideas, Not Capital or Institutions, Enriched the World*. Chicago: University of Chicago Press.

McCusker, John J. 2005. "The Demise of Distance: The Business Press and the Origins of the Information Revolution in the Early Modern Atlantic World." *American Historical Review* 110(2): 295–321.

McKitterick, David. 1998. "The Beginning of Printing". In *The New Cambridge Medieval History*, ed. Christopher Allmand. Vol. 7: c.1415–c.1500. Cambridge: Cambridge University Press, pp. 287–98.

McNeill, William H. 1982. *The Pursuit of Power: Technology, Armed Force and Society since A.D. 1000*. Chicago: University of Chicago Press.

Menaldo, Victor. 2012. "The Middle East and North Africa's Resilient Monarchs." *Journal of Politics* 74(3): 707–22.

Michalopoulos, Stelios. 2012. "The Origins of Ethnolinguistic Diversity." *American Economic Review* 102(4): 1508–39.

Miguel, Edward. 2004. "Tribe or Nation? Nation Building and Public Goods in Kenya versus Tanzania." *World Politics* 56(3): 323–62.

Miguel, Edward, and Mary Kay Gugerty. 2005. "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya." *Journal of Public Economics* 89(11–12): 2325–68.

Minahan, James B. 2000. *One Europe, Many Nations: A Historical Dictionary of European National Groups*. Westport, CO: Greenwood Press.

Mistral, Frédéric. 1986. *The Memoirs of Frédéric Mistral*. Translated by George Wickes. New York: New Directions.

Mitchell, Brian R. 2003*a*. *International Historical Statistics: Africa, Asia, and Oceania, 1750–2000*. 4th ed. New York: Palgrave Macmillan.

Mitchell, Brian R. 2003*b*. *International Historical Statistics: Europe, 1750–2000*. 5th ed. New York: Palgrave Macmillan.

Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton: Princeton University Press.

Mokyr, Joel. 2006. "Long-term Economic Growth and the History of Technology". In *Handbook of Economic Growth*, ed. Philippe Aghion and Steven N. Durlauf. Vol. 1B. New York: Elsevier, pp. 1113–80.

Mokyr, Joel. 2016. *A Culture of Growth: The Origins of the Modern Economy*. Princeton: Princeton University Press.

Murphy, Alexander B. 2002. "National Claims to Territory in the Modern State System: Geographical Considerations." *Geopolitics* 7(2): 193–214.

Mynolas, Harris. 2013. *The Politics of Nation-Building: Making Co-Nationals, Refugees and Minorities*. New York: Cambridge University Press.

Nexon, Daniel H. 2009. *The Struggle for Power in Early Modern Europe: Religious Conflict, Dynastic Empires, and International Change*. Princeton: Princeton University Press.

Nordhoff, Sebastian, Harald Hammarström, Robert Forkel, and Martin Haspelmath, eds. 2013. *Glottolog 2.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
URL: `http://glottolog.org`

North, Douglass C. 1981. *Structure and Change in Economic History*. New York: W.W. Norton.

North, Douglass C., and Barry R. Weingast. 1989. "Constitutions and Commitment: The Evolution of Institutional Governing Public Choice in Seventeenth-Century England." *The Journal of Economic History* 49(4): 803–31.

North, Douglass C., and Robert Paul Thomas. 1977. "The First Economic Revolution." *The Journal of Economic History* 30(2): 229–41.

Nunn, Nathan. 2014. "Historical Development". In *Handbook of Economic Growth*, ed. Philippe Aghion and Steven N. Durlauf. Vol. 2A. New York: Elsevier, pp. 347–402.

Nunn, Nathan, and Diego Puga. 2012. "Ruggedness: The Blessing of Bad Geography in Africa." *Review of Economics and Statistics* 94(1): 20–36.

Nüssli, Christos. 2011. "Euratlas.".
URL: `http://www.euratlas.com/about.html`

Oakes, Leigh. 2001. *Language and National Identity: Comparing France and Sweden*. Amsterdam and Philadelphia: John Benjamins Publishing.

O'Brien, Patrick. 1977. *The New Economic History of the Railways*. London: Croom Helm.

O'Brien, Patrick, ed. 1983. *Railways and the Economic Development of Western Europe, 1830–1914*. London and Basingstoke: Macmillan.

Olsson, Ola, and Christopher Paik. 2016. "Long-run Cultural divergence: Evidence from the Neolithic Revolution." *Journal of Development Economics* 122: 197–213.

Olsson, Ola, and Douglas A. Hibbs. 2005. "Biogeography and Long-run Economic Development." *European Economic Review* 49(4): 909–38.

Osiander, Andreas. 2001a. "Before Sovereignty: Society and Politics in *Ancien Régime* Europe." *Review of International Studies* 27(5): 119–45.

Osiander, Andreas. 2001b. "Sovereignty, International Relations, and the Westphalian Myth." *International Organization* 55(2): 251–87.

Ostler, Nicholas. 2005. *Empires of the Word: A Language History of the World*. New York: HarperCollins.

Paterson, Linda M. 1993. *The World of the Troubadours: Medieval Occitan Society, c. 1100–c. 1300*. Cambridge: Cambridge University Press.

Pettegree, Andrew. 2007. *The French Book and the European Book World Renaissance*. Leiden and Boston: Brill.

Pettegree, Andrew. 2010. *The Book in the Renaissance*. New Haven: Yale University Press.

Pfaff, Steven, and Katie E. Corcoran. 2012. "Piety, Power, and the Purse: Religious Economies Theory and Urban Reform in the Holy Roman Empire." *Journal for the Scientific Study of Religion* 51(4): 757–76.

Philpott, Daniel. 2001. *Revolutions in Sovereignty: How Ideas Shaped Modern International Relations*. Princeton: Princeton University Press.

Pleiades. 2015.
URL: `http://pleiades.stoa.org/home`

Poggi, Gianfranco. 1978. *The Development of the Modern State: A Sociological Introduction*. Stanford: Stanford University Press.

Pounds, Norman J.G. 1951. "The Origin of the Idea of Natural Frontiers in France." *Annals of the Association of American Geographers* 41(2): 146–57.

Pounds, Norman J.G. 1954. "France and 'Les Limites Naturelles' from the Seventeenth to the Twentieth Centuries." *Annals of the Association of American Geographers* 44(1): 51–62.

Price, Glanville, ed. 1998. *Encyclopedia of the Languages of Europe*. Malden, MA: Blackwell.

Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Regimes and Material Well-Being in the World, 1950–1990*. New York: Cambridge University Press.

Pzreworski, Adam, and Fernando Limongi. 1997. "Modernization: Theories and Facts." *World Politics* 49(2): 155–83.

Reske, Chrisoph. 2007. *Die Buchdrucker des 16. und 17. Jahrhunderts im deutschen Sprachgebiet. Auf der Grundlage des gleichnamigen Werkes von Josef Benzing*. Wiesbaden: Harrassowitz.

Riker, William H. 1964. *Federalism: Origins, Operation, Significance*. New York: Little, Brown.

Riley, Shawn J., Stephen D. DeGloria, and Robert Elliot. 1999. "A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity." *Intermountain Journal of Sciences* 6(1–4): 23–7.

Roche, Alphonse V. 1954. *Provençal Regionalism: A Study of the Movement in the* Revue félibréenne*,* Le Feu *and other Reviews of Southern France*. Evanston: Northwestern University Press.

Rodden, Jonathan A. 2008. "Federalism". In *Oxford Handbook of Political Economy*, ed. Barry R. Weingast and Donald A. Wittman. New York: Oxford University Press, pp. 357–70.

Roeder, Philip G. 2007. *Where Nation-States Come From: Institutional Change in the Age of Nationalism*. Princeton: Princeton University Press.

Rokkan, Stein. 1999. "Conditions of State Formation and Nation-Building". In *State Formation, Nation-Building, and Mass Politics in Europe: The Theory of Stein Rokkan*, ed. Peter Flora with Stein Kuhnle and Derek Urwin. Oxford: Oxford University Press, pp. 149–69.

Root, Hilton L. 1987. *Peasants and King in Burgundy: Agrarian Foundations of French Absolutism*. Berkeley and Los Angeles: University of California Press.

Ross, Marc Howard. 2007. *Cultural Contestation in Ethnic Conflict*. New York: Cambridge University Press.

Roza, Joseph P. 2003. French Languages and French Nationalism: The Félibrige, Occitan, and the French Identity of Southern France, 1854–1914 PhD thesis Department of History, University of Washington.

Rubin, Jared. 2014. "Printing and Protestants: An Empirical Test of the Role of Printing in the Reformation." *Review of Economics and Statistics* 96(2): 270–86.

Rüegg, Walter, ed. 2011. *A History of the University in Europe*. Vol. 1–4. Cambridge: Cambridge University Press.

Rydell, Robert W., and Nancy E. Gwinn, eds. 1994. *Fair Representations: World's Fairs and the Modern World*. Amsterdam: VU University Press.

Safran, William, and Jean A. Laponce, eds. 2005. *Language, Ethnic Identity and the State*. New York: Routledge.

Sahlins, Peter. 1989. *Boundaries: The Making of France and Spain in the Pyrenees*. Berkeley and Los Angeles: University of California Press.

Sahlins, Peter. 2004. *Unnaturally French: Foreign Citizens in the Old Regime and After*. Ithaca: Cornell University Press.

Sasaki, Yu. Forthcoming. "Publishing Nations: Technology Acquisition and Language Standardization for European Ethnic Groups." *The Journal of Economic History* .

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

Sekhon, Jasjeet S. 2004. "Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals." *Perspectives on Politics* 2(2): 281–93.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Shelef, Nadav G. 2016. "Unequal Ground: Homelands and Conflict." *International Organization* 70(1): 33–63.

Singh, Prerna, and Matthias vom Hau. 2016. "Ethnicity in Time: Politics, History, and the Relationship between Ethnic Diversity and Public Good Provision." *Comparative Political Studies* 49(10): 1303–40.

Smith, Anthony D. 1986. *The Ethnic Origins of Nations*. Malden, MA: Blackwell.

Smith, Nathaniel B. 1973. "Historical Symbolism and Mythology in the Occitanist and Félibrige Movements." *Contemporary French Civilization* 3(1): 47–69.

Spolaore, Enrico, and Romain Wacziarg. 2009. "The Diffusion of Development." *Quarterly Journal of Economics* 124(2): 469–529.

Spolaore, Enrico, and Romain Wacziarg. 2013. "How Deep Are the Roots of Economic Development?" *Journal of Economic Literature* 51(2): 325–69.

Spruyt, Hendrik. 1994. *The Sovereign State and Its Competitors: An Analysis of Systems Change*. Princeton: Princeton University Press.

Spruyt, Hendrik. 2009. "War, Trade, and State Formation". In *Oxford Handbook of Comparative Politics*, ed. Carles Boix and Susan C. Stokes. New York: Oxford University Press, pp. 211–35.

Stasavage, David. 2010. "When Distance Mattered: Geographic Scale and the Development of European Representative Assemblies." *American Political Science Review* 104(4): 625–43.

Stasavage, David. 2014. "Was Weber Right? The Role of Urban Autonomy in Europe's Rise." *American Political Science Review* 108(2): 337–54.

Steinberg, Philip E. 2009. "Sovereignty, Territory, and the Mapping of Mobility: A View from the Outside." *Annals of the Association of American Geographers* 99(3): 467–95.

Steinberg, S.H. 1974. *Five Hundred Years of Printing*. 3rd ed. Middlesex: Penguin Books.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1): 1–21.

Swann, Julian. 2012. "Parlements and Provincial Estates". In *The Oxford Handbook of the Ancien Régime*, ed. William Doyle. New York: Oxford University Press, pp. 93–108.

Tabellini, Guido. 2008. "Institutions and Culture." *Journal of European Economic Association* 6(2–3): 255–94.

Tabellini, Guido. 2010. "Culture and Institutions: Economic Development in the Regions of Europe." *Journal of European Economic Association* 8(4): 677–716.

Talbert, Richard J.A., ed. 2000. *Barrington Atlas of the Greek and Roman World*. Princeton: Princeton University Press.

Thongchai, Winichakul. 1994. *Siam Mapped: A History of the Geo-Body of a Nation*. Honolulu: University of Hawai'i Press.

Tilly, Charles. 1975. "Reflection on the History of European State-Making". In *The Formation of National States in Western Europe*, ed. Charles Tilly. Princeton: Princeton University Press, pp. 3–83.

Tilly, Charles. 1985. "War Making and State Making as Organized Crime". In *Bringing the State Back In*, ed. Peter B. Evans, Dietrich Rueschemeyer and Theda Skocpol. New York: Cambridge University Press, pp. 169–91.

Tilly, Charles. 1992. *Coercion, Capital, and European States, AD 990–1992*. Malden, MA: Blackwell.

Tilly, Charles. 1994. "States and Nationalism in Europe, 1492–1992." *Theory and Society* 23(1): 131–46.

Tilly, Charles. 1994b. "Entanglements of European Cities and States". In *Cities and the Rise of States in Europe, A.D. 1000 to 1800*, ed. Charles Tilly and Wim P. Blockmans. New York: Cambridge University Press, pp. 1–27.

Tilly, Charles, and Wim P. Blockmans, eds. 1994. *Cities and the Rise of States in Europe, A.D. 1000 to 1800*. New York: Cambridge University Press.

van Zanden, Jan Luiten, Eltjo Buringh, and Maarten Bosker. 2012. "The Rise and Decline of European Parliaments, 1188–1789." *Economic History Review* 65(3): 835–61.

Walsby, Malcolm. 2011a. "The Vanishing Press: Printing in Provincial France in the Early Sixteenth Century". In *The Book Triumphant: Print in Transition in the Sixteenth and Seventeenth Centuries. Library of the Written Word*, ed. Malcolm Walsby and Graeme Kemp. Vol. 15: *The Handpress World*, vol. 9. Leiden and Boston: Brill, pp. 97–111.

Walsby, Malcolm. 2011b. *The Printed Book in Brittany, 1484–1600*. Leiden and Boston: Brill.

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

Weber, Max. 1978. *Economy and Society: An Outline of Interpretive Sociology*. Edited by Guenther Roth and Claus Wittich. Berkeley: University of California Press.

Weil, Patrick. 2008. *How to Be French: Nationality in the Making since 1789*. Translated by Catherine Porter. Durham: Duke University Press.

White, George W. 2000. *Nationalism and Territory: Constructing Group Identity in Southeastern Europe*. Lanham, MD: Rowman and Littlefield.

Wimmer, Andreas. 2002. *Nationalist Exclusion and Ethnic Conflict: Shadows of Modernity*. New York: Cambridge University Press.

Wimmer, Andreas. 2013. *Waves of War: Nationalism, State Formation, and Ethnic Exclusion in the Modern World*. New York: Cambridge University Press.

Wimmer, Andreas. 2015. "Nation-Building: A Long-Term Perspective and Global Analysis." *European Sociological Review* 31(1): 30–47.

Wimmer, Andreas, and Yuval Feinstein. 2010. "The Rise of the Nation-State across the World, 1816 to 2001." *American Sociological Review* 75(5): 764–90.

Wong, Edward. 2008. "The Dead Tell a Tale China Doesn't Care to Listen To." *New York Times,* November 18.
URL: `http://www.nytimes.com/2008/11/19/world/asia/19mummy.html`.

Woodberry, Robert D. 2012. "The Missionary Roots of Liberal Democracy." *American Political Science Review* 106(2): 244–74.

Woodward, David. 2007. "Cartography and the Renaissance: Continuity and Change". In *The History of Cartography*, ed. David Woodward. Vol. 3: *Cartography in the European Renaissance*. Part I. Chicago and London: University of Chicago Press, pp. 3–24.

Wright, Julian. 2003. *The Regionalist Movement in France, 1890–1914: Jean Charles-Brun and French Political Thought*. Oxford: Clarendon.

Yanagizawa-Drott, David. 2014. "Propaganda and Conflict: Evidence from the Rwandan Genocide." *Quarterly Journal of Economics* 129(4): 1947–94.

Yiftachal, Oren. 2001. "The Homeland and Nationalism". In *Encyclopedia of Nationalism*, ed. Alexander J. Motyl. Vol. 1: *Fundamental Themes*. San Diego: Academic Press, pp. 359–83.

Zerubavel, Yael. 1995. *Recovered Roots: Tradition, Collective Memory and the Making of the Israeli National Tradition*. Chicago and London: University of Chicago Press.

Ziblatt, Daniel. 2004. "Rethinking the Origins of Federalism: Puzzle, Theory, and Evidence from Nineteenth-Century Europe." *World Politics* 57(1): 70–98.

Ziblatt, Daniel. 2006. *Structuring the State: The Formation of Italy and Germany and the Puzzle of Federalism*. Princeton: Princeton University Press.

Appendix A

APPENDIX FOR CHAPTER 2: "ETHNIC AUTONOMY"

A.1 Descriptive statistics

Table A.1: Summary statistics.

	N	type	mean	median	sd	min	max
dependent variables*							
Proportion of non-French-speaking communes	214	pct	0.03	0	0.15	0	1
Proportion of non-French-speaking populations	214	pct	0.02	0	0.13	0	1
geography variables							
Distance from Paris (km)	214	con	372.7	349.6	197.55	0	985.7
Elevation (m)	214	con	180.1	131.5	194.2	2	1,304
Terrain ruggedness (m)	214	con	64.54	39.88	80.23	1.41	556.81
institutions variable*							
Years of provincial estates held	214	con	68	0	103.8	0	398
economics variables							
Urban potential	214	con	107.5	107.5	61.79	1	214
controls							
Ethnicity	214	bin	0.57	1	0.49	0	1
Cumulative years under French rule since 1400	214	con	453	393.2	137.72	60	500
Weighted language distance	214	pct	0.329	0.246	0.216	0	0.607
Agricultural onset (1,000 years)	214	con	5.9	5.9	0.177	5.44	6.36
Distance from Wittenberg (km)	214	con	938.9	943.9	223.7	483.2	1,400.8
Year of printing press adoption	60	con	1526	1496	56	1460	1700
Number of printing press*	214	con	0.53	0	0.62	0	3
University	214	bin	0.1	0	0.31	0	1
Number of university*	214	con	0.27	0	0.47	0	2
Bishopric	214	bin	0.43	0	0.5	0	19
Diocese	214	bin	0.32	0	0.47	0	1
Number of diocese*	214	con	0.92	0	0.87	0	4
Archdiocese	214	bin	0.11	0	0.31	0	1
Number of archdiocese*	214	bin	0.35	0	0.61	0	3
Hub Roman roads	214	bin	0.29	0	0.45	0	1
Roman roads	214	bin	0.7	1	0.46	0	1
Years under Roman rule	214	con	384	503	205	0	503
Access to oceanic coast under Rome	180	bin	0.28	0	0.45	0	1
Access to navigable river under Rome	180	bin	0.83	1	0.37	0	1

Note: * placed at the end of variable description denotes that data is observed at the provincial level. I treat missing values as zeros in analysis. "Type" indicates data types, where "con" denotes continuous variables, "pct" percentage, "fac" factor (or ordinal) variables, and "bin" binary (or dummy) variables.

A.2 Regression tables

In this section, I report all regression outputs used in Chapter 2. The first subsection documents the OLS results and the second, the instrumental-variable regression results.

A.2.1 OLS results

Table A.2 is the basis for the quantities of interest presented in Figure 2.4. Panels (a) through (d) of Figure 2.4 correspond to Models 1 through 4, respectively. In the paper, I focus on the magnitude of the *Ethnicity*, *Log distance to Paris*, *Ethnicity* $\times$ *Log distance to Paris*, and *Linguistic distance* variables.

Figures 2.5 and 2.6 of the paper are based on the simulation analysis of Model 1 (panel (a) for each figure) and Model 3 (panel (b) for each figure), respectively. Figures 2.7 and 2.8 are the simulation results of Model 2 (panel (a) for each figure) and Model 4 (panel (b) for each figure), respectively. (The results of the "no interaction" scenarios presented in Figure 2.8 is the same as Figure 2.5—shown again to highlight the effect of the interaction model.)

Table A.2: OLS regression outputs on ethnic groups' ability to retain autonomy, saturated models.

Dependent variable	Proportion non-French-speaking communes (1)	Proportion non-French-speaking communes (2)	Proportion non-French-speaking populations (3)	Proportion non-French-speaking populations (4)
Ethnicity	0.858**	−2.333**	0.681**	−2.210***
	(0.371)	(0.899)	(0.329)	(0.826)
Log distance to Paris	0.113*	0.040	0.104*	0.038
	(0.064)	(0.032)	(0.060)	(0.031)
Ethnicity × Log distance to Paris		0.490***		0.444***
		(0.116)		(0.110)
Years under French rule	−0.0005*	−0.0001	−0.0005*	−0.0002
	(0.0003)	(0.0002)	(0.0002)	(0.0002)
Elevation	−0.0001	−0.0002	−0.0001	−0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Terrain ruggedness	−0.0005	−0.0004	−0.0004	−0.0003
	(0.0004)	(0.0004)	(0.0003)	(0.0004)
Years of provincial estates held	−0.0002	−0.0002	−0.0003	−0.0002
	(0.0005)	(0.0004)	(0.0004)	(0.0004)
Log urban potential	0.012	0.025	0.010	0.022
	(0.027)	(0.024)	(0.022)	(0.019)
Linguistic distance	1.392*	0.600	1.045	0.327
	(0.763)	(0.737)	(0.685)	(0.652)
Log distance to Wittenberg	0.151	0.009	0.163	0.034
	(0.208)	(0.173)	(0.192)	(0.161)
Number of printing press	−0.001	−0.014	−0.005	−0.018
	(0.032)	(0.028)	(0.030)	(0.027)
Number of university	0.043	0.048	0.029	0.033
	(0.054)	(0.050)	(0.048)	(0.044)
Number of bishoprics	−0.009	−0.003	−0.013	−0.008
	(0.038)	(0.038)	(0.036)	(0.037)
Roman hub	−0.022	−0.029	−0.001	−0.007
	(0.048)	(0.042)	(0.041)	(0.035)
Roman road	−0.090	−0.095	−0.093	−0.098
	(0.085)	(0.078)	(0.075)	(0.069)
Years under Roman rule	0.0002	0.0002	0.0002	0.0001
	(0.0002)	(0.0001)	(0.0002)	(0.0001)
Access to river	0.005	−0.003	−0.025	−0.032
	(0.084)	(0.082)	(0.077)	(0.074)
Access to ocean	−0.035	−0.036	−0.032	−0.033
	(0.107)	(0.097)	(0.091)	(0.083)
Intercept	−2.146*	−0.533	−1.950*	−0.489
	(1.150)	(1.071)	(1.065)	(0.999)
Adjusted R^2	0.34	0.45	0.33	0.44
Observations	214	214	214	214

Note: Robust standard errors clustered by province for all models. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table A.3 documents the OLS outputs of the non-saturated specifications as a robustness check. Here I focus on the effect of the three competing models—geography, institutions, and urbanization—along with all controls used in Table A.2. This shows that the *Log distance to Paris* variable is positively and significantly associated with the two dependent variables (Models 1 and 2), while the other geography variables are negatively and not significantly correlated. *Ethnicity* is also positive and significant across all models. The result is consistent with that of Table A.2.

Table A.3: OLS regression outputs on ethnic groups' ability to retain autonomy, three competing models (geography, institutions, urbanization).

model	geography		institutions		urbanization	
Dependent variable	Proportion non-French-speaking communes	Proportion non-French-speaking communes	Proportion non-French-speaking populations	Proportion non-French-speaking populations	Proportion non-French-speaking populations	Proportion non-French-speaking populations
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnicity	0.811**	0.622**	0.943***	0.760***	0.917***	0.721**
	(0.350)	(0.313)	(0.328)	(0.292)	(0.317)	(0.285)
Log distance to Paris	0.109*	0.099*				
	(0.065)	(0.060)				
Elevation	−0.0001	−0.0001				
	(0.0002)	(0.0002)				
Terrain ruggedness	−0.0005	−0.0004				
	(0.0004)	(0.0003)				
Years of provincial estates held			−0.0001	−0.0002		
			(0.0005)	(0.0004)		
Log urban potential					0.004	0.002
					(0.028)	(0.024)
Years under French rule	−0.0005*	−0.0005*	−0.001**	−0.001**	−0.001**	−0.001**
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Linguistic distance	1.366*	1.012	1.595**	1.232*	1.583**	1.214*
	(0.757)	(0.683)	(0.710)	(0.643)	(0.713)	(0.647)
Log distance to Wittenberg	0.144	0.157	0.278	0.280	0.276	0.275
	(0.206)	(0.190)	(0.202)	(0.191)	(0.200)	(0.187)
Number of printing press	0.002	−0.003	−0.009	−0.013	−0.009	−0.013
	(0.031)	(0.029)	(0.031)	(0.030)	(0.031)	(0.029)
Number of university	0.043	0.030	0.055	0.040	0.056	0.041
	(0.055)	(0.049)	(0.057)	(0.050)	(0.056)	(0.049)
Number of bishoprics	−0.008	−0.013	0.005	−0.001	0.004	−0.002
	(0.039)	(0.037)	(0.039)	(0.037)	(0.039)	(0.037)
Roman hub	−0.025	−0.005	−0.022	−0.001	−0.024	−0.005
	(0.047)	(0.040)	(0.050)	(0.043)	(0.049)	(0.042)
Roman road	−0.086	−0.088	−0.118	−0.118	−0.115	−0.113
	(0.081)	(0.072)	(0.085)	(0.076)	(0.080)	(0.072)
Years under Roman rule	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Access to river	0.003	−0.028	0.001	−0.029	−0.001	−0.032
	(0.084)	(0.078)	(0.082)	(0.076)	(0.083)	(0.077)
Access to ocean	−0.037	−0.035	0.015	0.014	0.011	0.006
	(0.101)	(0.086)	(0.114)	(0.098)	(0.110)	(0.093)
Intercept	−2.005*	−1.815*	−2.412**	−2.204**	−2.404**	−2.161**
	(1.104)	(1.021)	(1.186)	(1.117)	(1.160)	(1.081)
Adjusted R^2	0.35	0.33	0.32	0.30	0.32	0.30
Observations	214	214	214	214	214	214

Note: Robust standard errors clustered by province for all models. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

A.2.2 Endogeneity concerns

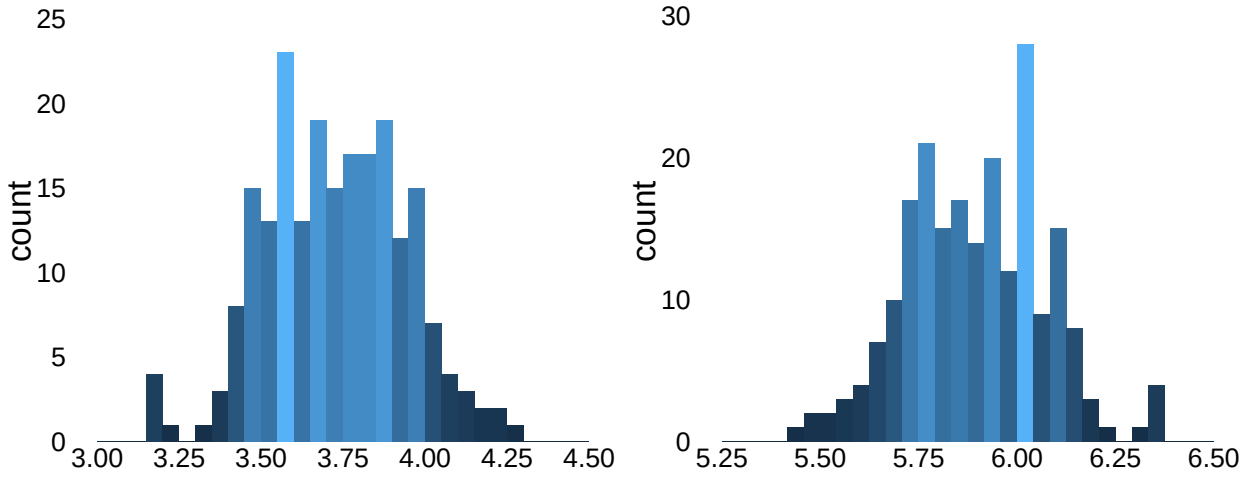
The OLS results reported in Chapter 2 and in Appendix A above should allay concerns that geography scale, in particular the distance-to-Paris measure, is positively and significantly associated with ethnic autonomy. They are robust to the inclusion of a host of covariates that may be connected to the outcome variable. Yet one might challenge the claim that distance and ethnic self-rule are *causally* connected. More specifically, there is a concern that ethnic autonomy might be a product not of the institutional development different from that of the ruling state, but of cultural differences between the ruling group and the ruled. Figure 2.4 points to this challenge in that linguistic distance between ethnic groups is positively and significantly associated with the proportion of non-French-language users. If this variable captures the hypothesis that culture jointly determines ethnic groups' capacity to resist the ruling state and distance from Paris, the impact of the distance parameter is confounded. To address this endogeneity concern, I turn to the instrumental variables (IV) approach. Specifically, I rely on the exogenous variation of the timing of the Neolithic Revolution (or agriculture) as a proxy for the start of institutional development.

The timing of the Neolithic Revolution is an ideal instrument for this paper's empirical analysis. An adequate instrument should be correlated with the distinct institutional development at the local level (distance to Paris) but not with ethnic autonomy itself. At the same time, it should affect the outcome variable only through the proposed causal mechanism. The onset-of-agriculture variable meets these criteria. The Neolithic Revolution, in general, denotes the transition from nomadic to sedentary life approximately 10,000 years ago, when hunter-gatherers began crop cultivation and animal domestication as the primary means of subsistence. In the social sciences, it represents the onset of institution-building such as property rights in a path to independent statehood, because efficient, sustained food production under the condition of growing population density demanded rule-making and enforcement (Chanda and Putterman 2007; Hariri 2012; Menaldo 2012; North and Thomas 1977; Olsson and Hibbs 2005). Although the state has been the default unit of analysis in the literature, the analytical framework is applicable to my argument of the distance-to-Paris

variable as a proxy for distinct institutional development for ethnic groups.

I use the geographical distance to the Fertile Crescent as an instrument and construct a measure of the time elapsed since agriculture onset. The "Fertile Crescent" encompasses an expansive region that commonly refers to today's Iraq, Israel, Jordan, Lebanon, and Syria (Olsson and Hibbs 2005, 913). I take Baghdad as one of the centers of this area and a proxy for the origin of the spread of agriculture. As extant research has shown that agriculture spread from the Middle East through Europe with an estimated pace of 1.2km a year (Bellwood 2001, 186), it is possible to calculate the timing of agriculture onset by taking the distance between Baghdad and a given French city *i* in the data set divided by 1.2. I then calculate the time elapsed since the onset. Figure A.1 plots the distribution of distance from Baghdad and of agriculture onset. Both histograms appear normally distributed and would be adequate for least-squares estimation.

Figure A.1: Distribution of distance to Baghdad and of the time elapsed since the Neolithic Revolution.



(a) Distance to Baghdad (1,000km) **(b) Time elapsed since agriculture onset (1,000 years)**

To examine the causal impact of the time since the Neolithic Revolution as an IV, I adopt a two-stage least-squares (2SLS) model. In the first stage, I estimate distance to Paris as the outcome

variable, using the instrument as the main predictor along with the same set of covariates in the previous models. In the second stage, I use the predicted values of the first-stage regression to estimate the proportion of non-French-language speaking communes and populations with the same set of controls. Figure A.2 shows the 2SLS regression outputs.

Figure A.2: 2SLS (two-stage least-squares) model with the onset of agriculture as IV.

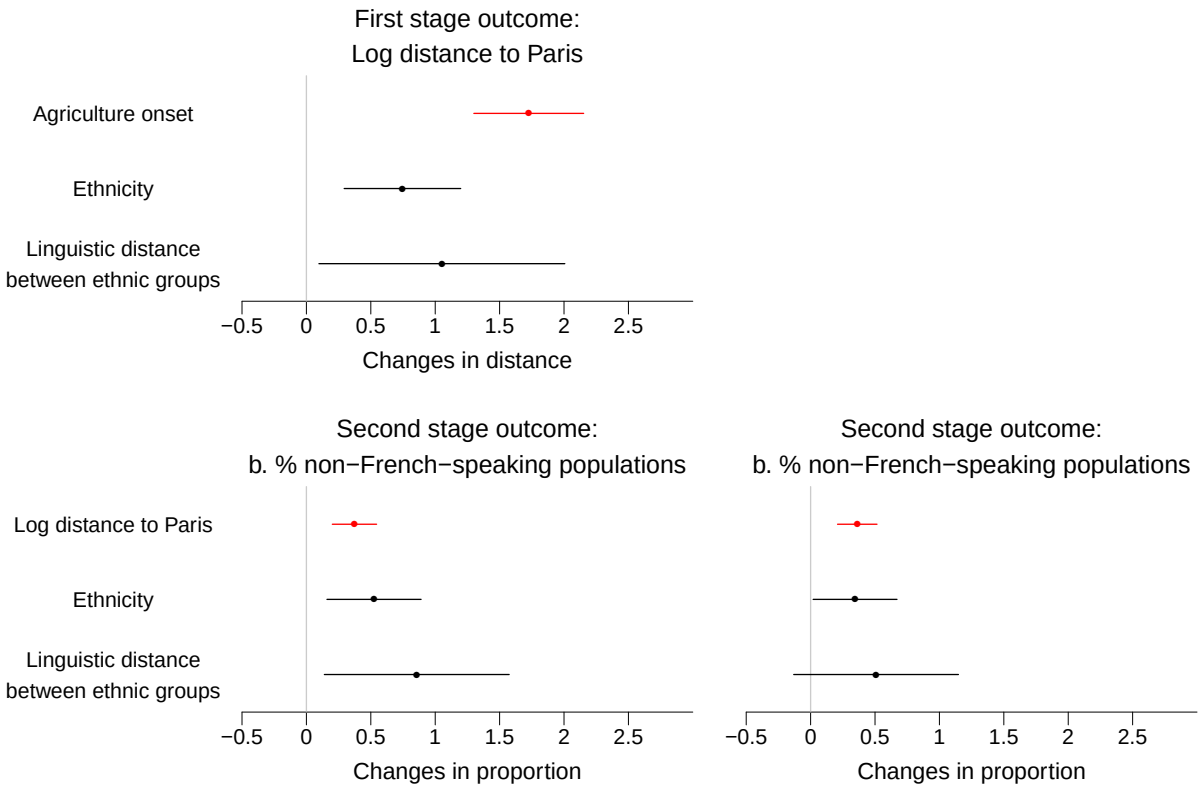


Figure A.2 indicates that the time elapsed since the Neolithic Revolution is positively correlated with the distance-to-Paris variable whose effects remain positive in the two outcome variables on ethnic autonomy. In the first stage, when agriculture starts early by 1,000 years, the institutional development as captured by the distance parameter increases by 5.75. The F-statistic test for the weak instrument is 63. In the second stage, the effect of the early development of institutions remains

consistent, where the proportion of use in non-French tongues increases by approximately 0.005. This value is 3 to 3.5 times *greater* than the OLS model presented in Figure 4. These results suggest that agriculture onset is a valid instrument and supports the hypothesis that ethnic groups' ability to resist the ruling state is based on institutional development through the Neolithic Revolution.

Finally, to check the robustness of the IV approach, I estimate an alternative model, where the linguistic distance between groups as a proxy for culture as the first-stage outcome variable. There might be a concern that while agriculture triggered institutional development, it could affect the degree of language sophistication. This is because crop and animal husbandry may demand efficient communication, which in turn promotes language development. Figure A.3 reports the 2SLS model output.

Figure A.3: 2SLS (two-stage least-squares) model with the onset of agriculture as IV.

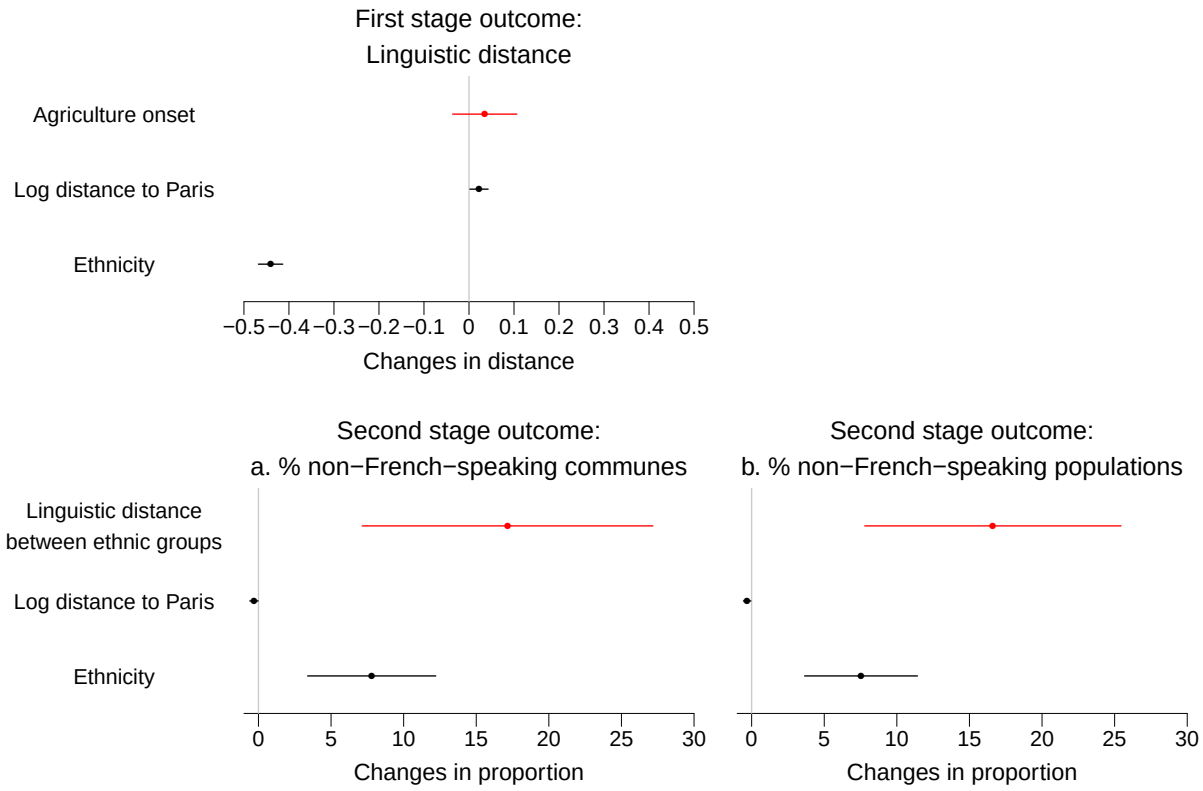


Figure A.3 suggests that the time since agriculture onset does not seem to affect language development. Although linguistic distance is positively and significantly associated with ethnic autonomy, agriculture does not have a significant impact on language as shown in the first stage. This result suggests that the onset of the Neolithic Revolution has an effect on ethnic groups' ability to resist the ruling state primarily through institutional development as captured in geographical scales.

The tables below report the full results of the 2SLS regressions, where Figures A.2 and A.3 correspond to Tables A.4 and A.5, respectively.

Table A.4 documents that the onset of the Neolithic Revolution is positively and significantly linked to the distance-to-Paris measure in the first stage. The second stage uses the predicted val-

ues of the distance variable as the main explanatory variable for reported in the second stage with the two outcome variables on ethnic autonomy (columns 2 and 3). This geography scale remains positive and significant, while ethnicity exhibits the similar pattern. The *F*-statistic test on weak instruments is above 63, where 10 is generally considered the cutoff for "weakness."

Table A.5 reports the same specifications except for the first stage, where the dependent variable is linguistic distance, a proxy for culture. The onset of the Neolithic Revolution is positive but not significant. The *F*-statistic test is less than 1, suggesting that agriculture onset is a weak predictor for the development of distinct culture for minority ethnic groups in France. Nevertheless, I used the predicted values of this IV as the main explanatory variable in the second stage. Although columns 2 and 3 indicate that cultural differences seem highly positively correlated with ethnic autonomy, it is statistically difficult to tell if this proxy for culture is exogenous to observed or unobserved variables.

Table A.4: Two-stage least-square (2SLS) regression outputs on ethnic groups' ability to retain autonomy, 1400–1900, saturated models, Onset of Neolithic Revolution as IV.

	First stage	Second stage	Second stage
dependent variable	Log distance to Paris	Prop. non-French-speaking communes	Prop. non-French-speaking populations
	(1)	(2)	(3)
Onset of Neolithic Revolution	1.73***		
	(0.22)		
Log distance to Paris		0.373***	0.362***
		(0.088)	(0.078)
Ethnicity	0.755**	0.525***	0.345*
	(0.23)	(0.186)	(0.166)
Elevation	0.0003	−0.0003	−0.0003†
	(0.0003)	(0.0002)	(0.0002)
Terrain ruggedness	−0.0007	−0.0002	−0.0001
	(0.0005)	(0.0004)	(0.0003)
Years of provincial estates held	0.0004	−0.0003	−0.0004*
	(0.0003)	(0.0002)	(0.0002)
Log population	−0.014	−0.004	−0.009
	(0.034)	(0.017)	(0.015)
Log urban potential	0.005	0.014	0.011
	(0.034)	(0.023)	(0.02)
Urbanization rate	−122.5***	22.5	24.59†
	(14.31)	(14.99)	(13.36)
Weighted average economic growth	0.06†	0.017	0.024
	(0.033)	(0.024)	(0.2)
Linguistic distance	1.05*	0.86*	0.51
	(0.49)	(0.37)	(0.33)
Log distance to Wittenberg	1.12***	−0.057	−0.037
	(0.19)	(0.13)	(0.11)
Number of printing press	−0.005	0.001	−0.004
	(0.054)	(0.038)	(0.034)
Number of university	0.17*	0.029	0.014
	(0.07)	(0.05)	(0.04)
Number of bishoprics	−0.045	−0.011	−0.016
	(0.057)	(0.04)	(0.04)
Roman hub	0.02	0.005	0.03
	(0.082)	(0.057)	(0.05)
Roman road	−0.18†	−0.05	−0.05
	(0.1)	(0.07)	(0.07)
Years under Roman rule	0.0001	0.0001	0.0001
	(0.0003)	(0.0001)	(0.0002)
Access to river	−0.052	0.02	−0.002
	(0.091)	(0.064)	(0.057)
Access to ocean	0.38***	−0.12	−0.12†
	(0.093)	(0.07)	(0.07)
Intercept	−12.79***	−2.07**	−1.93**
	(2.15)	(0.778)	(0.693)
Observations	214	214	214
Adjusted R^2	0.64	0.37	0.36
F-statistic on weak instrument	63.02		

*** denote $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and † $p < 0.1$.

Table A.5: Two-stage least-square (2SLS) regression outputs on ethnic groups' ability to retain autonomy, 1400–1900, saturated models, Onset of Neolithic Revolution as IV.

	First stage	Second stage	Second stage
dependent variable	Linguistic distance	Prop. non-French-speaking communes	Prop. non-French-speaking populations
	(1)	(2)	(3)
Onset of Neolithic Revolution	0.035		
	(0.036)		
Linguistic distance		17.16***	16.61***
		(5.11)	(4.5)
Log distance to Paris	0.022*	−0.31*	−0.31*
	(0.01)	(0.15)	(0.13)
Ethnicity	−0.44**	7.8***	7.53***
	(0.014)	(2.25)	(1.98)
Elevation	0.0001**	−0.002**	−0.002***
	(0.0003)	(0.0006)	(0.0005)
Terrain ruggedness	−0.0002**	0.003**	0.003**
	(0.00008)	(0.001)	(0.001)
Years of provincial estates held	0.00009*	−0.002***	−0.002***
	(0.00004)	(0.0005)	(0.0004)
Log population	−0.007*	−0.11*	0.1**
	(0.004)	(0.041)	(0.04)
Log urban potential	0.002	−0.019	−0.02
	(0.005)	(0.025)	(0.02)
Urbanization rate	7.2**	−132**	−128.1***
	(2.39)	(41.25)	(36.35)
Weighted average economic growth	−0.011*	0.22***	0.22***
	(0.005)	(0.06)	(0.054)
Log distance to Wittenberg	0.167***	−2.35**	−2.31***
	(0.03)	(0.78)	(0.69)
Number of printing press	−0.001	0.02	0.02
	(0.008)	(0.04)	(0.035)
Number of university	−0.014	0.31**	0.29***
	(0.01)	(0.09)	(0.08)
Number of bishoprics	0.001	−0.045	−0.048
	(0.008)	(0.04)	(0.04)
Roman hub	0.009	0.005	−0.11†
	(0.012)	(0.057)	(0.06)
Roman road	−0.002	−0.14†	−0.08
	(0.02)	(0.08)	(0.07)
Years under Roman rule	0.00003	−0.0003	−0.0003
	(0.00004)	(0.0002)	(0.0002)
Access to river	0.019	−0.31**	−0.33**
	(0.013)	(0.12)	(0.1)
Access to ocean	−0.038**	0.62**	−0.61**
	(0.014)	(0.22)	(0.19)
Intercept	−0.91**	8.33*	8.34**
	(0.33)	(3.29)	(2.9)
Observations	214	214	214
Adjusted R^2	0.9	0.37	0.34
F-statistic on weak instrument	0.93		

*** denote $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and † $p < 0.1$.

Appendix B

APPENDIX FOR CHAPTER 3: "PUBLISHING NATIONS"

B.1 List of 171 ethnic groups in the data set

Table B.1: List of 171 ethnic groups in the data set

	name	homeland city	printing press	dictionary
1	Abaza	Adyge-Khabl		
2	Abkhaz	Sukhumi		1986
3	Adyge	Maykop		
4	Aguls	Tpig, Dagestan		
5	Ajars	Batumi		
6	Akhvakh	Akhvakhsky district, Dagestan		
7	Alanders	Mariehamm		
8	Albanians	Tirana	1555	1980
9	Alsatians	Strasbourg	1460	1899
10	Andalusians	Seville	1477	1941
11	Andis	Khasavyurt, Dagestan		
12	Andorrans	Andorra La Vella		
13	Archis	Arsha-Makhi, Dagestan		1977
14	Armenians	Yerevan	1512	1971
15	Aromanians	Trikala		1963
16	Austrians	Vienna	1461	1963
17	Avars	Kunzakh, Khunzakhsky district, Dagestan		
18	Azeris	Baku	1870	1966
19	Azoreans	Ponta Delgada		
20	Bagulals	Tsumadinsky district		
21	Balkars	Karachayevsky district		1996
22	Bashkorts	Ufa		1993
23	Basques	Vitoria	1545	1989
24	Bats	Tusheti		2003
25	Bavarians	Munich	1482	1827
26	Belarussians	Minsk	1520	1977
27	Bezhetas	Bezhta village, Tsuntinsky district, Dagestan		
28	Bosnians	Sarajevo	1866	1992
29	Botligs	Botlikh village, Botlikhsky district, Dagestan		
30	Bretons	Rennes	1485	1958
31	Budugs	Budug		
32	Bulgarians	Sofia	1566	1895
33	Burgundians	Dijon	1491	
34	Canarians	Santa Cruz de Tenerife		1995
35	Carpatho-Rusyns	Uzhgorod	1572	

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
36	Catalans	Barcelona	1475	1930
37	Chamalals	Agvali village, Tsumadinsky district, Dagestan		
38	Chavash	Cheboksary		1928
39	Chechens	Grozny		1961
40	Cherkess	Cherkessk		1956
41	Cornish	Truno	1700	
42	Corsicans	Ajaccio	1769	
43	Crimean Tatars	Bakhchysarai		
44	Croats	Zagreb	1483	1901
45	Csangos	Miercurea Ciuc	1533	1936
46	Czechs	Prague	1487	1935
47	Danes	Copenhagen	1490	1918
48	Dargins	Makhachkala, Dagestan		2005
49	Didos	Kidero village, Tsuntinsky district, Dagestan		
50	Don Cossacks	Novocherkassk		1975
51	Dutch	Amsterdam	1477	1864
52	English	London	1476	1755
53	Estonians	Tallinn	1535	1918
54	Faroese	Torshavn		1966
55	Finns	Helsinki	1542	1953
56	Flemish	Brussels	1475	1873
57	French	Paris	1470	1858
58	Frisians	Leeuwarden	1630	1874
59	Friulis	Udine	1484	1985
60	Gagauz	Comrat		
61	Galicians	Santiago de Compostela	1483	1986
62	Georgians	Tbilisi	1709	1884
63	Germans	Berlin	1540	1852
64	Gibraltarians	Gibraltar	1878	
65	Ginugs	Tsuntinsky district, Dagestan		
66	Godoberis	Godoberi village, Botlikhsky district, Dagestan		
67	Greeks	Athens	1759	1936
68	Guernseians	St. Peter Port		
69	Hungarians	Budapest	1473	1890
70	Icelanders	Reykjavik	1540	1912
71	Ingrians	Saint Petersburg	1711	1953
72	Ingush	Nazran		
73	Irish	Dublin	1550	1913
74	Istrians	Pulj	1905	1979
75	Italians	Rome	1467	1861

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
76	Jerseyites	St. Helier		
77	Kabards	Nalchik		
78	Kalmyks	Elista		
79	Karachais	Mikoyan Shakhar		1996
80	Karatas	Karata village, Akhvakhsky district, Dagestan		
81	Karels	Petrozavodsk	1930	1968
82	Kashubians	Gdansk	1499	1994
83	Khinalugs	Quba		
84	Khunzals	Nakhada village, Tsuntinsky district, Dagestan		
85	Khwarshia	Khvarshi village, Tsumadinsky district, Dagestan		
86	Kists	Kakheti		
87	Komis	Syktyvkar		1959
88	Kosovars	Pristina		
89	Kryts	Jek		
90	Kuban Cossacks	Krasnodar		
91	Kumyks	Buynaksk, Dagestan		
92	Ladins	Belluno	1475	
93	Laks	Vachi, Dagestan		
94	Latvians	Riga	1525	1972
95	Lezgins	Mamash, Dagestan		
96	Liechtensteiners	Vaduz		
97	Ligurians	Genoa	1472	1985
98	Lithuanians	Vilnius	1525	1956
99	Livonians	Ventspils	1542	1938
100	Lombards	Milan	1471	1835
101	Luxembourgers	Luxembourg City	1598	1950
102	Macedonians	Skopje	1840	1961
103	Madeirans	Funchal		
104	Maltese	Valletta	1642	1987
105	Manx	Douglas	1707	1835
106	Maris	Yoshkar-Ola	1767	1990
107	Megleno-Romanians	Kilkis	1512	1963
108	Meskhtekians	Akhaltsikhe		
109	Moldovans	Chisinau	1642	1977
110	Monegasques	Monaco-Ville	1795	

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
111	Montenegrins	Podgorica	1494	
112	Moravians	Brno	1486	1982
113	Mordvins	Saransk		1973
114	Mountain Jews	Majalis, Dagestan		
115	Nenets	Naryan-Mar		
116	Nogais	Kizlyar, Dagestan		
117	Normans	Rouen	1485	
118	Northumbrians	Newcastle	1646	
119	Norwegians	Oslo	1643	1930
120	Occitans	Toulouse	1476	1875
121	Ossetians	Vladikavkaz		1958
122	Piedmontese	Turin	1474	1983
123	Poles	Warsaw	1522	1807
124	Pomaks	Edirne	1554	
125	Portuguese	Lisbon	1489	1852
126	Rhinelanders	Cologne	1464	1928
127	Romands	Geneva	1478	1926
128	Romanians	Bucharest	1708	1907
129	Romansh	Chur	1557	1939
130	Romas	Soroca		
131	Russians	Moscow	1564	1891
132	Rutuls	Rutul, Dagestan		
133	Samis	Kautokeino	1649	1939
134	Sanjakis	Novi Pazar		
135	San Marinese	San Marino		
136	Sards	Cagliari	1493	
137	Savoyards	Chambéry	1484	1969
138	Saxons	Dresden	1524	1953
139	Scanians	Malmo	1528	
140	Scots	Edinburgh	1508	1931
141	Serbs	Belgrade	1552	1818
142	Sicilians	Palermo	1478	1977
143	Silesians	Wroclaw	1475	1962
144	Slovaks	Bratislava	1500	1959
145	Slovenes	Ljubljana	1575	1976
146	Sorbs	Bautzen	1555	1978
147	Spaniards	Madrid	1566	1886
148	Swabians	Stuttgart	1486	1901
149	Swedes	Stockholm	1483	1893
150	Swiss-Germans	Zürich	1479	1881

continued on next page

continued from previous page

	name	homeland city	printing press	dictionary
151	Szeklars	Cluj-Napoca	1578	1975
152	Tabasarans	Khiv, Dagestan		
153	Talysh	Astara		
154	Tats	Shabran		
155	Tatars	Kazan	1804	1969
156	Terek Cossacks	Stavropol		
157	Tindis	Tindi village, Tsumadinsky district, Dagestan		
158	Tsakhurs	Balakan, Dagestan		
159	Turks	Ankara	1726	1963
160	Tyroleans	Innsbruck	1547	1866
161	Udis	Qabala district		
162	Udmurts	Izhhevsk		1890
163	Ukrainians	Kiev	1616	1886
164	Venetians	Venice	1469	1829
165	Veps	Sheltozero district		1972
166	Volga Germans	Saratov		1974
167	Vorarlbergers	Bregenz	1658	1955
168	Waldensians	Torre Pellice	1479	1973
169	Walloons	Namur		1927
170	Welsh	Cardiff	1546	1887
171	Zetlanders	Lerwick		

B.2 List of 47 European states in the data set

This table provides a list of 47 European states in my data set, which describes the earliest date of the printing press, the name of the city where the first printing press was set up (whenever the name is available), the language with the largest population, and the date of the vernacular dictionary. The *Language* column denotes the official languages; for countries with multiple official languages, the language with the largest population is listed on the first row and those with fewer populations are on the second row.

The data is sorted by the date of print acquisition.

Table B.2: List of 47 European states in the data set

	country	printing press	city	dictionary	language
1	Germany	1450	Mainz	1852	German (standard)
2	Austria	1461	Vienna	1963	German (regional)
3	Italy	1464	Subiaco	1861	Italian
4	Netherlands	1465	Haarlem	1864	Dutch
5	France	1470	Paris	1858	French
6	Switzerland	1470	Basel, Beromünster	1881	Swiss-German, French, Italian, Romansh
7	Belgium	1473	Bruges	1927	French (regional), Dutch
8	Hungary	1473	Buda (Budapest)	1890	Hungarian
9	Spain	1473	Valencia, Barcelona	1886	Spanish
10	Poland	1474	Kraków (Cracow)	1807	Polish
11	Czech Republic	1476	Plzeň (Pilsen)	1935	Czech
12	United Kingdom	1476	London	1755	English
13	Denmark	1482	Odense	1918	Danish
14	Croatia	1483	Zagreb	1901	Croatian
15	Sweden	1483	Stockholm	1893	Swedish
16	Portugal	1487	Faro	1852	Portuguese
17	Turkey	1493	Istanbul	1963	Turkish
18	Montenegro	1494	Podgorica		Montenegrin
19	Slovakia	1500	Bratislava	1959	Slovak
20	Romania	1508	Târgovişte	1907	Romanian (standard)
21	Armenia	1512	Yerevan	1971	Armenian
22	Greece	1512	Salonica (Thessaloniki)	1936	Greek
23	Belarus	1520		1977	Belarusian, Russian
24	Lithuania	1525	Vilnius	1956	Lithuanian
25	Iceland	1530	Hollar	1912	Icelandic

continued on next page

continued from previous page

	name	printing press	city	dictionary	language
26	Bosnia and Herzegovina	1531	Goražde	1992	Bosnian
27	Ireland	1550	Dublin	1913	Irish, English
28	Serbia	1552	Belgrade	1818	Serbian
29	Albania	1555	Tirana	1980	Albanian
30	Russia	1564	Moscow	1891	Russian
31	Ukraine	1574	Lviv (Lvov)	1886	Ukrainian
32	Slovenia	1575	Ljubljana	1976	Slovene
33	Latvia	1587	Riga	1972	Latvian
34	Luxembourg	1598	Luxembourg	1950	Luxemburgish, French, German
35	Estonia	1630	Tallinn	1918	Estonian
36	Finland	1642	Turku	1953	Finnish, Swedish
37	Malta	1642		1987	Maltese
38	Moldova	1642	Iaşi	1977	Romanian (regional)
39	Norway	1643	Olso	1930	Norwegian
40	Georgia	1709	Tbilisi	1884	Georgian
41	Monaco	1795			Monégasque
42	Bulgaria	1828	Samokov	1895	Bulgarian
43	Macedonia	1840		1961	Macedonian
44	Azerbaijan	1870	Baku	1966	Azeri
45	Andorra				
46	San Marino				
47	Liechtenstein				

B.3 Regression tables

B.3.1 Cox proportional hazards and logistic regression models used in paper

Table B.3 reports the complete regression estimates of Table 3.3.

Table B.3: Regression outputs of the printing press' impact on language standardization.

	bivariate		demand side		supply side		fully specified	
model	Cox	logit	Cox	logit	Cox	logit	Cox	logit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Printing press	2.159***	2.933***	1.600***	2.178***	1.455***	1.932***	1.450***	2.083***
	(0.269)	(0.342)	(0.356)	(0.456)	(0.354)	(0.503)	(0.339)	(0.556)
University			0.0004***	0.001***	0.0004***	0.001***	0.0004***	0.001***
			(0.0001)	(0.0002)	(0.0001)	(0.0003)	(0.0001)	(0.0002)
Bishop			−0.000002	−0.0001	0.00003	0.00004	−0.00002	−0.00002
			(0.0001)	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.0003)
Log urban potential			0.114	0.098	−0.002	−0.307	0.186	−0.276
			(0.226)	(0.365)	(0.607)	(0.793)	(0.636)	(0.821)
Log distance to Wittenberg/Zürich					0.0001	−0.0001	0.0004	0.0002
					(0.0003)	(0.001)	(0.0003)	(0.001)
War frequency					−0.031**	−0.043***	−0.030**	−0.048***
					(0.014)	(0.016)	(0.013)	(0.018)
Elevation					−0.0001	−0.00002	0.0001	0.0005
					(0.0004)	(0.001)	(0.0004)	(0.001)
Terrain ruggedness					−0.003**	−0.004**	−0.002**	−0.004**
					(0.001)	(0.002)	(0.001)	(0.002)
Hub Roman road					0.156	0.494	0.129	0.431
					(0.296)	(0.634)	(0.337)	(0.7)
Roman road					−0.009	−0.023	0.081	0.151
					(0.247)	(0.467)	(0.304)	(0.603)
Oceanic port					−0.415**	−0.712*	−0.477*	−0.721
					(0.212)	(0.411)	(0.26)	(0.539)
Island					−0.057	0.185	−0.038	0.2
					(0.475)	(0.85)	(0.52)	(0.965)
Ottoman Empire							−0.621	−0.822
							(0.812)	(0.959)
Russian Empire							0.243	0.354
							(0.237)	(0.473)
Habsburg Empire							−0.075	−0.252
							(0.33)	(0.704)
Orthodox							0.279	0.725
							(0.284)	(0.614)
Protestant							0.383	0.613
							(0.235)	(0.474)
Islam							−0.939**	−1.072
							(0.446)	(0.694)

continued on next page

continued from previous page

model	bivariate		demand side		supply side		fully specified	
	Cox	logit	Cox	logit	Cox	logit	Cox	logit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Northern Europe							−0.246	−0.371
							(0.35)	(0.719)
Southern Europe							0.017	−0.264
							(0.317)	(0.695)
Eastern Europe							−0.194	−0.39
							(0.363)	(0.768)
Western Asia							−0.621	−1.207
							(0.595)	(1.043)
t		0.024		0.021		0.04		0.04
		(0.017)		(0.017)		(0.029)		(0.029)
t^2		0.00001		0.00001		−0.00001		−0.00001
		(0.00002)		(0.00002)		(0.00004)		(0.00004)
Intercept		−16.151***		−16.159***		−19.244***		−19.811***
		(3.697)		(3.767)		(6.554)		(6.452)
AIC	1664.72	516.44	1650.75	494.78	1645.65	484.26	1645.01	480.83
Observations	2,052	2,052	2,052	2,052	2,052	2,052	2,052	2,052

Note: Robust standard errors clustered by ethnic groups for all models. *Western Europe* is used as the reference category for region fixed effects and thus omitted. *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

B.3.2 Robustness checks I: state as the unit of analysis

State-only sample

Table B.4 reports the complete output of the abridged Cox proportional hazards models shown in Table 3.4 of the main body. In this series of robustness checks, I use states as the unit of analysis. "List" refers to a set of alternative dates of dictionary publications for countries with multiple official languages in the dependent variable.

Table B.4: Cox proportional hazards model outputs of the printing press' impact on language standardization.

model	List 1	List 2	List 3	List 4
	(1)	(2)	(3)	(4)
Printing press	14.711***	14.984***	14.725***	14.814***
	(0.98)	(0.92)	(1.1)	(0.97)
First year of printing press	−0.002	−0.001	−0.002*	−0.002
	(0.001)	(0.001)	(0.001)	(0.001)
University	0.0002	0.0003	0.0001	0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Bishop	−0.0003**	−0.0002	−0.0001	−0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Log urban potential	0.932	1.335**	1.254**	0.937
	(0.595)	(0.532)	(0.58)	(0.62)
Log distance to Wittenberg/Zürich	0.0007*	0.001*	0.001**	0.001*
	(0.0004)	(0.0003)	(0.0004)	(0.0004)
War frequency	−0.003	−0.001	−0.002	−0.006
	(0.013)	(0.012)	(0.013)	(0.013)
Elevation	−0.002***	−0.001	−0.002***	−0.002***
	(0.0005)	(0.0006)	(0.0005)	(0.0005)
Terrain ruggedness	−0.005*	−0.006**	−0.005*	−0.005*
	(0.003)	(0.003)	(0.003)	(0.003)
Hub Roman road	0.336	0.468	0.547	0.462
	(0.364)	(0.399)	(0.427)	(0.377)
Roman road	0.402	−0.028	0.234	0.378
	(0.324)	(0.338)	(0.38)	(0.324)
Oceanic port	−0.8**	−0.59*	−0.886***	−0.805**
	(0.337)	(0.351)	(0.344)	(0.336)
Island	0.115	0.268	0.160	0.118
	(0.258)	(0.271)	(0.262)	(0.25)
Ottoman Empire	−1.822*	−2.127**	−1.949*	−1.702
	(1.1)	(1.06)	(1.09)	(1.1)
Russian Empire	−0.244	−0.054	−0.225	−0.216
	(0.289)	(0.251)	(0.295)	(0.298)
Habsburg Empire	−0.642	−0.925**	−0.863*	−0.641
	(0.441)	(0.455)	(0.503)	(0.45)

continued on next page

continued from previous page

model	List 1 (1)	List 2 (2)	List 3 (3)	List 4 (4)
Orthodox	−0.095	0.0002	−0.164	−0.092
	(0.462)	(0.453)	(0.535)	(0.472)
Protestant	1.250***	0.268	0.885***	1.136***
	(0.303)	(0.31)	(0.316)	(0.317)
Islam	−0.371	−0.267	0.020	−0.389
	(1.163)	(0.948)	(1.051)	(1.173)
Northern Europe	−0.378	0.110	−0.343	−0.259
	(0.303)	(0.311)	(0.312)	(0.307)
Southern Europe	0.685*	0.537	0.629	0.780**
	(0.359)	(0.395)	(0.4)	(0.374)
Eastern Europe	1.105**	0.640	0.758	1.074**
	(0.5)	(0.526)	(0.593)	(0.513)
Western Asia	1.592*	0.757	1.076	1.646*
	(0.963)	(0.965)	(0.958)	(0.94)
AIC	538.29	596.87	558.7	533.87
Observations	564	564	564	564

Note: Robust standard errors clustered by country for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects so they are omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Stateless ethnic groups-only sample

Table B.5 documents Cox regression estimates using stateless ethnic groups (n = 124). This is reported in Column 5 of Table 3.5.

Table B.5: Regression outputs of the printing press' impact on language standardization.

model	Cox
	(1)
Printing press	4.915**
	(2.293)
Print year	−0.002*
	(0.001)
University	0.0002
	(0.0002)
Bishop	−0.00005
	(0.0002)
Log urban potential	0.700
	(0.967)
Log distance to Wittenberg/Zürich	0.001
	(0.0005)
War frequency	−0.032
	(0.02)
Elevation	0.0005
	(0.0006)
Terrain ruggedness	−0.004**
	(0.002)
Hub Roman road	−0.254
	(0.396)
Roman road	0.021
	(0.369)
Oceanic port	−0.434
	(0.395)
Island	−0.438
	(0.966)
Ottoman Empire	−15.591***
	(0.875)
Russian Empire	0.979***
	(0.297)
Habsburg Empire	0.543
	(0.571)

continued on next page

continued from previous page

model	Cox (1)
Orthodox	0.202
	(0.403)
Protestant	0.469
	(0.297)
Islam	−1.295**
	(0.571)
Northern Europe	−0.413
	(0.537)
Southern Europe	0.136
	(0.48)
Eastern Europe	−1.079***
	(0.412)
Western Asia	−2.881**
	(1.328)
AIC	815.41
Observations	1,488

Note: Robust standard errors clustered by country for all models. *Western Europe* and *Catholic* are used as the reference category, respectively, for geographic region fixed effects and for religion fixed effects and are omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

B.3.3 Robustness checks II: The role of human capital

Table B.6 documents the full regression estimates of Table 3.5.

Table B.6: Regression outputs of the printing press' impact on language standardization.

model	Cox (1)	logit (2)	Cox (3)	logit (4)
Printing press			1.394***	2.023***
			(0.342)	(0.558)
Vernacular Bible	0.531**	0.745**	0.461*	0.676*
	(0.251)	(0.356)	(0.257)	(0.365)
University	0.0005***	0.001***	0.0003**	0.001***
	(0.0001)	(0.0002)	(0.0001)	(0.0003)
Bishop	0.00006	0.0001	−0.00004	−0.0001
	(0.0001)	(0.0003)	(0.0001)	(0.0003)
Log urban potential	−0.157	−0.554	0.168	−0.259
	(0.613)	(0.805)	(0.616)	(0.818)
Log distance to Wittenberg/Zürich	−0.0001	−0.0004	0.0004	0.0002
	(0.0003)	(0.001)	(0.0003)	(0.001)
War frequency	−0.026**	−0.038**	−0.03**	−0.048***
	(0.013)	(0.016)	(0.013)	(0.018)
Elevation	0.0003	0.001	0.0002	0.001
	(0.0004)	(0.001)	(0.0004)	(0.001)
Terrain ruggedness	−0.003***	−0.005**	−0.003**	−0.004**
	(0.001)	(0.002)	(0.001)	(0.002)
Hub Roman road	0.217	0.531	0.054	0.267
	(0.341)	(0.693)	(0.336)	(0.712)
Roman road	0.007	0.082	0.107	0.252
	(0.29)	(0.582)	(0.31)	(0.62)
Oceanic port	−0.213	−0.317	−0.436*	−0.658
	(0.252)	(0.487)	(0.263)	(0.541)
Island	−0.36	−0.374	−0.115	0.063
	(0.507)	(0.888)	(0.511)	(0.949)
Ottoman Empire	−0.666	−0.914	−0.589	−0.834
	(0.792)	(0.907)	(0.792)	(0.892)
Russian Empire	0.018	−0.046	0.173	0.244
	(0.252)	(0.491)	(0.235)	(0.472)
Habsburg Empire	0.075	0.007	−0.088	−0.265
	(0.341)	(0.646)	(0.337)	(0.7)
Orthodox	0.467	0.963	0.286	0.786
	(0.291)	(0.588)	(0.286)	(0.602)
Protestant	0.337	0.561	0.345	0.527
	(0.248)	(0.494)	(0.236)	(0.482)
Islam	−0.913**	−1.031	−0.870*	−1.016
	(0.454)	(0.661)	(0.448)	(0.679)

continued on next page

continued from previous page

model	Cox (1)	logit (2)	Cox (3)	logit (4)
Northern Europe	−0.179	−0.259	−0.188	−0.304
	(0.346)	(0.699)	(0.347)	(0.717)
Southern Europe	0.006	−0.271	0.001	−0.363
	(0.314)	(0.700)	(0.313)	(0.72)
Eastern Europe	−0.190	−0.343	−0.102	−0.281
	(0.376)	(0.747)	(0.371)	(0.753)
Western Asia	−0.374	−0.763	−0.650	−1.296
	(0.658)	(1.025)	(0.594)	(1.039)
t		0.044		0.041
		(0.03)		(0.029)
t^2		−0.00001		−0.00001
		(0.00004)		(0.00004)
Intercept		−19.325***		−20.064***
		(6.777)		(6.476)
AIC	1660.38	495.57	1642.95	478.43
Observations	2,052	2,052	2,052	2,052

Note: Robust standard errors clustered by ethnic groups for all models. *Western Europe* is used as the reference category for region fixed effects and thus omitted. *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

B.3.4 Instrumental variables (IV) approach: Distance from Mainz as IV

Table B.7 reports the full regression estimates of Table 3.6 of Chapter 3.

Table B.7: IV probit regression outputs of the printing press' impact on language standardization.

	First stage	Second stage			
		language standardization			
dependent variable	printing press	by 1850	by 1900	by 1950	by 2000
N cumulative dictionary		n = 7	n = 27	n = 48	n = 104
	(1)	(2)	(3)	(4)	(5)
Log distance to Mainz	−0.159***				
	(0.060)				
Printing press		11.822	4.639	4.741**	4.018**
		(37.038)	(3.051)	(2.284)	(1.828)
Print adoption year	0.0003***	0.001	−0.001	−0.001	−0.0004
	(0.00004)	(0.006)	(0.001)	(0.001)	(0.001)
University	0.0001***	0.0001	0.0002	0.0001	0.0002
	(0.00003)	(0.005)	(0.0003)	(0.0003)	(0.0002)
Bishop	0.00001	−0.001	−0.0001	−0.0002	−0.00003
	(0.00003)	(0.004)	(0.0002)	(0.0002)	(0.0001)
Log urban potential	−0.004	7.111**	2.221***	1.280***	1.275***
	(0.073)	(3.010)	(0.217)	(0.099)	(0.074)
Log distance to Wittenberg/Zürich		−0.005	0.001**	0.001***	0.001***
		(0.005)	(0.001)	(0.0003)	(0.0003)
War frequency	−0.002	0.007	−0.011	−0.022*	−0.025***
	(0.003)	(0.092)	(0.022)	(0.013)	(0.009)
Elevation	0.0001**	0.005	−0.0002	−0.0004	−0.0001
	(0.00003)	(0.011)	(0.001)	(0.001)	(0.0004)
Terrain ruggedness	−0.0002**	−0.044	−0.002	−0.001	−0.001
	(0.0001)	(0.070)	(0.003)	(0.002)	(0.001)
Hub Roman road	0.131**	−6.338	−0.458	−0.480	−0.315
	(0.052)	(4.421)	(0.596)	(0.456)	(0.390)
Roman road	−0.008	1.358	0.473	0.268	0.141
	(0.040)	(3.645)	(0.576)	(0.449)	(0.349)
Oceanic port	0.041	0.444	−0.380	−0.527	−0.538*
	(0.043)	(3.262)	(0.459)	(0.366)	(0.286)
Island	−0.005	6.790***	1.158**	0.452	0.392
	(0.053)	(2.019)	(0.502)	(0.429)	(0.417)
Ottoman Empire	−0.049	4.178**	−0.220	−0.151	−0.322
	(0.045)	(1.643)	(0.611)	(0.480)	(0.467)
Russian Empire	0.020	−2.738***	0.253	0.262	0.320
	(0.026)	(0.766)	(0.526)	(0.351)	(0.275)
Habsburg Empire	0.118**	−3.027**	−0.673	−0.668	−0.590
	(0.046)	(1.300)	(0.777)	(0.596)	(0.445)
Orthodox	−0.144**	2.110	0.865	0.544	0.526
	(0.060)	(3.658)	(0.594)	(0.452)	(0.352)
Catholic/Protestant	−0.213***	−0.390	−0.390	−0.454	−0.269
	(0.043)	(5.030)	(0.697)	(0.468)	(0.372)
Islam	−0.124**	0.158	−4.899***	−5.652***	−0.680*
	(0.056)	(4.829)	(0.984)	(0.609)	(0.386)
Northern Europe	−0.001	6.131***	0.020	0.270	0.109
	(0.062)	(1.878)	(0.480)	(0.400)	(0.387)
Southern Europe	0.156**	5.603***	0.102	−0.297	−0.058
	(0.067)	(2.136)	(0.549)	(0.395)	(0.315)
Eastern Europe	0.060	9.491**	0.614	0.044	0.017
	(0.070)	(4.449)	(0.846)	(0.591)	(0.444)
Western Asia	0.071	11.974*	1.308	0.198	−0.566
	(0.078)	(6.995)	(1.532)	(1.097)	(0.622)
Intercept	1.28***	−28.929	−9.522***	−6.967***	−6.992***
	(0.413)	(18.163)	(1.138)	(0.749)	(0.598)
Observations	2,052	1,539	1,710	1,881	2,052
p-value, Wald exogeneity test		0.75	0.12	0.038*	0.028*
F-statistic on weak instrument	24.99				

Note: Robust standard errors clustered by ethnic groups. IV probit estimation: first stage is OLS, second stage is probit, regressed on predicted values from the first stage. In the first stage, *log urban potential* and *war frequency* are for the fourteenth centuries. *Western Europe* is used as the reference category for region fixed effects and thus omitted. In the second stage, *Catholic* is used as the reference category for religion fixed effects and thus omitted. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Appendix C

APPENDIX FOR CHAPTER 4:"DOES CULTURAL CONSOLIDATION LEAD TO PUBLIC GOODS PROVISION? "

C.1 Coding rules for the variables used in regression analysis

This section discusses my coding strategies and the sources of the variables used in the parametric models (shown in Table 4.2).

Unit of analysis

The ethnic group is the unit of analysis in my paper. I built a data set on 171 European ethnic groups that existed for 1400–2000. I draw primarily on Minahan (2000) for entries on these ethnic groups and add others referred to, but not entered, in Minahan's volume. I double-check them using *Ethnologue* compiled by Lewis, Simons, and Fennig (2013). I identify the main homeland city of each group and collect relevant attributes based on it.

Dependent variable

The outcome variable is the timing of language standardization. As discussed in the main body, I proxy it for the first publication date of a comprehensive, monolingual vernacular dictionary. By "comprehensive" I mean dictionaries that try to cover the entire alphabets. Dictionaries are an adequate proxy, because they require high fixed costs of understanding and developing a tongue. They also require literacy on the part of the users. The dictionary data rely mainly on Burke (2004), Dalby (1998), and Price (1998). For some ethnic groups in the Caucasus, I use Nordhoff et al. (2013).

Covariates

A main covariate of interest discussed in Chapter 4 is the time elapsed since the adoption of the printing press. The print variable comes from various sources but primarily from Clair (1976) and

Febvre and Martin (1976), which cover Western, Southern, and Central Europe. I also use Burke (2004), Price (1998), and Steinberg (1974) as supplementary sources. For the ethnic groups in the German-speaking area, I use Reske (2007).

The sources of the year of universities' foundation come from Darby and Fullard (1970) and Rüegg (four volumes, 1992–2011) on the history of European universities. Those of the bishoprics come from Chaney (2015).

Urban potential is a proxy for economic growth and often used in economic history works. It is originally constructed by de Vries (1984) as the sum of the populations in the given period divided by the geographical distance between a city and all others in my data set. More formally, this variable is obtained by $\sum_{j=1}^{N} \frac{\text{population}_j}{\text{distance}_{ij}}$. This variable gives a sense of whether a city is surrounded by competing urbanizing towns or located in a more sparsely-populated area. Higher values indicate greater potential for urbanization. The standard source on population size in preindustrial Europe is Bairoch, Batou, and Chévre (1988), which covers the period 800 through 1850 for hundreds of European cities. Bosker, Buringh, and van Zanden (2013) correct some of the data in Bairoch et al. and I follow their updates in compiling my data. For the period after 1850, I use several statistical handbooks including Mitchell (2003*a*,*b*). I also include an indicator for an *oceanic port*, taking the value of 1 if an ethnic group's homeland is located near or on the coast.

Geographical distance to Wittenberg or Zürich captures the impact of the Protestant Reformation. These two cities played a crucial role in the spread of Protestantism, in which shorter distance indicates higher likelihood that ethnic groups would be exposed to this religious movement. Following Pfaff and Corcoran (2012), I construct this variable by taking the nearest distance between the given ethnic homeland and either of the Protestant centers.

War frequency represents the effect of war. I draw on the database compiled by Brecke (1999) for war-related data. It records any conflict in the world with the minimum of 32 casualties in the period between 1400 and 2000. As Brecke uses states as the unit of analysis, I take care to localize the incidence of war at the ethnic-group level to the extent possible. For instance, if a war occurred

in the Habsburg Empire, I regard that war as affecting all ethnic groups within the imperial domain. The frequency is coded at a fifty-year interval, and I take the average to permit cross-sectional analysis.

I also collect data on other geographical, institutional, and political variables. One is a set of geographical attributes which include *elevation above the sea level* and *terrain ruggedness*. These data are found in the GLOBE (Global Land One-kilometer Base Elevation project) database (GLOBE Task Team and others 1999). Values for terrain ruggedness in the GLOBE data set come from the Terrain Ruggedness Index, originally devised by Riley, DeGloria, and Elliot (1999). A third geographical variable is an indicator on the island, taking the value of 1 if an ethnic group's homeland is located on an island.

I also include a set of institutional variables related to the Roman Empire's influence. Romans built roads across the cities they reached, which gave these cities an opportunity to develop institutional capacity in economic activity and political organization. I use Talbert (2000), Hammond (1981), Åhlfeldt (2015), and *Pleiades* (2015) to collect information about the Roman legacy. More specifically, I create an indicator taking the value of 1 if a city had major or minor Roman roads. Following Bosker, Buringh, and van Zanden (2013), those cities with two or more major roads are coded "Roman hub" and those with one major road or one or more minor roads are coded cities with "Roman roads."

Finally, I account for the impact of major European empires. I include *time elapsed under imperial rule* for the Ottoman, Russian, and Habsburg Empires based on Nüssli (2011) which shows the political "owner" of each European city that existed at the beginning of each century. I also include indicators for religion (Catholicism, Protestantism, Orthodoxy, and Islam) and for geographical region as fixed effects.

C.2 Summary statistics

Table C.1: Summary statistics.

	N	mean	median	sd	min	max
Outcome variable						
Year of vernacular dictionary	104	1935	1954	51	1755	2005
Main explanatory variable						
Year of printing press adoption	96	1566	1526	115	1460	1930
Control variables						
Year of vernacular Bible	144	1789	1837	193	1380	2012
University	67	1672	1735	208	1150	1925
Bishop	99	1250	1400	608	100	2000
Elevation (m)	171	405.3	145	581	−24	2,823
Terrain ruggedness (m)	171	137.51	59.87	205	0	1,125
Urban potential	171	8.43	2.89	14.99	0.35	99.45
War frequency (in 50-year periods)	171	8.05	5	8.75	0	37
Distance from Wittenberg (km)	171	1,669	1,429	948	89	5,292
Distance from Zürich (km)	171	1,711	1,516	1,048	0	3,350
Hub Roman road	171	0.12	0	0.33	0	1
Roman road	171	0.32	0	0.46	0	1
Oceanic port	171	0.38	0	0.49	0	1
Island	171	0.1	0	0	0	1
Russian Empire/Muscovy	171	0.15	0	0.36	0	1
Ottoman Empire	171	0.105	0	0.31	0	1
Habsburg Empire	171	0.08	0	0.27	0	1
Orthodox	171	0.16	0	0.37	0	1
Catholic	171	0.35	0	0.47	0	1
Protestant	171	0.17	0	0.1	0	1
Islam	171	0.29	0	0.45	0	1
Western Europe	171	0.16	0	0.36	0	1
Northern Europe	171	0.14	0	0.35	0	1
Southern Europe	171	0.16	0	0.40	0	1
Eastern Europe	171	0.40	0	0.49	0	1
Western Asia	171	0.10	0	0.30	0	1

C.3 Full regression output used in the main body

Table C.2: OLS and logistic regression outputs of language standardization by 1900 among 171 European ethnic groups.

model	logit (1)	OLS (2)
Time since print adoption	0.006*	0.0005*
	(0.003)	(0.0003)
Time since university founded	0.003*	0.001**
	(0.002)	(0.0002)
Time since bishopric erected	−0.001	−0.0001
	(0.001)	(0.0001)
Mean urban potential	0.074	0.001
	(0.349)	(0.041)
Log distance to Wittenberg/Zürich	−0.0002	−0.00003
	(0.001)	(0.0001)
Mean war frequency	0.071	0.009
	(0.080)	(0.010)
Elevation	−0.0002	0.00002
	(0.002)	(0.0001)
Terrain ruggedness	−0.002	−0.0001
	(0.005)	(0.0002)
Hub Roman road	0.344	0.039
	(0.861)	(0.107)
Roman road	0.345	0.051
	(0.736)	(0.081)
Oceanic port	−0.395	−0.048
	(0.909)	(0.081)
Island	0.941	0.100
	(1.120)	(0.111)
Time under Ottoman Empire	−0.003	−0.0004
	(0.004)	(0.0004)
Time under Russian Empire	−0.001	−0.0002
	(0.004)	(0.0004)
Time under Habsburg Empire	−0.003	−0.0005*
	(0.003)	(0.0003)
Orthodox	1.745	0.193
	(1.469)	(0.123)
Protestant	0.801	0.086
	(0.852)	(0.099)
Islam	−15.527	0.032
	(1, 395)	(0.126)

continued on next page

continued from previous page

model	logit	OLS
	(1)	(2)
Northern Europe	−0.754	−0.147
	(1.080)	(0.126)
Southern Europe	−0.573	−0.108
	(1.020)	(0.115)
Eastern Europe	0.316	−0.014
	(1.166)	(0.135)
Western Asia	1.066	−0.0001
	(2.568)	(0.168)
Intercept	−4.270	0.051
	(2.867)	(0.307)
Observations	171	171
F-statistic		2.49^{***}
Wald Chi square	3.58^{*}	
R^2	0.342	0.162

Note: Western Europe is used as the reference category for region fixed effects and thus omitted. *Catholic* is used as the reference category for religion fixed effects and thus omitted. The Wald Chi Square is the effect of time since print adoption. R^2 for logistic regression is pseudo-R^2. *** denote $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

VITA

Yu Sasaki was born in Sapporo, Japan. He earned a BA in International Studies at American University, Washington, D.C., in August 2002. He then worked at several research institutions in Washington, D.C., and Tokyo on international and homeland security. He earned an MA in Global, International, and Comparative History at Georgetown University in May 2009. He also holds an MA in Political Science at the University of Washington in June 2012 and the Political Methodology Certificate from the Center for Statistics and the Social Sciences. One of his dissertation essays, "Publishing Nations," will appear in *The Journal of Economic History* in early 2018.