19-21 de Octubre 2022 | Granada
INTERNATIONAL CONFERENCE ON REGIONAL SCIENCE
Challenges, policies and governance of the territories in the post-covid era
Desafíos, políticas y gobernanza de los territorios en la era post-covid
XLVII REUNIÓN DE ESTUDIOS REGIONALES
XIV CONGRESO AACR

COMUNICACIÓN

Título: State capacity and the uneven cost of nation building:

language mismatch and literacy levels in Valencia

Autores y e-mails de todos:

Francisco Beltrán-Tapia (NTNU, Noruega)
Alfonso Diez-Minguela (U. Valencia)
Alicia Gómez-Tello (U. Valencia)
Julio Martínez-Galarraga (U. Barcelona)
Daniel A. Tirado-Fabregat (U. Valencia)
email.- daniel.tirado@uv.es

S07 – New approaches to the historical economic geography of Spain

This paper studies the impact on human capital formation generated by the existence of a mismatch between the native language and the language of instruction in the context of the construction of the liberal state in Spain. Relying on a novel dataset with information on literacy rates and institutional and economic characteristics of the 524 municipalities that make up the region of Valencia and the use of Propensity Score Matching techniques, the analysis shows that the mismatch effect is only visible when the Spanish state enjoyed the capacity to force compliance with language regulations in public schools, in parallel with the advance of its financial and administrative capacity and the incipient advance of a democratic regime.

Keywords.- Language mismatch, literacy, State capacity, Nation building, Spain.

JEL Classification.- I28, O17, N33, N43, J24

Introduction

The so-called "*nation building*" literature stresses that heterogeneities within a territory in terms of ethnicity, language or religion, could harm its economic development by reducing the provision of public goods (Alesina and La Ferrara 1999; Miguel 2004) or by increasing the probability of conflict and social unrest (Alesina, Reich, and Riboni, 2020). Under these circumstances, central states often implement policies aimed at cultural homogenization, such as compulsory schooling in an specific language.[1] In this regard, it has been recently suggested that the threat of democratization encouraged cultural homogenization in the past (Alesina, Giuliano, and Reich, 2021).

Similarly, it has been noted that the effectiveness of cultural homogenization largely depends on the financial and administrative capacity of a political body or State (Alesina, Giuliano, and Reich 2021). Then, nation building and State capacity, understood as the ability to carry out the political will of a government (Skocpol 1985), appear to be strongly related. State capacity therefore arises as a crucial element to secure the success of the aforementioned policies. In the context of human capital accumulation, if the State lacked the capacity to enforce cultural homogenization and there are training costs associated with a language mismatch, then diversity could give rise to uneven development.

Yet, not much has been discussed on the limitations, challenges and the effectiveness of cultural homogenization policies. For the United States, Bandiera *et al.* (2019: 43) found that compulsory schooling laws were passed "*significantly earlier*" in states where European migrants from countries where compulsory schooling was not present by mid-19th century predominated. Interestingly,

[1] The use of a common language in compulsory education has long been regarded as the main tool or mechanism of homogenization in the history of education, political science and sociology.

English was imposed in some states whereas bilingualism remained in others, which was only banned during the 20th century. Still, English-only laws appear to have a moderate effect on immigrants' literacy (Lleras-Muney and Shertzer, 2015).[2]

The relevance of the language of instruction for enrolment and performance in basic instruction has also been explored in other contexts, especially regarding the existence of a training cost associated with a language mismatch in elementary instruction. Analysing India, Jain (2017) shows that districts where a language mismatch was present had significantly lower levels of literacy because being schooled in a different language requires a greater effort. As shown in investigations like that by Williams and Cooke (2012), aimed at identifying the best strategies for spreading literacy, educational outcomes were worse when these projects involve teaching in a language other than that of the potential beneficiaries. Similar conclusions are found when looking at immigrants in receiving countries. Bleakley and Chin (2004), for example, show that the education results for children who migrated to the US from English-speaking territories were significantly better than those for children with other mother tongues.

Additionally, the language mismatch can also affect the provision of education. This supply-side perspective finds an economic foundation in works such as those by Alesina, Baqir, and Easterly (1999), that designed a theoretical framework with which to analyze the provision of certain public services in the presence of heterogeneous agents. This literature has pointed out that the rules governing the education system can affect performance if they give rise to a mismatch

[2] Lleras-Muney and Shertzer (2016:286) concluded that the Americanization movement could have contributed to the assimilation and integration of immigrants but "*we find no evidence that the laws regulating English in schools contributed significantly to that process*". In other contexts, such as 19th century Germany, it seems that nation building policies had an impact on voting preference for nationalist parties and identity (Cinnirella and Schueler, 2018; Kersting and Wolf, 2019). Also, it has been examined to what extent nation building policies could have an impact on the provision of educational or sanitation infrastructure between countries in the context of the decolonization processes in Africa (Miguel 2004).

between the language of instruction and the prevailing language (or mother tongue).[3] More specifically, it has been posed that this effect could be especially relevant in cases where funding was a local affair and the elites regarded the use of a different language as a threat to their position of privilege.

In this line of thinking, Cvrcek and Zajicek (2013) find evidence for the Austro-Hungarian Empire that in territories where the language of schooling was the same as that of the local elites, education received greater financial support than in those where there was a distance between the two. Likewise, Cinnirella and Schueler (2016) show that the provision of primary education in Prussia was lower in those territories where German-speaking citizens lived alongside speakers of other Slavic languages (mainly Polish) in a context in which a Germanization policy had imposed German as the only language of schooling throughout the Empire after its foundation in 1871.

Overall, these studies stress that educational outcomes improved once the existing mismatch was eradicated. Although it is worth noting that higher enrolment rates or better academic performance in basic instruction might not necessarily guarantee better labor prospects, they can act as a catalyst in places where reading and writing are not widespread.[4] Furthermore, if the language mismatch only occurs in some territories, this might lead to an uneven regional development.

[3] Particularly, Alesina and La Ferrara (2005) carry out an in-depth survey of empirical works that analyse the existence of a relationship between the ethnolinguistic diversity of a territory and the provision of public goods.

[4] Postelementary instruction appears to be more closely related to labor market outcomes. Angrist and Lavy (1997), found that the introduction of Arabic in Morocco for grades 6 and above negatively affected the returns to schooling and attributed this to a loss of French writing skills.

This article contributes to this literature by examining the effectiveness of cultural homogenization, such as compulsory education in a particular language, in territories where there is a mismatch, that is, where the prevailing language was not the language of instruction. For this, we examine the region of Valencia between 1860 and 1930, a context where two languages have long coexisted. While Catalan (locally known as *valencià*) has prevailed in around two thirds of the municipalities, the remainder are Spanish-speakers. This linguistic diversity goes back to the medieval and modern periods when the territory was settled and, in some cases, re-settled.[5] More importantly, and except for some major cities, these two languages did not coexist within-municipalities at the onset of compulsory mass schooling. As a result, the linguistic frontier is clear-cut. Although Spanish was introduced as official language during the Bourbon reforms of the second half of the 18th century, its use as language of instruction was not enforced until 1857, when the Public Instruction Act (or Moyano Act) instituted compulsory elementary education.[6] Yet, the funding of elementary instruction remained a local affair until 1902 when the expenses were incorporated into the budget. In this way, it can be argued that, within our period of study, Spanish was first introduced (1857) and then enforced (1902) as language of instruction. Using this case study therefore offers a riveting opportunity to better understand the challenges and limitations that weak States faced.

The analysis relies on a dataset for the universe of municipalities that made up the region of Valencia (currently known as *Comunitat Valenciana*) between 1860 (earliest census available) to 1930, just before the 2nd Spanish Republic (1931) and the Civil War (1936-39). The dataset thus

[5] The expulsion of the Moriscos offers further insight. In the case of lordships that were formerly inhabited by Moriscos, however, the language of use depends not on the origins of the settlers or their lords after the Reconquista, but on the origins of the new colonists who settled after 1609 (Casanova 2001).

[6] The compulsory use of Spanish for teaching throughout the entire kingdom of Spain dates back to 1768, as established by Royal Decree by Carlos III.

offers ample municipal-level information at different points in time. Our results show that, although the presence of a language mismatch hampered human capital formation, this outcome is only visible when the liberal State had the capacity to enforce the Spanish-only schooling which essentially occurred in the early 20th century when the financial and administrative capacity of the central administration increased. As a corollary it can also be concluded that compulsory schooling in Spain brought about a territorial uneven distribution of the costs associated with nation building.

Historical background

The creation of Spanish and Catalan linguistic areas in the Valencia region

The analysis of the determinants of literacy levels in the municipalities of Valencia in the period 1860-1930 provides us with a good opportunity to identify the existence of a mismatch effect for at least two reasons. On the one hand, two main languages, Spanish and Catalan, have long coexisted in the region. According to the law of 1983 governing the use and teaching of *valencià* (as it is referred to locally), Catalan is the main language of use in two thirds of municipalities, while Spanish predominates in the other third. The situation in the mid-nineteenth century was very similar, since the territorial spread of one language or the other mainly depended on what happened before the development of the liberal state and the situation has hardly changed up to the present day.[7]

[7] Possibly the most significant change that occurred during the contemporary era was the incorporation into the province of Valencia of the municipalities located in the *comarca* or district of Utiel-Requena. This territory, which until 1851 was part of the kingdom of Castile, was Spanish-speaking. The same thing happened with the municipalities of Sax, which was transferred to the province of Alicante in 1836 from the province of Murcia (the provinces having only been created in 1833), and Villena, which that same year became part of the province of Alicante having previously belonged to Albacete. Another change occurred in Paterna, a Catalan-speaking municipality that was joined by a local council area called San Antonio de Benagéber in 1957. This small settlement was home to a group of people who came from a Spanish-speaking municipality, Benagéber, which

Particularly, according to Casanova (2001), the establishment of one language or the other is related with two main historical events. Firstly, it can be traced back to the granting of lordships by the Crown of Aragon to the nobles who took part in the conquest of Valencia (Aragonese or Catalans) and who in exchange were given rights over the various territories and therefore guided the settling of new inhabitants. As far as the geographical distribution is concerned, Catalan lords tended to be located in territories nearest the coast, while the Aragonese were in the interior (see Figure 1).[8]

Secondly, the map was partially redefined by the shock caused to the Valencia area by the expulsion of the Moriscos in 1609 and the subsequent repopulation process. Historiographers have analysed the origins of the settlers and recorded that most of them came from within the kingdom of Valencia, generally from neighbouring districts, although some groups have also been identified as coming from coastal towns and even from outside the region (from Mallorca, La Mancha, France and Italy).[9] The language that ultimately became the language of use in each territory was therefore not necessarily the one spoken by the local lord, but that of the new settlers in each of these municipalities.[10]

---

disappeared after the construction of a reservoir (of the same name) in 1952. In 1989 San Antonio de Benagéber and Paterna separated and the new municipality, San Antonio de Benagéber, became a Spanish-speaking island surrounded by Catalan-speaking municipalities.

[8] This aspect needs to be taken into account in the subsequent analysis given that it could generate an element of endogeneity in the distribution of the languages over the territory, thus introducing a bias into the study of its impact on literacy levels.

[9] The repopulation of the municipality of Tàrbena with settlers from Mallorca is particularly well-known, resulting in the village belonging to the Catalan linguistic domain. Even today it still retains vocabulary and certain cultural aspects typical of the Balearic Islands.

[10] Until 1609, for example, the linguistic frontier in the *comarques* of the Sierra de Espadán and La Plana Baixa was marked by the frontier between the Diocese of Tortosa (Catalan) and the Diocese of Segorbe (Spanish). However, after the expulsion of the Moriscos, repopulation took the Catalan language to the municipalities of Aín, Artana, Tales, Suera, Betxí and Veo. In the Camp de Morvedre, despite the general predominance of Catalan, municipalities such as Gátova and Marines became Spanish-speaking areas after 1609 (Casanova 2001). According to Sanchis Guarner (1973), Catalan used to be spoken in the Vega Baja del Segura due to the origins

It should be stressed, as noted by linguists, that in Valencia there are no transitional dialects separating Spanish-speaking and Catalan-speaking areas.[11] Naturally the two languages have had an influence on each other in frontier areas: basically Catalan to Spanish until the modern era, and Spanish to Catalan since then. Nonetheless, it can safely be assumed that, apart from among certain local elites (justice, administration), the languages used by the vast majority of people in territories identified as Catalan-speaking or Spanish-speaking were Catalan and Spanish respectively.[12]

All in all, it can be considered that the map that draws the use of Catalan or Spanish in the region of Valencia was completely defined before the period under analysis here and that the use of one or the other language was related to historical processes that occurred independently of the level of economic development of the territories occupied during the Reconquest. This element limits the relevance of the potential problem of endogeneity that the subsequent empirical analysis must face, although, as can be seen in Figure 1, it points to the existence of a relationship between linguistic dominance and geographical aspects.

<Insert Figure 1>

of its settlers in the fourteenth and fifteenth centuries, but from the sixteenth century it gradually became a Spanish-speaking area because of repopulation by settlers from Murcia, which increased after the expulsion of the Moriscos.

[11] Aragonese, a dialect spoken by some of those who settled in the kingdom of Valencia after the conquest, left a linguistic trace in some of these territories, but it gradually disappeared as a language of use. In territories that today are Catalan-speaking it was absorbed into Catalan for a number of reasons, including the demographic hegemony of the settlers from Catalonia and the prestige of a language that became the official language of the kingdom of Valencia due to its use by the political authorities in administering the municipalities and the incipient State. In what are today's Spanish-speaking areas it was absorbed in the same way by the Spanish language over the centuries following the conquest, especially during the modern era (Guinot 1999, p.262).

[12] This means that when studying the case of Valencia, we do not need to estimate the percentage of the population that had one or the other language as their language of use, as was indeed necessary in other research aimed at analysing the impact of the linguistic mismatch in Prussia (Cinnirella and Schueler 2016) and India (Jain 2017).

The enforcement of the compulsory use of Spanish language in primary education

An analysis of the existence of a linguistic effect that may have affected educational demand in the case of Catalan speaking municipalities in Valencia should take the preliminary step of proving that during the period analysed teaching the three Rs in schools (reading, writing and arithmetic) was actually done in Spanish. From a formal perspective, the obligation for the language of schooling to be Spanish dated back almost a century, to 1768. This was done in a royal decree by Carlos III as part of the centralization of the State imposed by the Bourbon dynasty after their victory in the Spanish War of Succession (1714).

It is important to point out that the Royal Decree establishing the obligatory use of Spanish as a formative language in primary schools also sanctioned two other measures aimed at building a more homogeneous society. On the one hand, in the same royal decree, the tariffs levied on judicial processes for the whole territory were unified and, on the other hand, the Castilian *real de vellón* was imposed as the unit of account for their payment. These are three measures that jointly promoted monetary, fiscal and language unification, with the aim of fostering the integration of the Spanish market and to increase the national mindset among its citizens.

This commitment to the cultural and economic unification of the country was in the philosophy of the enlightened liberals of the eighteenth century, was given shape and form in the Constitution of 1812, and was continued during the transition from the Ancien Regime to the nineteenth century liberal state. With regard to education, the compulsory use of Spanish as the common language appeared in all the documents involving projects and plans put forward from that time onwards (the Quintana Report of 1814, the Plan Rivas of 1836, the Primary Instruction Plan of 1838). Nevertheless, it was not until 1857, a time marked by great socioeconomic change in which various countries were taking their first steps towards mass schooling and the use of

education as a nation-building policy, that the Public Instruction Act (PIA), commonly known as the Moyano Law, was passed. This regulated the Spanish education system from 1857 to 1970, when it was replaced by the General Education Act (GEA).[13]

Indeed, the Moyano Law was one of the great reforms introduced in Spain in the nineteenth century.[14] Education was split into primary education and higher education (Art. 1), with a curriculum being established for both. At the same time, primary education would become "compulsory for all Spaniards" (Art. 7) between the ages of 6 and 9, but was only free in cases where the "parents, guardians or providers are unable to pay for it" (Art. 9).[15] Besides, according to Art. 2 of the PIA, primary education comprised 6 subjects: Christian doctrine and basic scripture, reading, writing, principles of Spanish grammar, principles of arithmetic, and basic knowledge of agriculture, industry and commerce (this latter subject depending on location), reaffirming the compulsory nature of the study of the first letters in Spanish.

To summarize, the regulatory framework for primary education in force in Spain in 1860 served to organize and homogenize the education system, set out its stages and establish the content of

[13] Although the PIA continued until 1970, successive changes were introduced during the 113 years that it remained in force. Compulsory education, for example, initially from ages 6 to 9, was extended to age 12 in 1909 and age 14 in 1964.

[14] This was not the only area in which the deployment of the liberal state gradually promoted administrative, territorial or economic homogenization, as well as the economic integration of Spanish regions. In a non-exhaustive manner, we could highlight the establishment of a homogeneous territorial administrative structure with the creation of the provinces (1833) and the organization of the territory into judicial districts (1834) and municipalities (1835), the reform of the Treasury (1845), the deployment of the postal service, the law for the organization of the telegraph network (1855), the law to promote the construction of railway infrastructures (1855) or the monetary unification around the peseta (1869). All of them, as a whole, would make up a broad program aimed at the creation of a homogeneous national conscience and the economic impulse through the integration of the Spanish market.

[15] The compulsory nature of education was not absolute, since pupils could ask to be excused when they were "*sufficiently provided with this type of education in their homes or in a private establishment*" (Art. 7). And to obtain free primary education, a "*certificate issued by the relevant parish priest and endorsed by the town mayor*" had to be provided (Art. 9).

each subject. It also listed the obligations that fell to families (the schooling of their children) and municipalities (the opening of schools). However, from the point of view of funding, the reality that was characteristic of the Spanish education system under the Ancien Regime remained unchanged. The provision of educational infrastructures in 1860 was the responsibility of the local authorities and funding was still expected to come from local entities and families and, to a lesser extent, from secular or religious foundations. In these circumstances, the State's capacity to promote primary education and force the compliance with the regulations that stablished Spanish as the compulsory learning language was very limited.

In fact, the widespread use of local languages in many of the regions that made up the kingdom of Spain (also visible in many documents of the time) leads us to think that perhaps the vernacular languages were used for teaching in municipal schools, especially by parish priests or charitable bodies. Although it is impossible to quantify how widespread this practice was, it has been confirmed by various investigations on the history of education in Spain (Escolano 1997). In any case, it seems a reasonable hypothesis that the construction of the liberal State and the growing interest in creating a kind of national mindset would have favoured the establishment of Spanish as the language of teaching in municipal schools and that the use of regional languages would have gradually diminished.[16]

Historians of education have defended the growing use of Spanish in schools on the basis of evidence relating to teacher training. De Gabriel (1994), for example, has shown that, since the Normal Schools for teacher-training were created in 1849, their curricula only ever considered

[16] This process of linguistic unification would come about contemporaneously in other European states that were in the process of consolidation over the same period. Fouret and Ouzof (1980) describe the process, seemingly very similar to that of Spanish in Spain, whereby French was introduced in schools in territories that had their own languages such as Corsican and Breton. Cinnirella and Schueler (2016) describe the introduction and spread of German in Polish-speaking territories during the time of the Empire.

Spanish as the language in which to teach grammar or the history and geography of Spain. Another stream of the literature has tackled the problem by analysing the materials and textbooks used in primary and secondary teaching, confirming that all of them were published in Spanish.[17]

The compulsory use of Spanish in primary education gained momentum with the creation of the Ministry of Public Instruction and Fine Arts in 1900.[18] From the mid-nineteenth century and during the early twentieth century, the Spanish economy and society underwent a number of important changes, such as rapid population growth, accelerated economic structural change and the extension of the electoral franchise.[19] It was in this context of far-reaching socioeconomic change that the Ministry of Public Instruction and Fine Arts was created, followed two years later by the central government taking over the funding of primary education. In particular, since 1902, the state took over the salaries of public-school teachers, who were then considered civil servants of the administration. Under these conditions, the construction of a centralized and publicly financed educational system provided the liberal state with the necessary legitimacy to force greater compliance with linguistic regulations. Proof of this are the words of the Minister of the Interior, Eduardo Dato, in his speech to the Senate in 1900:

"It is sensible that not all Spaniards know the national language; but it is an undeniable fact about which nothing can be done other than constant propaganda aimed at extending the knowledge of the official language, preventing the teaching of a

[17] It is significant that the widely-used *El instructor de la Juventud* (Young People's Instructor) was published by Esteban Paluzie in 1845 in Barcelona but was of course in Spanish. Textbooks from this publisher, such as the *Manual de Escritura y Lenguaje* (Manual of Writing and Language), were used in the Normal School of Barcelona.

[18] With the Finance Act of 31 March 1900 came the reorganization of the Ministry of Public Works, and after the Royal Decree of 18 April 1900 came the creation of the Ministry of Agriculture, Industry, Trade and Public Works and the Ministry of Public Instruction and Fine Arts (1900-1936). The Ministry of Agriculture, Industry, Trade and Public Works would revert to being called the Ministry of Public Works from 1905 onwards.

[19] General male suffrage replaced census suffrage in 1890. On the social and economic changes recorded in this period see Pérez Moreda, Reher, and Sanz (2015) and Prados de la Escosura (2017).

language other than Spanish in State Schools, not allowing texts written in any dialect as teaching books, and enforcing the authorities the measures that have been adopted for this purpose."

With the legitimacy granted by the financing of primary education in public schools from the general State budget, the Government decreed a new regulation that imposed sanctions on teachers, now State officials, who failed to comply with the linguistic regulations. Thus, the Royal Decree of 21 November 1902, signed by King Alfonso XII, established in its article 2:

"Teachers of primary instruction who teach their disciples the Christian doctrine or any other subject in a language or dialect other than Spanish will be punished for the first time with a warning by the Provincial Inspector of primary education, who will report the fact to the Ministry of the branch; and if they reoffend, after having received a reprimand, they will be separated from the official Magisterium, losing all the rights recognized by law."

The approval of this decree raised a wave of protests from regionalist political circles or from the Catholic Church itself, responsible for a relevant fraction of the primary schools. In the parliamentary debates opened after its promulgation, the Minister of Public Instruction, *Conde de Romanones*, vehemently exposed the reasons why he considered the pertinence of the Decree:

"To be a teacher you need neither to be nor to speak Catalan, but rather to speak Spanish (Castilian). […] Is it possible to consent to good principles, especially now that public teachers are State officials, for having passed the passage of this attention to the State; Can it be allowed that schools in Spain teach a language or dialect other than Spanish? This is the question that I ask all the Deputies. […] Within families, in the domestic home, all the languages or dialects that one wants can be spoken; In the State school, supported and supervised by the action of the State, it is only possible to teach in the national language, and the national language in Spain is Spanish..."

Therefore, the existing qualitative evidence points to the strengthening of state pressure to enforce the obligatory use of Spanish in primary education. Moreover, the relevance of the tightening of regulations was justified on two fundamental bases: that primary education was now financed by the State and that the State only recognized one national language, Spanish.[20] As a result of this process, historians of education have argued that Spanish gradually became the only language of schooling in primary education across Spain (Escolano 2012) and therefore the use of regional languages practically disappeared during the first years of the twentieth century.

Based on this evidence, it is plausible to argue that the increasing imposition of the compulsory use of Spanish in public schools, in a context in which great advances are made in mass schooling (Beltrán Tapia et al. 2019), would have increased the educational cost of those Valencian children who lived in Catalan-speaking territories and who were now forced to receive their instruction in Spanish. In these circumstances, it could be hypothesized the potential existence of a relationship between linguistic mismatch and literacy levels in Valencian municipalities and its growing relevance once the liberal state was able to enforce the regulations regarding the language of instruction.

[20] In this same period, the liberal State limited the use of regional languages in communications with the public administration by postal mail or telegraph. Royal Decree of 4 June 1904, signed by King Alfonso XIII at the proposal of the Minister of the Interior, José Sánchez Guerra.

Sources, data description and descriptive analysis

The empirical analysis of the hypothesis posed in the previous section relies on an original dataset containing information on all Valencian municipalities measured at three points in time: 1860, 1900 and 1930. One of the challenges to build the dataset is to make the units of observation comparable over time since the number of municipalities changed due segregation and fusion phenomena that took place during the period of study.[21]

The main variable of the analysis, the literacy rate, is obtained from the population censuses and is measured as the percentage of the total male population who could read and write[22]. Table 1 reports basic descriptive statistics of men literacy rate by municipality for the three censuses. The average of male literacy rate increased from 14.3 percent in 1860, to 26.4 in 1900, to 53.6 percent in 1930. However, one of the most striking characteristics of Spanish literacy is that it varied greatly across the territory. This is true not only when comparing the regions that made up the kingdom of Spain (Núñez 1992), but also when comparing municipalities within the same region or province (Beltrán Tapia et al. 2019). A good illustration of this variability across municipalities is the region of Valencia, as shown in Table 1 and Figure 2.

<Insert Table 1>

<Insert Figure 2>

The other key variable in this analysis is the one that classifies municipalities according to their linguistic domain, namely between Catalan or Spanish. Following the *Llei d'ús i ensenyament del valencià* published by the Generalitat Valenciana in 1983 and assuming that this situation is similar

[21] Appendix B provides a detailed explanation about the way in which the changes of the number of municipalities among the different censuses has been managed. Procedure follows that proposed in Beltrán Tapia et al. (2019) for the whole of Spain.

[22] Female literacy rates are not considered because the significant literacy gender gap could complicate the interpretation of our results.

to the one existing at the end of the Old Regime, 381 and 143 municipalities are classified as speaking Catalan or Spanish, respectively (Figure 1). As explained above, these linguistic domains follow a clear geographical divide and Catalan-speaking municipalities are situated along the coast. As shown in Table 1, when we discriminate municipalities by language, a new point of interest emerges. In 1860 the average literacy rate in Catalan municipalities was lower than in Spanish municipalities (13.4 versus 16.6). This difference decreased in 1900, where the difference between Catalan and Spanish municipalities was only 1.6 percentage points (26 versus 27.6). Nevertheless, this scenario changed in 1930, where the Catalan municipalities had, on average, a higher male literacy rate (54 versus 52.4).

The basic descriptive statistics for male literacy are complemented with the kernel densities, which raise some interesting issues (Figure 3). On the one hand, both distributions move to the right and, on the other hand, distributions were left-skewed in 1860 and 1900, but not in 1930. These two facts indicate that there was a generalized increase in male literacy rates of the municipalities of the region of Valencia along the period considered. Concerning the effect of the language, the distribution of Spanish-speaking municipalities was located to the right of Catalan-speaking municipalities in 1860, indicating that the latter had lower literacy level. The situation is not so clear in 1900, where the right tail of the Catalan-speaking municipalities is longer that the one for the Spanish-speaking municipalities; namely, municipalities with highest male literacy rates were Catalan-speaking. It seems that the advantage of the Spanish-speaking municipalities decreases still more in 1930, where a considerable part of the distribution of Spanish-speaking is located to the left.

<Insert Figure 3>

In any case, a precise empirical analysis of the existence of a relationship between language mismatch and literacy rates by municipality needs to include additional elements that generate

variability in the costs and benefits of education – and thus in the observed literacy rates. In this respect, the analysis considers the territorial variability on the institutional characteristics of Valencian municipalities during the Ancien Regime, obtained from the Census of Floridablanca (1787). This source provides information on the size, structure and occupation of the population by settlement. In addition, these settlements are classified by category (city, town, village, hamlet and so on) and jurisdiction. The census classifies the type of jurisdiction of local entities as *royal*, *ecclesiastical lordship*, *secular lordship* and *military orders*. Using the information on the kingdom of Valencia, we have been able to identify the jurisdictional regime of all 524 municipalities. With this information we create the dummy variable *lordship*, which is equal to one when the municipality is classified as lordship and equal to zero when classified as royal. Besides, we have also identified those municipalities inhabited by Moriscos until they were expelled in 1609.[23]

In addition, we have also taken into account the total population as well as a proxy for the settlement pattern in each municipality. This means we can analyse whether the inhabitants were concentrated in the core population entity (towns, villages) or scattered across minor entities (hamlets, homesteads, mills, rural dwellings and isolated buildings). To do this we have created a variable that measures the percentage of population living in the main population centre. In this case the information comes from the *Nomenclator* of Spain for 1887. Besides, since geographical characteristics could also affect literacy levels, the analysis includes controls for both first-nature (temperature, rainfall, altitude and ruggedness) and second-nature geography (distance to roads).[24] Finally, the economic structure of the municipalities could have an effect

[23] A specific analysis of the effects of differences in institutional setting on literacy levels across Valencian municipalities in Beltran Tapia et al. (2020).

[24] We have also constructed additional variables representative of the second-nature geographic characteristics of the municipalities (distance to the coast, distance to the capital city). Since taking these alternative measures into account does not alter the main results of the empirical exercise, they are not described in the text.

on the demand for skilled work, so the share of manufacturers and artisans on total employment in 1787 has also been computed.

<Insert Table 2>

Empirical strategy and results

The descriptive analysis confirms the existence of a language effect but, a priori, it seems in favour of the Catalan-speaking municipalities, at least from 1900 onwards. Nevertheless, other things beyond the language mismatch could be explaining this preliminary result. Then, we propose an econometric analysis that allows us to control for the additional elements that affect literacy rate. The econometric specification for a specific year *t* is the following:

$$male\_literacy_{it} = \alpha_0 + \alpha_1 Catalan_i + \beta X_{it} + \delta Z_i + \lambda_j + u_{it} \quad (1)$$

where the subscript *i=1, …, 524* refers to municipalities, *j=1, …, 17* refers to districts, and *t=1860, 1900, 1930* refers to the year of the census. The dependent or endogenous variable, the male literacy rate, is explained by the language of the municipality ($Catalan_i$), a group of control variables—both time-variant (X_{it}) and time-invariant (Z_i) control variables—, districts fixed effects (λ_i), and an error term (u_{it}).

Our main objective is to analyse the impact generated by the existence of a mismatch between the formative language and the language used by the population; namely, we want to know the male literacy rate of a specific Catalan-speaking municipality if in that municipality the official language would have been Spanish. This hypothetical situation, known as counterfactual, is unknown. In any case, if the distribution of the language were exogenous, the estimated coefficient α_1 would inform about the language effect: a negative (positive) coefficient means

that Catalan-speaking municipalities has, on average, lower (higher) male literacy rates than Spanish-speaking municipalities.

Nevertheless, as it has been shown in Figure 1, the distribution of language was not random since Catalan-speaking municipalities are located in territories near to the coast. In this regard it could be argued that access to the coast enabled greater specialization in commercial activities and therefore foster investments in human capital in Catalan-speaking territories. Alternatively, it could also be hypothesised that in areas specialized in agricultural production, which were more intensive in unskilled work, families faced a greater opportunity cost when it came to sending their children to school. Whatever the direction of these effects, it is possible that Catalan-speaking municipalities showed different educational attainments than those in Spanish-speaking municipalities, regardless of the mismatch between the language of speaking and instruction.

Therefore, as the official language of a municipality was not a random phenomenon, using the ordinary least square estimation method will produce a biased estimated coefficient α_1. In order to overcome this issue, an alternative estimation technique is proposed, the propensity score matching (PSM). This technique, very common to assess the impact of specific programs or public policies, consists of comparing the literacy rate of a treated unit (in this case a Catalan-speaking municipality) with a very similar one untreated unit (in this case, a Spanish-speaking municipality). Averaging all these comparisons provides the average impact, known as the average treated effect of the treated (ATET).

The technique consists of several steps. In the first step it is computed the propensity score or the probability of being treated, namely, to be a Catalan-speaking municipality. The choice of variables to compute this propensity score is one of the more controversial aspects of this technique because the inability to capture well the characteristics that determine whether a unit

is treated generates biased estimators (Heinrich et al. 2010). We assume that all the characteristics in which the treated and control groups differ are observed, which is known as conditional independence assumption or unconfoundedness assumption (see the Appendix A for a brief explanation). As this assumption cannot be tested, the validity of the model relies on the economic theory and previous empirical findings (Cappelli and Vasta 2020).

In our case, the historical evidence shows that the official language of a specific municipality was related to the geographical characteristics of that municipality. Then, we use the geographical variables (both first and second nature)[25] in order to compute the probability of being a Catalan-speaking municipality (Table A1 in the Appendix A provides the estimated results). The other assumption that cannot be violated is the common support or overlap condition, which implies that there must be enough municipalities (Catalan or Spanish speaking) with similar propensity scores. The distribution of propensity scores confirms that there is wide common support (see Figure A1 in the Appendix A), which ensures that each treated municipality has at least one untreated municipality that behaves as counterfactual.

Finally, we match the similar units in the treatment and control group using a matching technique. Firstly, we rely on the Kernel matching, a non-parametric matching estimator that, in order to construct the counterfactual outcome, uses a weighted average of all individuals in the control group. As it uses all the available information, the estimator has a low variance, although the bias could be high. In order to compute the ATET, we take into account, apart from the

[25] The geographical variables are the following: temperature, rainfall, altitude, ruggedness, distance to the coast, distance to the capital city, and distance to main roads.

geographical variables, the other control variables presented in section 3 and considered in equation 1.[26]

Table 3 presents the magnitude and significance of the average treatment effect of the treated for three different years (1860, 1900, and 1930). The results show that, in terms of male literacy rates, the language did not have any significant effect in 1860 or 1900. In 1930, nevertheless, the language mismatch was statistically different at the 10 percent significance level. If Catalan-speaking municipalities were Spanish-speaking municipalities, the male literacy rate would have been 5.9 percentages points higher.

<Insert Table 3>

As far as there is not a best matching procedure (it depends on the data structure), we have also made use of two additional matching techniques (nearest neighbour and stratification) as robustness checks.[27] On the one hand, the nearest neighbour compares each treated municipality with the municipality in the control group that has the more similar propensity score. As not many observations are used (not all the units in the control groups are used), this method computes estimators with low bias, but high variance. On the other hand, the stratification matching consists of dividing the common support into a set of intervals or groups. Then, the technique compares treated and control units within each group and compute an average impact. The optimal number of intervals depend on the data, and it is very important to test that, within

[26] The (non-geographical) control variables are: population, settlement pattern, lordship, morisco, the percentage of manufacturers and artisans, and districts fixed effects.

[27] See Caliendo and Kopeining (2005) for a description of the matching techniques.

each group, the mean propensity score is not different for treated and controls units. In our case, the model has generated eight different groups.[28]

Tables 4 shows the results of these two techniques, being the panel a for the neighbour matching and panel b for the stratification matching. The results are consistent with the previous one: the language mismatch was not relevant neither 1860 nor 1900 but was statistically significant (at the 5 percent level) in 1930.

<Insert Table 4>

Apart from the matching estimator, there are other estimators to compute treatment effects. To complete the robustness analysis, we present the results of the inverse-probability-weighted regression adjustment (IPWRA) estimator. This estimator uses two models, one to predict the treatment status and other to predict outcomes, and one of its advantages is that only one of the models must be correctly specified in order to get a consistent estimator (see the Appendix A for a brief description of this estimator). As before we estimate the treatment model based on the geographical variables (both first- and second- nature) and the outcome model that includes all the control variables. Table 5 shows the results, which confirm the existence of a language mismatch in 1930. Nevertheless, unlike the previous results, with this procedure this mismatch was already statistically significant in 1900.

<Insert Table 5>

The results presented in Tables 3-5 therefore support the existence of a negative effect on the percentage of male literacy in the municipalities treated, which is noted in the case of the use of the IPWRA estimator in 1900 and which is robustly observed in the time cut-off corresponding

[28] We use the Stata command "pscore" (StataCorp 2021b).

to 1930. In this case, the effect is robust regardless of the estimation method used and ranges from 5.9% when the Kernel method is used to determine the matched observations to 9.9% in the case of an IPWRA estimation of the treatment effect. Summing up, the literacy gap between the Valencian municipalities that formed part of the Catalan lisguistic domain and those that had similar attributes but which spoke Spanish grew during our period of study, and especially so after 1900. The language mismatch effect thefore materialized as soon as the state was able to effectively enforced Spanish as the language of instruction in school.[29]

Concluding remarks

Using the case of Valencia, this article confirms the negative the impact on literacy rates caused by the introduction of an education system that generated a linguistic distance between the language of schooling and the language of use. These results support those found by Cinnirella and Schueler (2016) and Jain (2017) regarding the importance of the language mismatch effect on results in education. Nevertheless, we have shown that, this effect is only visible at the end of the analysed period. In this sense, if we consider that the regulations imposing Spanish as the language of instruction in primary schools originated in the Ancien Régime, the absence of effects on educational levels would be an indication of the lack of effectiveness of the regulations. Only in the period that begins with the entry into the twentieth century it is possible to identify the expected effects of the implementation of a nation-building policy such as the one described here.

[29] Jain (2017) provides values for the mismatch effect in Indian districts that would range between 18 and 22%.

This temporal evolution could be related to different aspects that deserve further consideration. On the one hand, the effectiveness of this type of policy is related to the capacity of the state to enforce its compliance. Throughout the text it has been shown that in the Spanish case this registered a key change around 1900, when the state assumed the financing of public primary schools, thus legitimizing the forced compliance with the regulations in spite of the opposition from social segments such as the Church or some local authorities. Our results obtained would support the hypothesis that state capacity acted as a contingent on the effective implementation of nation-building policies. In the Spanish case, this capacity was very limited until the first third of the twentieth century, endorsing a gradual and slow vision of the deployment of the liberal state throughout the nineteenth century and the first third of the twentieth century.

On the other hand, considering that in the Spanish case universal male suffrage was introduced in 1890, the identification of this effect in 1930 could be linked to the growing interest of the state in achieving linguistic unification. In this sense, Alesina and Reich (2013) and Alesina, Giuliano and Reich (2019) put forward a theoretical model from which it is derived that the advance of democracy would drive the use of homogenization policies in the face of the growing risk of social unrest. The results obtained in this paper provide partial evidence in favor of this hypothesis in the context of the incipient advance of democracy in Spain at the turn of the century. Linking both aspects, this result would be in line with the thesis sustained in Alesina, Giuliano and Reich (2019), which would point to the advance of democracy as a relevant element for the understanding of the growth of state capacity, necessary for the effective implementation of public policies of national homogenization.

Finally, the evidence presented here contributes to the knowledge of the determinants of the unequal progress of mass literacy in the Spanish territory. The confluence of different languages and the compulsory use of Spanish as the only language of instruction would have meant a

differential educational cost for those citizens and territories that did not have Spanish as their mother tongue. In this sense, the consideration of the unequal distribution of the cost associated with the implementation of this nation-building policy can contribute to a better understanding of the origins of the economic and territorial inequality that characterizes Spanish society.

References

Alesina, Alberto, and Eliana La Ferrara. 2005. "Ethnic diversity and economic performance", *Journal of Economic Literature* 43, no. 3 (2005): 762-800.

Alesina, Alberto, Baqir, Reza, and William Easterly. "Public goods and ethnic divisions", *Quarterly Journal of Economics* 114, no. 4 (1999): 1243-84.

Alesina, Alberto, Paola Giuliano, and Bryony Reich. "Nation-building and education", Economic Journal 131, Issue 638 (2021): 2273-2303.

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

Angrist, Joshua D., and Victor Lavy. "The effect of a change in language of instruction on the returns to schooling in Morocco", *Journal of Labor Economics*, 15 (1997): S48-S76.

Bandiera, Oriana, Myra Mohnen, Imran Rasul, and Martina Viarengo. "Nation-building through compulsory schooling during the age of mass migration", *Economic Journal* 129, no. 617 (2019): 62-109.

Beltrán Tapia, Francisco J., Alfonso Díez-Minguela, Julio Martinez-Galarraga, and Daniel A. Tirado-Fabregat. *Capital humano y desigualdad territorial. El proceso de alfabetización en los municipios españoles desde la Ley Moyano hasta la Guerra Civil*. Estudios de Historia Económica 74, Madrid: Banco de España, 2019.

Beltrán Tapia, Francisco J., Alfonso Díez-Minguela, Alicia Gómez-Tello, Julio Martinez-Galarraga, and Daniel A. Tirado-Fabregat. "Lordships, state capacity and beyond: literacy rates in mid-nineteenth-century Valencia", European Historical Economics Society Working Paper No 196, Setember 2020.

Bleakley, Hoit, and Aimee Chin. 'Language skills and earnings: evidence from childhood emigrants', *Review of Economics and Statistics*, 86, no. 2 (2004): 267-98.

Caliendo, Marco, and Sabine Kopeining. "Some practical guidance for the implementation of Propensity Score Matching", *Journal of Economic Surveys* 22, no. 1 (2008): 31-72.

Cappelli, Gabriele, and Michelangelo Vasta. "Can school centralization foster human capital accumulation? A quasi-experiment from early twentieth-century Italy," *Economic History Review* 73, no. 1 (2020):159-84.

Casanova, Emili. "La frontera lingüística castellano-catalana en el País Valenciano", *Revista de Filología Románica* 18 (2001): 213-60.

Cinnirella, Francesco, and Ruth Schueler. "Nation building. The role of central spending in education", *Explorations in Economic History* 67 (2018): 18-39.

Cinnirella, Francesco, and Ruth Schueler. 'The cost of decentralization: linguistic polarization and the provision of education', CESifo Working Paper No 5894, 2016

Clots-Figueras, Irma, and Paolo Masella. "Education, language and identity", *Economic Journal* 123, (2013): 332-57.

Cvrcek, Tomas, and Miroslav Zajicek. "School, what is it good for? Useful human capital and the history of public education in Central Europe", NBER Working Paper 19690, Cambridge, MA, 2013.

De Gabriel, Narciso. "La formación del magisterio." In Jean-Louis Guereña, Alejandro Tiana and Julio Ruiz, coords. *Historia de la educación en la España contemporánea: diez años de investigación*. Madrid: Centro de Investigación y Documentación Educativa, 1994: 215-66

Escolano, Agustín. *Historia ilustrada del libro escolar en España: del Antiguo Régimen a la Segunda República*. Madrid: Fundación Germán Sánchez Ruipérez, 1997.

Escolano, Agustín. *Historia de la educación (Edad Contemporánea)*. Madrid: Universidad Nacional de Educación a Distancia, 2012.

Fouret, François, and Jacques Ouzof. *Lire et ècrire: l'alphábetisation des Français de Calvin à Jules Ferry*. Paris: Editions de Minuit, 1980.

Guinot, Enric. *Els fundadors del Regne de València,* vol. I. València: Tres i Quatre, 1999.

Heinrich, Caroline, Alesssandro Maffioli, and Gonzalo Vázquez. 2010. "A Primer Applying Propensity-Score Matching", Inter-American Development Bank Technical Notes 161, 2010.

Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis, A. "Very high resolution interpolated climate surfaces for global land areas", *International Journal of Climatology* 25, no. 15, (2005): 1965-78.

Instituto Nacional de Estadística. *Censo de 1787 "Floridablanca",* tomo VI. Madrid: Instituto Nacional de Estadística, 1991.

Jain, Tarun. "Common tongue: the impact of language on educational outcomes', *Journal of Economic History*, 77, (2017): 473-509.

Kersting, Felix, and Nikolaus Wolf. "On the origins of national identity", Rationality and Competition, Discussion Paper 217, Munich, Germany, 2019.

Lapeyre, Henri. *Géographie de l'Espagne morisque*. Paris: SEVPEN, 1959.

Lapeyre, Henri. *Geografía de la España morisca*. Valencia: Ediciones Alfonso el Magnánimo, 1986.

Livi Bacci, Massimo. "Il Censimento di Floridablanca nel contesto dei censimenti europei", *Genus*, 43, (1987): 137-51.

Lleras-Muney, Adriana, and Allison Shertzer. "An evaluation of the effect of English-only and compulsory schooling laws on immigrants", *American Economic Journal: Economic Policy*, 7(3): 258-290.

Miguel, Edward. "Tribe or nation? Nation building and public goods in Kenya versus Tanzania", *World Politics*, 56, (2004): 327-62.

Miguel, Edward, and May Kay Gugerty, M.K. "Ethnic diversity, social sanctions and public goods in Kenya", *Journal of Public Economics*, 89, no. 11-12 (2005): 2325-68.

Núñez, Clara Eugenia. 1992. *La fuente de la riqueza. Educación y desarrollo económico en la España contemporánea*. Madrid: Alianza, 1992.

Pérez Moreda, Vicente, David S. Reher and Alberto Sanz. *La conquista de la salud. Mortalidad y modernización en la España contemporánea*. Madrid: Marcial Pons, 2015.

Prados de la Escosura, Leandro. *Spanish economic growth, 1850-2015*. Cham: Palgrave- Macmillan, 2017.

Ramachandran, Rajesh. 'Language use in education and human capital formation: Evidence from the Ethiopian educational reform', *World Development*, 98 (2017): 195-213.

Sanchis Guarner, Manuel. 1973. 'La frontera lingüística entre Alicante y Murcia', *Cuadernos de Geografía*, 13 (1973): 15-29.

StataCorp. "Stata treatment-effects reference manual: Potential outcomes/counterfactual outcomes. Release 17". College Station, TX: Stata Press, 2021a.

StataCorp. "Stata: Release 17. Statistical Software". College Station, TX: StataCorp LLC, 2021b.

Seida, Yared. 'Does learning in mother tongue matter? Evidence from a natural experiment in Ethiopia', *Economics of Education Review*, 55, (2016): 21-38.

Swee, Eik Leong. "Together or separate? Post-conflict partition, ethnic homogenization, and the provision of public schooling", *Journal of Public Economics* 64, (2005): 415-24.

Williams, Eddie, and James Cooke, J. "Pathways and labyrinths: language and education in development", *TESOL Quarterly* 36, no. 3, (2002): 297-322.

Tables

Table 1. Main descriptive statistics of male literacy rate (percentages)

Sample	N	mean	Sd	Min	p25	p50	p75	Max
Census 1860								
All municipalities	524	14.3	6.5	0.9	10.1	13.5	17.7	44.0
Catalan-speaking	381	13.4	6.1	0.9	9.4	12.9	16.5	39.7
Spanish-speaking	143	16.6	7.2	3.5	11.9	15.2	20.6	44.0
Census of 1900								
All municipalities	524	26.4	9.2	7.4	19.8	25.5	32.2	68.9
Catalan-speaking	381	26.0	8.9	7.4	19.8	25.4	31.6	68.9
Spanish -speaking	143	27.6	9.7	9.7	20.6	26.4	34.1	55.7
Census of 1930								
All municipalities	524	53.6	9.6	25.2	47.6	54.2	60.2	82.5
Catalan-speaking	381	54.0	9.4	25.2	48.1	54.6	60.2	82.5
Spanish -speaking	143	52.4	10.3	29.3	45.0	52.1	60.8	77.1

Notes: The percentage of male literacy in any municipalities is computed as the percentage of men that know read and write respect to the total male population of each municipality. This mean is an unweighted average.
Sources: Censuses of 1860, 1900, and 1930, *Llei d'ús i ensenyament del valencià* (Generalitat Valenciana, 1983), and own elaboration.

Table 2. Main descriptive statistics of the control variables

Variable	N	Mean	sd	Min	p25	p50	p75	Max
Population 1860	524	2,434	6,795	90	709	1,229	2,386	140,61
% population in the core entity	524	86.03	18.14	4.64	80.68	94.21	98.15	100.00
Lordship	431	13.5	6.0	0.9	9.5	13.1	16.6	44.0
Royal	93	17.8	7.8	5.1	12.0	17.2	21.5	39.7
Morisco	216	11.9	5.4	0.9	7.8	11.6	15.2	32.0
Non-Morisco	308	16.0	6.7	3.4	11.4	14.6	19.9	44.0

Distance to main road (km)	524	17.8	15.8	0.06	4.5	13.6	27.7	69.1
Temperature (degrees)	524	15.7	2	9.7	14.2	16.2	17.3	18.3
Rainfall (mm)	524	470.2	62.9	280.9	439.9	464.4	507.7	627.3
Altitude (m)	524	372.5	316.4	1.2	90	298	624.1	1330.6
Ruggedness	524	91.7	62.5	1.5	34.8	90.6	136.6	289.8
Share of manufacturers and artisans	474	8.5	9.4	0	2.5	6.4	10.5	66.3

Sources: Own elaboration based on the 1787 Census of Population; the 1860 Census of Population; Lapeyre (1959), Lapeyre (1986), *Nomenclator de España*, 1887 and GIS.

Table 3. Average treatment effect of the treated

Matching method: Kernel

Number of treated units: 118, Number of untreated units: 112

Year	Difference (ATET)	Std. Error (Bootstrap)	t-stat
1860	-0.776	3.109	-0.249
1900	-3.638	3.196	-1.139
1930	-5.859	3.862	-1.517

Notes: Common support: [0.00804, 0.9787].

Table 4. Average treatment effect of the treated

Panel a: Matching method: Neighbour

Number of treated units: 118, Number of untreated units: 30

Year	Difference (ATET)	Std. Error (Bootstrap)	t-stat
1860	1.260	3.243	0.389
1900	-4.323	3.826	-1.130
1930	-7.599	4.584	-1.658

Notes: Common support: [0.00804, 0.9787].

Panel b: Matching method: Stratification

Number of treated units: 118, Number of untreated units: 112

Year	Difference (ATET)	Std. Error (Bootstrap)	t-stat
1860	1.452	1.649	0.881
1900	-3.806	2.975	-1.279
1930	-7.615	4.179	-1.822

Notes: Common support: [0.00804, 0.9787]. Number of groups: 8

Table 5. Average treatment effect of the treated

Estimator: IPWRA

Outcome model: linear

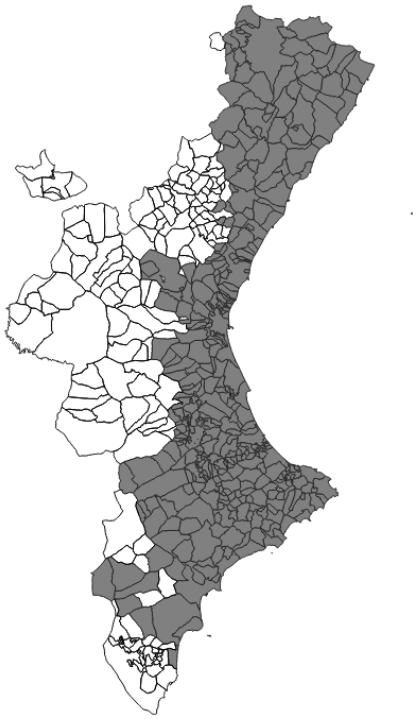
Treatment model: logit

Number of observations: 140

Year	ATET	Robust SE	Z
1860	1.1635	2.4425	0.48
1900	-6.8871	3.2450	-2.12
1930	-9.9130	4.0137	-2.47

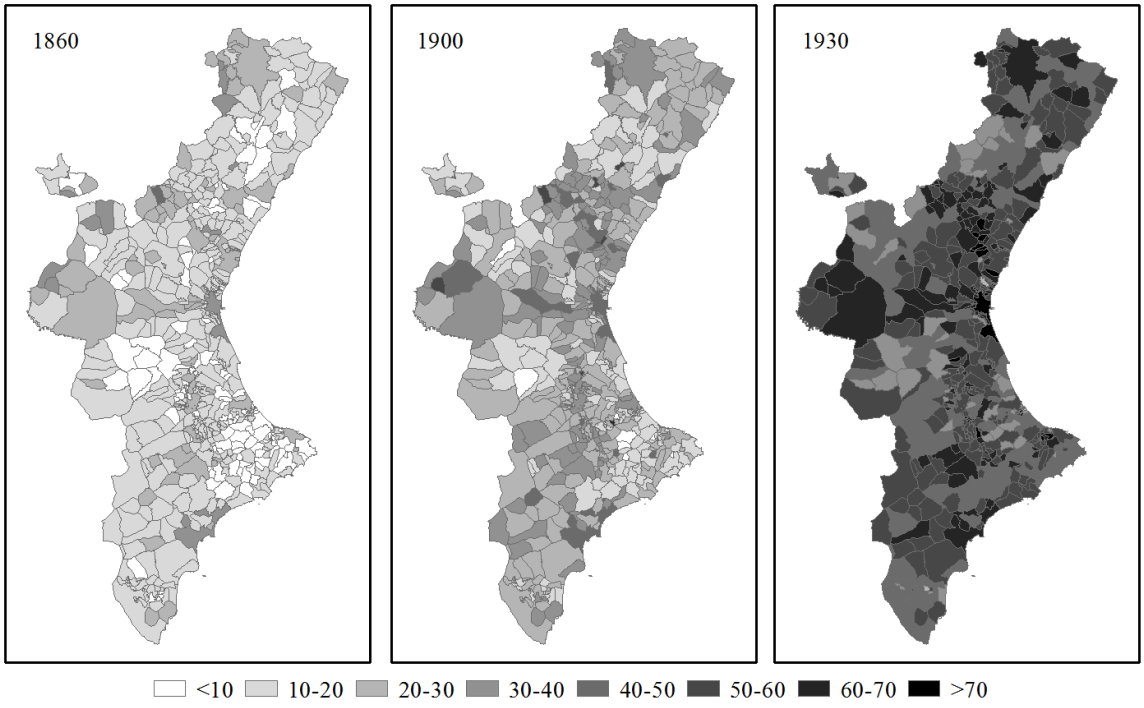
Figures

Figure 1. Valencian municipalities by language



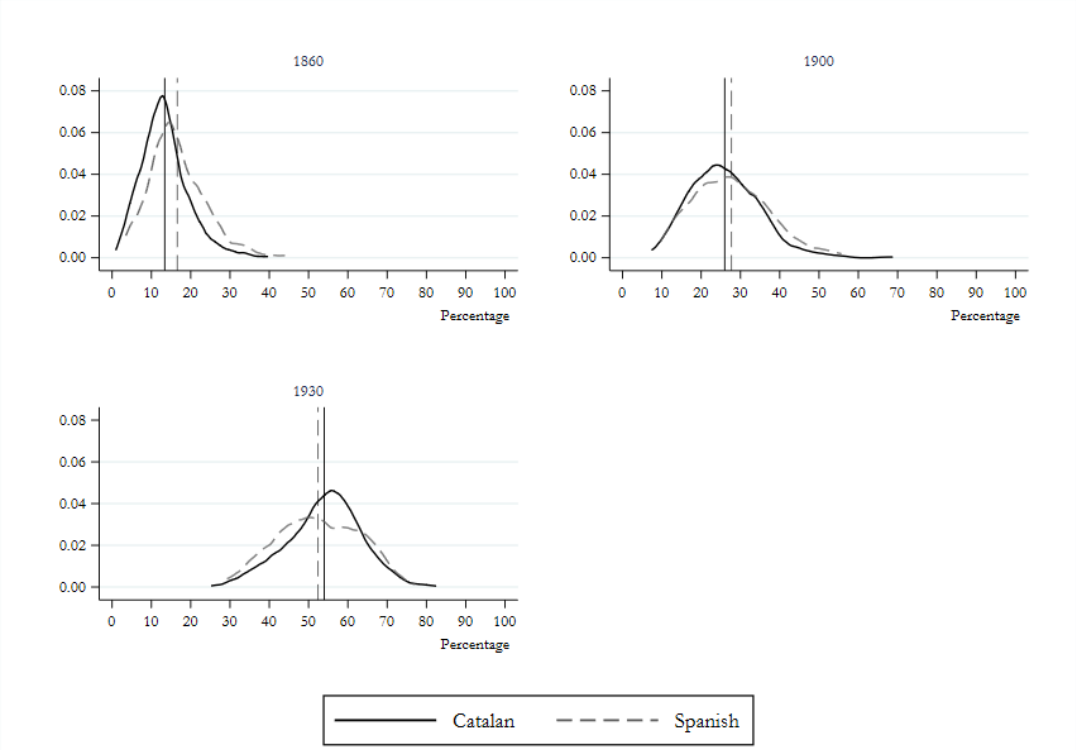
Source: see text.

Figure 2. Male literacy rates in 1860, 1900 and 1930 by municipality (%)



Source: own elaboration based on the 1860, 1900 and 1930 Censuses of Population.

Figure 3. Kernel distribution for male literacy



Sources: Censuses of 1860, 1900, and 1930 and own elaboration.

Appendix A

Conditional Independence Assumption or Unconfoundedness assumption

This appendix provides a brief explanation about the importance of the conditional independence assumption or unconfoundedness assumption. For further details, see Caliendo and Kopeining (2008) and Heinrich et al (2010).

In a random assignment, the treatment status (T) is uncorrelated with any other variable (both observable and unobservable) and, as result, the potential outcomes will be statistically independent of the treatment status. In mathematic terms: $(Y_1, Y_0) \perp T$, where Y_1 is the potential outcome of a treated unit and Y_0 is the potential outcome of an untreated unit.

Without random assignment, treatment may be correlated with factors influencing the potential outcomes. Then, treated and untreated units not only differ in the treatment status but also in other characteristics. If the researchers are not able to control for the characteristics that determine the selection process, the impact of the program will not be correctly estimated.

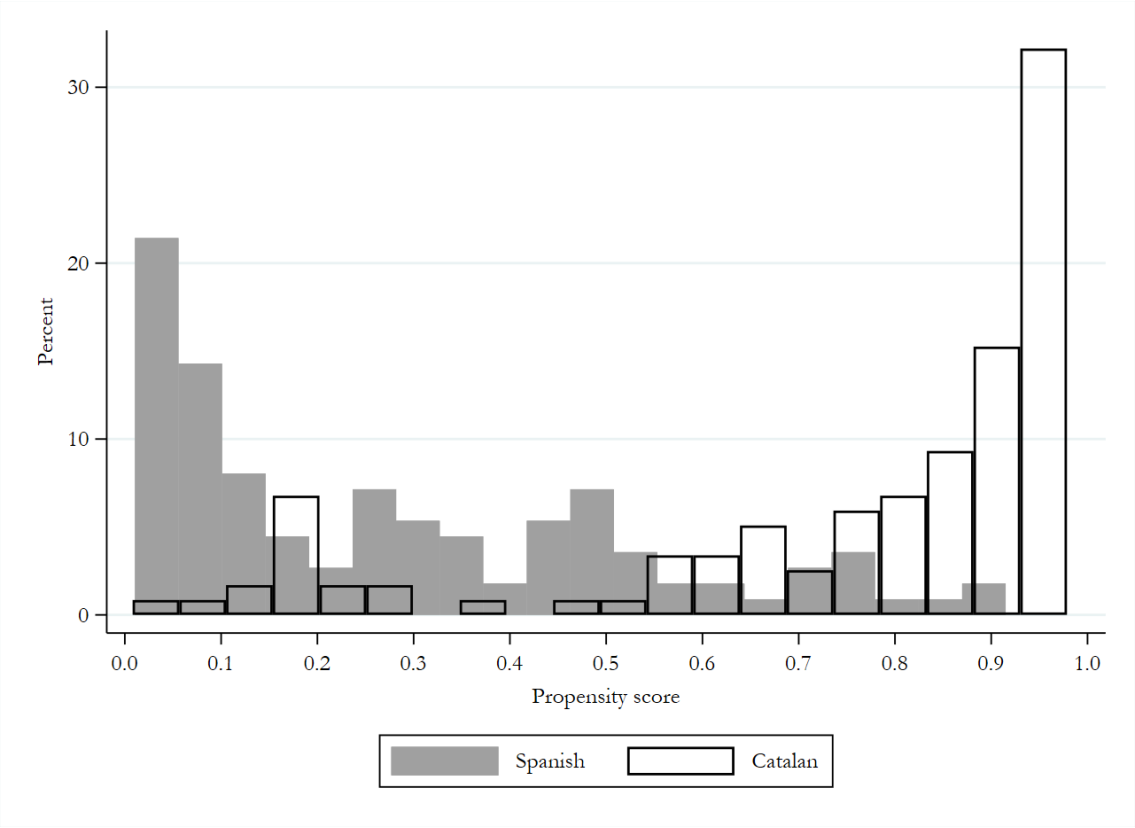
We assume that the characteristics on which the treated and untreated units differ are observable. Given the propensity score (that depends on these observable variables), the potential outcomes are independent of the treatment status, namely the treatment status is as good as random: $(Y_1, Y_0) \perp T|p(X)$.

This assumption allows us to correctly identify the impact of the program because it ensures that the differences between the treated and untreated units are taken into account diminishing, in this way, the selection bias. Then, the untreated observations can be used as counterfactual for the treated observations.

Table A1. Core probit model for estimating propensity scores, Catalan

	Coeff.	Maginal effects
Temperature (degrees)	1.9352***	0.1962***
	(0.3345)	(0.0290)
Rainfall (mm)	0.0421***	0.0043***
	(0.0043)	(0.0002)
Altitude (m)	0.0054***	0.0005***
	(0.0019)	(0.0002)
Ruggedness (u)	-0.0047*	-0.0005*
	(0.0026)	(0.0003)
ln(Distance to the coast) (km)	-0.9634***	-0.0977***
	(0.2900)	(0.0282)
ln(Distance to capital city) (km)	0.2637	0.0267**
	(0.3148)	(0.0319)
ln(distance to main road) (km)	0.1766*	0.0179*
	(0.0925)	(0.0093)
Intercept	-48.0418***	0.1962***
	(7.1136)	(0.0290)
Observations	524	

Figure A1. Visual analysis for common support for estimated propensity scores



Notes: Common support: [0.00804, 0.9787].

Inverse-probability-weighted regression adjustment (IPWRA) estimator

This Appendix provides a brief explanation about the IPWRA estimator. For further details, see StataCorp (2021a).

The IPWRA combines elements of the inverse-probability weighted (IPW) estimator and the regressions adjustment (RA) estimator. The advantage of the IPWRA estimator is that it is more robust to misspecification since only one of the models must be correctly specified in order to obtain a consistent estimator.

One of the most difficult challenges to analyze the impact of treatment is to construct a credible counterfactual, above all when the treatment status is not random and depends on some characteristics that also affect the outcome.

The IPW estimator consider these counterfactuals as missing values and weights the observed units by the inverse of the probability of being in a specific group (treatment or control). The objective is to correct the estimates of the treated and untreated sample means for the missing data (the counterfactuals). For instance, in order to compute the male literacy rate of Catalan-municipalities we had to apply more weight to the Catalan-municipalities further to the coast. On the other hand, the aim of the regression adjustment is to predict potential outcomes after considering the effect of relevant covariates.

The IPWRA estimators uses a three-step approach (although the Stata output only present the final step) to estimate treatment effects:

1) They estimate the treatment model in order to compute the inverse probability weights.

2) For each treatment level, they estimate the weighted regression models of the outcome and, for each subject, obtain the treatment-specific predicted outcomes.

3) They compute the means of the treatment-specific predicted outcomes. The differences of these averages provide the estimates of ATEs (considering both treated and untreated subjects) and ATETs (considering treated subject only).

Appendix B

The sample

Our sample is based on the number of municipalities located in the Valencia region today. According to the latest count, there were a total of 542 in the Comunitat Valenciana. However, there have been various changes and alterations, so we have had to make some adjustments in order to have consistent municipalities over time.

Endogenous variable: literacy (Population Censuses of 1860)

The Population Census of 1860 included a total of 570 municipalities in the Valencia region.[30] This means that we have to convert this historical figure into the current 542 municipalities.[31] To homogenize the data we assigned a current INE code to all the municipalities that appear in the census. First of all, following Goerlich et al. (2006: *Apéndice 2. Alteraciones de los municipios entre los censos de 1900 y 2001*) and using information from the document entitled "*Variaciones de los municipios de España desde 1842*" (Ministerio de Administraciones Públicas, 2008), the 570 municipalities of 1860 were converted into today's 542. These sources enabled us to assign a current INE code to the vast majority of the historical municipalities in the population census, following their historical pathways. However, in some cases complications arose. Most changes in municipalities between these dates are the result of mergers or separations, although the causes vary: towns absorbing neighbouring towns, towns that were merged together, towns that split from each other, etc.[32] In 1860, for example, today's municipality of l'Eliana was part of La Pobla de Vallbona. Hence no information on literacy rates for L'Eliana is available for 1860 (in fact the municipality of l'Eliana separated from la Pobla de Vallbona in 1955). To address this problem, we created a «pseudo-municipality» known as Pobla de Vallbona - l'Eliana. Thus the municipalities included within a pseudo-municipality form a new single entity and have the same literacy rate, or to be more precise, a joint literacy rate. We worked in a similar way with the other

[30] These 570 municipalities were distributed among the three provinces that make up the region as follows: 142 (Alicante), 144 (Castellón) and 284 (Valencia). They were further divided into 46 *partidos judiciales* or judicial districts: 14 (Alicante), 10 (Castellón) and 22 (Valencia). Since there were 4 districts in the city of Valencia, 43 cities and towns were district capitals.

[31] Benicull de Xúquer became the 542nd municipality in the early 2000s, when it split from Polinyà de Xúquer.

[32] There were also numerous changes in the names of the municipalities, with many of them applying to express the name of the municipality in the language of the region.

cases. Overall, due to border changes, there is no information on literacy in the population census of 1860 for 18 out of the 542 municipalities. To solve this problem, 17 artificial pseudo-municipalities (covering 35 municipalities) were created (see Table B1). If the current total number of municipalities in the region of Valencia is 542, once the pseudo-municipalities were created this figure was reduced by 18. Thus our sample consists of 524 municipalities, 17 of which are our artificial pseudo-municipalities.

Table B1. Pseudo-municipalities

INE code	Municipality	Province
3005	**Albatera**	Alicante
3904	San Isidro	Alicante
3014	**Alacant/Alicante**	Alicante
3050	Campello, El	Alicante
3015	**Almoradí**	Alicante
3903	Montesinos, Los	Alicante
3077	**Fondó de las Neus, El**	Alicante
3078	Hondón de los Frailes	Alicante
3093	**Novelda**	Alicante
3114	Romana, La	Alicante
3099	**Orihuela**	Alicante
3902	Pilar de la Horadada	Alicante
3013	Algueña	Alicante
3105	**Pinós, El/Pinoso**	Alicante
12101	San Rafael del Río	Castellón
12121	**Traiguera**	Castellón
12124	Vall d'Alba	Castellón
12128	**Vilafamés**	Castellón
12902	Sant Joan de Moró	Castellón
12135	**Villareal**	Castellón
12901	Alquerías del Niño Perdido	Castellón
12068	Herbés	Castellón
12080	**Morella**	Castellón
46116	Eliana, l'	Valencia
46202	**Pobla de Vallbona, la**	Valencia
46124	Fontanars dels Alforins	Valencia
46184	**Ontinyent**	Valencia
46190	**Paterna**	Valencia
46903	San Antonio de Benagéber	Valencia
46058	Benifairó de les Valls	Valencia
46122	**Faura**	Valencia
46007	**Albal**	Valencia
46065	Beniparrell	Valencia

INE code	Municipality	Province
46197	**Polinyà de Xúquer**	Valencia
46904	Benicull de Xúquer	Valencia

Note: The municipalities in bold are those that existed in 1860.

Census of Floridablanca (Census of 1787)

The Census of Floridablanca contains information on the ancien regime. With data for 1787, it includes a huge amount of information covering the total population by town (or population entity) and by gender, its structure by age groups and its distribution by professions. It also includes information on the administrative characteristics of the population entities, such as category (*ciudad, villa, lugar, aldea,*…), person in authority (*alcalde mayor, alcalde ordinario, gobernador*, …) and jurisdiction (royal or lordship).[33] On the former kingdom of Valencia, it provides information on a total of 550 population entities.[34] These include 9 that today belong to the Comunitat Valenciana but were then part of other administrations.[35]

As regards one of our main explanatory variables – jurisdiction – the census carried no information of any kind for 43 of today's municipalities. These mainly involve two types of cases: a) 15 of them are included in our pseudo-municipalities (see above), so we assume the same jurisdiction as their pseudo-municipality partner/s (see Table B2), and b) for the remaining 28 municipalities the strategy is twofold: in some cases we can identify the municipality they belonged to before becoming independent and then apply the same jurisdiction, otherwise we search for historical information from different sources to discover the type of jurisdiction in the past, assuming they existed in 1787 (see Table B3).[36] Once this is done, and taking into

[33] *Realengo, señorío secular, señorío religioso* and *órdenes militares* are the main types of jurisdiction.

[34] There were 129 in the today's province of Alicante, 141 in Castellón and 280 in Valencia. These towns or entities were in turn grouped into larger administrative entities called *partidos* or districts (13) and *intendencias* or regions (the kingdom of Valencia was an *intendencia)*. Moreover, some entities were «free» or exempted.

[35] Sax and Villena (Murcia); Camporrobles, Caudete de las Fuentes, Fuenterrobles, Requena, Utiel, Venta del Moro and Villagordo de Cabriel (Cuenca) joined Valencia in the mid-nineteenth century.

[36] Torrevieja, which was created from lands belonging to Orihuela, Guardamar, Rojals and Almoradí (all of them *realengos*), is considered to be royal jurisdiction, although before the nineteenth century there was no population, just surveillance towers (and salt mines). In the cases of Geldo and Serra d'En Galceran, for some reason the census does not supply any information about jurisdiction. Based on information obtained from alternative sources we assign to these two towns a lordship jurisdiction.

account the 17 artificial pseudo-municipalities, we have information on type of jurisdiction for all 524 municipalities in the sample.

Table B2. Municipalities without jurisdiction in the Census of Population of 1787 which belong to a pseudo-municipality

INE code	Municipality	Same jurisdiction as pseudo-municipality partner
3013	Algueña	**3105 Pinós, El/Pinoso (3089 Monòver)**
3050	el Campello	3014 Alicante/Alacant
3078	Hondón de los Frailes	**3077 Fondó de les Neus (3019 Asp)**
3114	la Romana	3093 Novelda
3902	Pilar de la Horadada	3099 Orihuela
3903	Los Montesinos	3015 Almoradí
3904	San Isidro	3005 Albatera
12101	San Rafael del Río	12121 Traiguera
12124	Vall d'Alba	12128 Vilafamés
12901	Alquerías del Niño Perdido	12135 Vila-Real
12902	Sant Joan de Moró	12128 Vilafamés
46116	l'Eliana	46202 la Pobla de Vallbona
46124	Fontanars dels Alforins	46184 Ontinyent
46903	San Antonio de Benagéber	46190 Paterna
46904	Benicull de Xúquer	46197 Polinyà de Xúquer

Table B3. Municipalities without jurisdiction in the Census of Population of 1787 which do not belong to a pseudo-municipality

INE code	Municipality	INE code	Municipality
3004	Aigües	12049	Costur
3011	Alfàs del Pi	12069	Higueras
3012	Algorfa	12102	Santa Magdalena de Pulpis
3016	Almudaina	12114	Torás
3051	Campo de Mirra/Camp de Mirra, El	46046	Barx
3052	Cañada	46048	Bellreguard
3062	Daya Vieja	46082	Canet d'En Berenguer
3077	**Fondó de las Neus, El/Hondón de las Nieves**	46087	Casas Altas

INE code	Municipality	INE code	Municipality
3105	**Pinós, El/Pinoso**	46088	Casas Bajas
3120	San Miguel de Salinas	46089	Casinos
3121	Santa Pola	46108	Chera
3122	San Vicente del Raspeig/Sant Vicent del Raspeig	46141	Higueruelas
3133	Torrevieja	46149	Losa del Obispo
3138	Verger, El	46224	Segart

Note: Both El Pinós and El Fondó de les Neus are the main population entities in their respective pseudo-municipalities. However, these two municipalities did not exist back in 1787, since they were created by breaking away from Monòver (1826) and Asp (1839) respectively.

As for the other variables included in our analysis apart from the administrative ones (i.e. total population, population by gender, population by age group and by profession), there are two different situations to be considered. On the one hand, we have the same problem as before given that the census has no information of any kind for the 28 municipalities shown in Table B3. Although we have been able to compile past administrative information for these places, we cannot do this for the demographic variables. In addition, there are 21 municipalities with administrative information (category, authority, jurisdiction, district, regional authority), but no information on population and its structure and professions (see Table B4). This means that with some econometric specifications we are left with a minimum sample of (524-28-21) = 475 municipalities.

Table B4. Municipalities in the Census of 1787 with incomplete information

INE code	Municipality	INE code	Municipality
3064	Dolores	46054	Benetússer
3118	San Fulgencio	46078	Burjassot
12028	Benicasim/Benicàssim	46098	Corbera
12060	Figueroles	46125	Fortaleny
12097	Sacañet	46152	Llocnou de la Corona
12110	Teresa	46186	Paiporta
46002	Ador	46196	Pinet
46013	Alboraia	46197	Polinyà de Xúquer
46014	Albuixech	46223	Sedaví
46022	Alfafar	46237	Tavernes Blanques
46032	Almàssera		

Census of 1887

The census of 1887 contains information not only about the population of the municipalities but also its distribution between the core entity (city, town) and the minor entities (village, hamlet, homestead, mills and isolated buildings). In this case there is information on all 524 municipalities in our sample.

Using data from the *Nomenclator* of Spain for 1887, we created a variable that measures the percentage of population living in the core entity. In the computation of this variable it is important to bear two things in mind.

(i) Pseudo-municipalities. In order to compute this dispersion measure, only the core entity of the main municipality is taken into account (see Table B1).

(ii) Municipalities that were independent in 1887 but today belong to other municipalities (see Table B5). In order to compute the dispersion measure, the core entities of both municipalities in 1887 are taken into account.

Table B5. Completing the information on jurisdiction

Independent municipalities in 1887	**INE code**	**Current municipality**
Adsubia Forna	**3001**	Adsubia
Daya Nueva o Vieja Puebla de Rocamora	**3061**	Daya Nueva
Sela y Mirarrosa Miraflor	**3901**	Poblets, Els
Alcudia de Veo Veo	**12006**	Alcudia de Veo
Montanejos Campos de Arenoso	**12079**	Montanejos
Pobla de Benifassà, La Ballestar Bojar Corachar Fredes	**12093**	Pobla de Benifassà, La
Rosell Bel	**12096**	Rossell
Albaida Aljorf	**46006**	Albaida

Albuixech Mahuella	**46014**	Albuixech
Gandia Beniopa Benipeixcar	**46131**	Gandia
Valencia Benifaraig Borbotó Campanar Masarrochos Villanueva del Grau	**46250**	Valencia

Other relevant information

As well as the information on literacy levels, type of jurisdiction, whether or not the municipality was Morisco and the official language, our data set also includes information on other demographic and geographical variables for the 524 municipalities of Valencia, taken mainly from the population censuses. Table B6 summarizes the sample of the data set.

Table B6. Summary

	Number of municipalities
Dependent variables	
Literacy (1860) Literacy (1900) Literacy (1930)	524
Variables of interest	
Jurisdiction	524
Morisco or non-Morisco	524
Official language	524
Control variables	
Population and professions (1787)	474
Population (1860)	524
Settlement patterns (1887)	524
Geographical information	524