UNIVERSITAT DE BARCELONA

Empirical Essays on Economics and Language

Bernat Mallén Alberdi

2023

PhD in Economics | Bernat Mallén Alberdi

ERSITAT DE
CELONA

UNIVERSITAT DE BARCELONA

PhD in Economics

Empirical Essays on Economics and Language

Bernat Mallén Alberdi

LIBERTAS PERFVNDET
OMNIA LVCE

UNIVE
BARC

PhD in Economics

Thesis title:

Empirical Essays on Economics and Language

PhD candidate:
Bernat Mallén Alberdi

Advisors:
Antonio Di Paolo
Ramon Caminal
Raul Ramos

Date:
June 2023

UNIVERSITAT DE BARCELONA

Acknowledgements

These past four years have been incredibly intense. I met Maria, we got married, and we welcomed our son, Roger. In addition to these life-changing events, I also embarked on a Ph.D. journey.

I am profoundly grateful to my supervisors, Antonio Di Paolo, Ramon Caminal, and Raul Ramos. They provided guidance and support throughout every aspect of my thesis without being overbearing. I truly appreciate the freedom they afforded me in selecting research topics while offering valuable insights and observations. Prior to pursuing my Ph.D., I held the belief that the marginal gain in knowledge diminishes with increased education. However, I can confidently say that this has not been the case in these past years, and I owe this largely to my supervisors. The skills I have acquired extend far beyond economic and econometric knowledge, including coding and computing. From Antonio, I have learned the utmost professionalism required to excel in one's work and to pay attention to the smallest details that make a difference.

I am also grateful for the open-mindedness of my supervisors, who understood my personal circumstances and interests and never pressured me to pursue an academic career. In fact, they actively supported me in seeking job positions aligned with my profile in the public sector. As a result, as I write these lines, I already have a job secured.

I would like to express my gratitude to the AGAUR of the Catalan Government for the FI scholarship, as well as to the Universitat de Barcelona and the Facultat d'Economia i Empresa.

Finally, I cannot conclude this section without thanking my parents for their unwavering moral and financial support during the first year of my Ph.D. and throughout my Master's program in Paris. I would not have reached this point without them. And last but not least, I want to express my heartfelt gratitude to my wife Maria, who has been my rock and constant source of love and support. Her unwavering belief in me and her understanding of the demands of this journey have made all the difference. And to my son Roger, the sleepless nights were greatly compensated by the happiness he brought to our home. Us estimo.

Contents

List of tables

List of Figures

1. Introduction

The presence of linguistic diversity and multilingualism, driven by globalization and the Internet, has profound implications for our daily lives. Even in historically monolingual nations, the exposure to foreign languages has become nearly unavoidable. Institutions of various types have responded to these developments by enacting language regulations, attracting the attention of economists as well. However, there is still much research to be conducted in the field of Language Policy Evaluation and, more broadly, the Economics of Language, to address the challenges posed by these new scenarios.

The Economics of Language is a small strand of the Economics literature that encompasses research focusing on language-related economic issues. It has largely been devoted to the study of the determinants and consequences of language proficiency using the methodology and tools of economics. Notably, economic research has clearly established that language skills matter for economic outcomes. For instance, studies have demonstrated that a shared language facilitates international trade and migration flows (e.g., Frankel and Rose, 2002; Melitz, 2008; Egger and Lassmann, 2015; Aparicio and Kuehn 2016). Furthermore, evidence from a variety of countries indicates that fluency in the host country's language impacts immigrants' performance in the labour market and their economic and social integration (e.g., Bleakley and Chin, 2004 and 2010; Chiswick and Miller, 2007; Lochmann et al, 2019). These findings can be attributed to the communicative use of languages, where languages serve as tools of communication. By minimizing or eliminating communication barriers, which can be viewed as transaction costs or frictions, the market can operate more efficiently, leading to optimal outcomes. Effective communication is indeed crucial for both production and trade.

The utilitarian view on languages as a neutral tool of communication has been challenged by recent research (Caminal & Di Paolo, 2019; Ginsburg & Weber, 2011). Efforts to promote minority languages, the resilience of linguistic groups, and the prevalence of multilingualism in supranational organizations (such as the EU Parliament) all exemplify the existence and significance of the subjective dimension of language, which extends beyond its communicative benefits. Individuals demonstrate through their behavior that they assign value to the goals they pursue, regardless of the reasons behind their attachment to a particular language (thereby undermining the notion of language neutrality).

This increased value attribution influences individuals' willingness to pay when presented with language choices (Gazzola, 2014), particularly in the context of cultural goods. Consequently, the economic value of languages comprises distinct and non-mutually exclusive components that can be categorized into two major classes: use value and non-use value (Grin and Vaillancourt, 1998).

In this dissertation, I study both the use and non-use value of Catalan language in the administrative region of Catalonia, in Spain. Regarding the first dimension, we evaluate an extensive language policy that aimed at enhancing the language knowledge among the adult population, focusing on the Metropolitan Area of Barcelona. Regarding the less studied second dimension, I look the at the non-use value of both Catalan and Spanish for a cultural good in which language plays a significant role: cinema. Specifically, I analyse whether there is a suboptimal provision of films in Catalan language in the movie theaters due to the nature of the industry. Additionally, I study the impact of the emerging paradigm brought about by the expansion of Video-on-Demand (VoD) platforms on language diversity in Catalonia.

There are several reasons why I chose Catalonia as the focus of my case study. Firstly, Catalonia provides ample availability of linguistic data. While multilingualism exists in many regions worldwide, the scarcity of data often hinders economists from conducting comprehensive studies. However, in this case, language knowledge is included in administrative data such as the census, as well as in various surveys. Secondly, because the university in which I am based, the University of Barcelona, and myself are Catalan, which gives me an advantage both to deeply understand the contextual background of my studies and also to obtain additional information on language data. In fact, both the first and second articles of this dissertation utilize new data that is not publicly available. Finally, I do believe that research should serve the taxpayers who fund it by having relevant social impact. The grant that covered my PhD, the FI, is issued by the Catalan government. Hence, this research also aims to add additional knowledge on the evaluation of language policies in Catalonia that can enhance the policy planning and make it more efficient, as well as to contribute to future debates about it, also in other multilingual countries.

This dissertation is organized as follows:

In the first article (chapter 2), together with my advisor Antonio Di Paolo, we investigate the effects of geographical exposure to local language training

centres in a bilingual urban labour market, the Metropolitan Area of Barcelona, exploiting the implementation of a language policy that provided publicly subsidized language courses for adults. Our variable of interest consists in a measure of spatial availability of language schools that captures potential exposure and its expansion over time. Therefore, our approach provides Intention-to-Treat (ITT) estimates, which represent the only policy-relevant parameter we are able to identify given the absence of information on participation in local language courses. First, we focus on the formation of local language skills, adopting a reduced-form approach. Second, we analyse whether accessibility to language centres also affects employment, working hours, employment sector, and occupation. We disentangle the effects by individual and neighbour characteristics to find some heterogeneous effects of the policy.

In the second article (chapter 3), I investigate the effect of competition on language diversity in a cultural market, the movies market, in which language is a relevant characteristic of the good. I analyse the case of the bilingual region of Catalonia to empirically test the effect of competition in two stages of the supply chain – the distribution and the exhibition – on the availability of films in the weaker language. I create a unique data set of all the screenings in the region over 10 months from different sources using advanced web-scraping techniques. The hypothesis to be tested is whether a higher concentration in each of the stages of the supply chain leads to a decrease of availability of films in Catalan language. An additional hypothesis and this paper aims to test is whether the market reacts differently to adults and the children. Thus, I also look for heterogeneous effects disentangling two types of audiences: children-targeted films and adult-targeted films. If children have a higher intensity of the preferences (or their parents, if they are who choose the film), we should observe that the market is more responsive to this type of consumer, that is, that the concentration has greater impact on the availability of Catalan language in children-targeted movies.

In the third and last paper (chapter 4), I investigate the effect of the introduction of VoD platforms on the preferences for different language versions of movies. By using survey data gathered from 2014 to 2019 in the bilingual region of Catalonia (Spain), I explore whether the exposure to VoD affected the likelihood of preferring the Catalan, Spanish or original versions. The hypothesis to be tested is whether people adapt their preferences to what they experience. Hence, the research question is whether the introduction of a new

technology into a market (in the case of this paper, the movie market) can change the preferences of consumers, who accustom, or accommodate, their preferences to the new paradigm.

Finally, in the fifth chapter some concluding remarks are presented.

2. Does Geographical Exposure to Language Learning Centres Matter in a Bilingual City?

2.1. Introduction

This work analyses whether geographical exposure to language learning centres affects the formation of language skills and labour market outcomes in a bilingual urban labour market, the Metropolitan Area of Barcelona. The empirical setup exploits the implementation of a language policy that provided publicly subsidized language courses to adult individuals, in the bilingual region of Catalonia, aimed at fostering proficiency in Catalan, the local language. Existing research in language economics provides compelling evidence in favour of the general idea that language matters, since better language proficiency improves several dimensions of individuals' economic and social wellbeing, similarly to other forms of human capital such as schooling (Chiswick and Miller, 2007). The majority of existing works focus on the behaviour of migrants, based on the underlying hypothesis that language proficiency in the host country language affects the economic and social assimilation of migrants positively (Bleakley and Chin, 2004, 2010). However, investigating the formation of language skills and their relationship with socioeconomic outcomes is not only relevant for the case of migrants learning the language of the host country or for the case of natives who acquire skills in foreign languages. Indeed, studying the social and economic returns to linguistic competences, and most importantly, the role of language policies, is of extreme importance in multilingual societies where different languages coexist and are entitled to an official status. Actually, the issue of linguistic diversity and the vitality and preservation of local or minority languages in multilingual countries or regions is gaining importance over time among policymakers and in society in general. In Europe alone, examples include some of the former soviet republics (like Ukraine, Estonia, and Kazakhstan), Belgium, Cyprus, Wales, the region of South Tyrol in northern Italy, and Spain. In this last country, for example, since the adoption of the Democratic Constitution of 1978, four distinct local languages (Basque, Catalan, Galician, and Aranese, a variety of Occitan) are entitled to co-official status in specific regions, together with Spanish, the official language of the whole country.

Consistently, a growing number of papers point out that skills in local languages are rewarded in the labour market and positively affect other non-monetary

outcomes. Moreover, language policies aimed at promoting the knowledge and use of local languages tend to have longstanding impacts on many dimensions of life. This is relevant not only for foreign migrants but also for internal migrants coming from non-bilingual regions of the same country, as well as for native-born individuals with limited competences in the local language (e.g. in the Spanish case, not all individuals born in bilingual regions are necessarily proficient in the corresponding local language). The case of the bilingual Spanish region of Catalonia, which implemented several language policies to foster the use of its local language (Catalan) after the end of the Franco regime, represents an ideal setting to investigate these issues. Indeed, some works examined the impacts of the language policy implemented in the 1980' by the Catalan Government to stimulate bilingualism and biliteracy among inhabitants of Catalonia regardless of language background and regional origin. The so-called Language Normalization Act (LNA) of 1983 was motivated by the high share of internal migrants from other Spanish regions with no knowledge of Catalan, by the lack of written skills of native Catalan speakers, and by the aspiration to recover and preserve the vitality of the local language. Indeed, the main feature of this reform consists in the introduction of Catalan as a medium of instruction at school together with Spanish (i.e. bilingual education), which has been investigated in recent works. This language policy and the increase in skills it induced (especially) among the non-Catalan speaking population have been found to impact different social and economic outcomes. For example, Cappellari and Di Paolo (2018) show that the exposure to bilingualism at school induced by the reform increased the returns to each additional year of schooling, especially among individuals with non-Catalan origins. Moreover, Caminal and Di Paolo (2019) highlight that the policy-induced increase in oral skills in Catalan among native Spanish speakers affected partnership formation, by raising the propensity to form a linguistically mixed couple (thus, reducing social segregation). Finally, a recent work by Caminal et al. (2021) reveals that the language-in-education reform, by improving skills in the local language among native Spanish speakers, generated intergenerational spillovers by affecting the language spoken with the children.

The findings from this recent research indicate that language policies that promote the formation of local language skills matter in multilingual societies. However, the evidence regarding the impact of local language policies that is available so far only refers to interventions targeted to school-age populations. That is, nothing has been said regarding the effects of the provision of language

training programmes targeted to the adult population in multilingual territories. To fill this gap in the literature, in this paper we investigate the effects of a language policy implemented in Catalonia at the end of the 1980' with the aim of providing publicly subsidized (local) language training programmes. These Catalan language courses were targeted not only to adult migrants coming from other Spanish regions or from foreign countries but also to native-born individuals with limited knowledge of the local languages who were schooled before their introduction as subjects and/or a medium of instruction at school.

This work also contributes to the emerging, although still scarce, literature on the impacts of language training provision programmes, to understand whether these kinds of policies are effective in fostering the linguistic, cultural, and socioeconomic assimilation of newcomers. The few recent works on the effects of participation in language courses have focused on the effects on adult migrants and refugees (Arendt et al., 2021; Åslund and Engdahl, 2018; Foged et al., 2022; Giesecke and Schuss, 2019; Heller and Slungaard Mumma, forthcoming; Kiwi et al., 2020; Lochmann et al., 2019; Lang, 2021; Pont-Grau et al., 2020), which generally report positive results. We are, however, the first to provide evidence regarding large-scale language training programmes in a bilingual setting. Moreover, our analysis is quite relevant for policymaking, since we provide evidence that is useful for determining the spatial distribution of language training centres as well as the extent to which different subgroups of individuals react to an increase in the local supply of language courses.

Specifically, we investigate the effects of geographical accessibility of language learning centres, considering first its impact on the formation of skills in Catalan and, second, whether it affects labour market outcomes in a bilingual urban labour market. We exploit historical administrative information about the geographical location of local language schools since their creation in 1989, focussing on the Metropolitan Area of Barcelona. This is a bilingual urban labour market, where a large majority of language training centres were concentrated during the initial phase of implementation of this language policy. Geolocation information about the location of Catalan language centres is merged with data from the Microcensus[1] of 2001, which contains individual-level information regarding several sociodemographic characteristics, local language skills, and labour market outcomes, plus the census tract identifier of

[1] As explained below, we also aggregate information from other statistics (censuses of 1991 and 2001 and the Population Statistic of 1996) to construct local-level control variables.

the current place of residence (which is the finest geographical unit of analysis for available data).

The empirical analysis is based on a reduced-form approach that aims at estimating the causal effects of geographical exposure to language schools since their creation as well as the expansion of its geographical coverage over time. Specifically, our measure of exposure to language learning opportunities consists in the number of language centres located within a given radius from the centroid of the census tract of residence between during the period 1989–2001. In this way, we estimate the effects of local availability of language schools by comparing skills in Catalan (and labour market outcomes) of individuals residing in neighbourhoods with different degrees of potential accessibility of language learning centres. Therefore, our approach provides Intention-to-Treat (ITT) estimates, which represent the only policy-relevant parameter we are able to identify given the absence of information on participation in local language courses.[2] Indeed, we provide credible evidence that corroborates the causal interpretation of the corresponding OLS coefficient as the causal effect of geographical exposure to local language centres.

We focus our analysis on both individuals born in Catalonia and internal migrants coming from other Spanish regions, who were the most relevant users of such public services during the first decade of implementation of the language policy. Our results indicate that proximity to language training opportunities has a modest, but highly robust, effect on the formation of local language skills, fostering the probability of being able to speak and write in the local language. The results are stable with respect to the inclusion of census-tract level controls as well as district fixed effects. Falsification exercises based either on younger cohorts who were exposed to Catalan at school (and were therefore not a "target" of adults' language schools) or on the random assignment to fake census tracts provide results that speak in favour of the causal interpretation of our main estimations. The effects are not driven by language centres located in specific types of hosting institutions (public schools, community centres, and other municipality buildings) and the spatial effect shows a clear distance-decay pattern. The impacts on language skills are strongly heterogeneous, since the

[2] Two other works on migrants also exploit information on the location of language schools to construct instruments for participation, both considering the German case. Specifically, Lang (2021) uses the number of participants in training programmes for each "job centre" over the number of foreign born in the corresponding area, while Giesecke and Schuss (2019) use the density of available slots in language courses at the county level.

results suggest that only native-born, young individuals with low education levels benefited from geographical accessibility to language training centres. Moreover, the heterogeneous analysis by local characteristics provides some mild evidence suggesting that the impact of exposure appears to be higher for individuals residing in areas with a higher fraction of internal migrants (or a lower fraction of individuals who are proficient in Catalan). Finally, the evidence regarding whether spatial exposure to language schools translated into better labour market outcomes is far from conclusive. In fact, although our results indicate that skills in Catalan are significantly associated with better labour market outcomes (especially employment and occupation), we did not detect any significant effect of exposure to language courses on the labour market outcomes considered in a reduced form framework. This is possibly because the effect of language skills is too modest, in quantitative terms, to induce any direct improvement in labour market performance in the bilingual urban labour market.

The rest of the paper proceeds as follows. In section 2.2, we provide the relevant institutional background, describing the context and implementation of the language policy under investigation. Section 2.3 includes an explanation of the different data sources we use in this work and descriptive statistics. Section 2.4 describes the empirical strategy followed to investigate the effect of exposure to language training centres. Section 2.5 provides the results of our analysis and several robustness checks, and section 2.6 concludes the paper.

2.2. Institutional background

In this paper, we analyse the effect of geographical exposure to (local) language courses on language skills formation and labour market outcomes by exploiting the public provision of Catalan courses in the Spanish region of Catalonia, starting at the end of the 1980'. Catalan belongs to the family of romance languages (together with French, Italian, Occitan, Portuguese, and Spanish) and has been the local language of the Spanish region of Catalonia since the early eleventh century. Starting with the War of Spanish Succession (1701–1714) and the subsequent incorporation of Catalonia within the Spanish Crown, the use of Catalan was progressively limited to domestic use and the language lost much of its social prestige. Although this trend reversed during the second half of the nineteenth century (the so-called "Renaixença"), when Barcelona became one

of the cultural capitals of Europe, the dramatic political events that occurred in Spain during the 1930s represented a major negative shock for the public use and vitality of Catalan. During the Franco dictatorial regime, Catalan was banned in the public milieu, its private use was prosecuted, and Spanish became the only official language. After Franco's death in 1975, the country went through democratic transition. The decentralization process that took place with the Democratic Constitution of 1978 recognized the co-officiality of Spanish and local languages in bilingual regions. Moreover, the new Constitution allowed regional governments to recover and stimulate the public and private use of their own languages. Several policies were implemented in Catalonia, as well as in other bilingual Spanish regions such as the Basque Country, to favour bilingualism of the population and to recover the use and vitality of local languages.

However, the sociodemographic landscape of Catalonia after the Franco regime represented the main challenge in pursuing these aims. Due to mass migration from the Spanish-speaking areas of the country towards Catalonia since the 1950s, a substantial share of Catalan residents (i.e. internal migrants and their offspring) were native Spanish speakers, with limited or no knowledge of Catalan. This was particularly the case in the periphery of the city of Barcelona, where most migrants were located. Instead, in Catalan-speaking families, Catalan represented the native language even for new generations born during the dictatorship. This means that individuals of Catalan origin were fluent, at least orally, in their native language, but writing skills were scarce or even absent for a large fraction of native Catalan speakers.[3] It was against this background of linguistic segmentation that the local government used language policies as the main instrument of "language normalization". The main target was to guarantee bilingualism and biliteracy regardless of language background or regional origins.

Immediately after the 1978 Constitution came into effect, Catalan language became a compulsory subject (for at least three hours per week) in non-tertiary education. Some years later, the Language Normalization Act (LNA)

[3] For example, in 1986, only 31% of the Catalan population was able to write in Catalan, and the share was substantially lower for individuals born in other Spanish regions (7% versus 44% among individuals born in Catalonia). Oral skills were also far from being widespread (64%). Five years later, in 1991, the percentage of individuals who were able to write (speak) in Catalan was 40% (69%), and in 2001 up to 52% (77%) of the population had written (oral) skills in the local language.

implemented by the Catalan Government (*Generalitat*) in 1983 introduced a sharp change by establishing Catalan as a medium of instruction in primary and secondary schools, alongside Spanish, making the education system effectively bilingual. The reform's objective was that, by the end of compulsory school, all pupils must have achieved complete proficiency in the four basic competences (understanding, speaking, reading, and writing) in both Catalan and Spanish. Indeed, this language-in-education policy significantly shaped the language skills of the affected cohorts, especially among native Spanish speakers, and had important implications for labour market and social outcomes (for more details, see Cappellari and Di Paolo, 2018; Caminal and Di Paolo, 2019; and Cappellari et al., 2021).

Thus, a huge effort was devoted to spreading the knowledge of Catalan among new generations. However, a significant portion of the adult population was not fully proficient in Catalan. This group included mainly monolingual Spanish speakers but also Catalan speakers who were not able to write in their native language. In the meanwhile, the regional labour market was characterized by a rising demand for local language skills. This was a result of the increasing use of the local language as a medium of communication in private business, on the one hand, and the establishment of Catalan as the main language in regional and local public institutions after the implementation of the LNA, on the other.[4]

Consequently, the availability of Catalan language courses for adults increased substantially during the 1980s, thanks to small public or even private initiatives in the city of Barcelona and its metropolitan area. However, this modest supply of language courses was clearly insufficient to guarantee equal chances to gain access to language training and to shape local language skills of the adult population. With the aim of stimulating the formation of local language skills among adults and mitigating the insufficient endowment of the workforce with

[4] The same reform also allowed for the creation of a public Catalan TV channel (TV3) and incentivized the presence of newspapers written in Catalan. The demand for Catalan skills in the local labour market was also enhanced by the subsequent Language Policy Law (LPL) of 1998, which affected the relevance of Catalan in the labour market. First, a proficiency of level C in Catalan was set as the prerequisite to enter public sector jobs. It is worth mentioning that this proficiency certificate was automatically awarded at the end of compulsory education to individuals who had entered school since the 1978/79 school year, i.e. those born since 1972. Second, it increased the incentives to foster the use of Catalan in private business, especially among those firms with direct commercial contacts with the Catalan public sector and/or service firms with a strong level of contact with the public (e.g. the restaurant and hotel industry). That is, the LPL introduced the institutional basis for the creation of a bilingual labour market.

this relevant asset, in 1988 the Catalan Government implemented a novel language policy aimed at providing language courses at different levels targeted to the adult population. The policy was implemented through the creation of the Consortium for the Linguistic Normalization (*Consorci per a la Normalització Lingüística*, CLN from now on), which centralized the organization of courses, under the supervision of the General Board of Language Policy (*Direcció General de Política Lingüística*). City councils, provincial deputations, and the Catalan government itself, which guaranteed the funding of the CLN, participated in the consortium. Until 1994, by law, 65% was funded by the *Generalitat* and 35% by other local entities; from 1995 the rule was relaxed and resulted in approximately 67% funding by the Catalan government and 33% by the local entities. To stimulate participation, language courses were generally offered free of charge except for advanced levels, although the price of these courses was highly publicly subsidized.[5]

In this paper, we focus on the first decade of existence of the CLN, covering the academic courses from 1989/90 to 2000/01 (starting in September). This choice is driven not only by the fact that the empirical analysis exploits data from the 2001 Microcensus, containing individual-level data on skills in Catalan, but also by other features that are useful for our identification strategy. A first interesting aspect of the language policy under investigation concerns the creation of a network of language courses. The initial deployment of CLN's Catalan courses was markedly concentrated in the Metropolitan Area of Barcelona. Among the 19 municipalities that participated in the CLN since 1989, 11 belong to the Metropolitan Area. Moreover, another 12 municipalities joined the consortium in 1990, of which seven also form part of this large agglomeration of cities around Barcelona. Therefore, the geographical delimitation of our empirical analysis coincides with the border of the Metropolitan Area of Barcelona, which also simplifies some technical issues related to the use of census tracts as the territorial unit of analysis (see section 4). Second, classes were held in public buildings belonging to public schools, community centres, and other municipality infrastructures, on the basis of agreements between the CLN and the hosting institutions. Therefore, the initial locations of language courses and their spatial distribution in general were generally driven by the availability of suitable infrastructures rather than by

[5] Here we refer to the general courses, open to everybody. In addition to the general courses, the CLN also offered sectorial courses, mainly to civil servants of regional and local institutions, who needed to attain the required level of Catalan language.

potential demand-side factors. This aspect speaks in favour of the exogeneity of the geographical availability of language courses with respect to individuals' residential locations. However, in the empirical analysis we still provide several checks regarding this possible issue as well as regarding the fact that hosting institutions could also provide other services, besides language courses, that might have affected the outcomes we investigate.

Third, until the beginning of the twenty-first century, the large majority of language students were born in Catalonia or in other Spanish regions.[6] This enables restricting the empirical analysis to individuals born in Catalonia or in other Spanish regions, thus limiting the presence of unobserved heterogeneity. Indeed, 10,173 out of 17,092 individuals (59%) who enrolled on the course in 1990/91 were born in the region. The share of foreigners in 1992/93 (the first year with available information) was just 10%. It is important to notice that the distribution across language levels varied substantially by origin: while non-Catalan students generally took beginner courses, Catalan students were mostly enrolled on intermediate and advanced courses. Indeed, among the latter there were, on the one hand, some monolingual Spanish speakers born in Catalonia (i.e. second-generation internal migrants) who had limited or no knowledge of Catalan because they were schooled before the LNA. On the other hand, a relevant share of individuals from Catalan-speaking families, who were already orally proficient, also wanted to improve their written skills in their native language. The proportion of students born in Catalonia did not change much during the decade. Actually, in the last school year we consider in this work (2000/01), students born in Catalonia represented 56% of those enrolled, while those born in other regions of Spain represented 19.78% and the rest were from other countries (7.68% from the European Union and 7.58% from outside it). However, the number of foreign students taking Catalan courses increased sharply afterwards. Indeed, in the academic year 2003/04, around 40% of enrolled students were migrants from other countries (mostly Latin American). The proportion of foreign migrants rose to 76% in 2005/06 and remained stable until today. Coinciding with the large inflow of foreign migrants, the CLN also started a general rationalization and (territorial) reorganization of language courses, increasing the supply in areas characterized by a high share of migrants as well as in other medium and small cities of Catalonia.

[6] Detailed information is contained in the yearly reports of the CLN, which are available at https://www.cpnl.cat/transparencia/memories/.

2.3. Data and Descriptive Statistics

The empirical analysis presented in this paper focuses on the Metropolitan Area of Barcelona, a wide agglomeration of municipalities located around Barcelona, the capital of the Spanish region of Catalonia. It represents the second most populated urban area of Spain and one of the densest agglomerations in Europe and is composed of 35 different municipalities. The Metropolitan Area of Barcelona is interesting for our analysis for several reasons besides its size. First, most of the internal migrants (and subsequently foreign migrants) who moved to Catalonia are located within the borders of the Metropolitan Area, mostly in specific neighbourhoods of the capital and in some surrounding municipalities. Second, it covers virtually all the municipalities that participated in the CLN since its creation in 1989 throughout the period we consider in this paper (academic years 1989/90 to 2000/01).[7] Third, focussing on the metropolitan area simplifies the process of tracking changes in the geographical definition of census tracts that occurred over time, which, as explained later, is fundamental for our work.

We combine several databases to answer our research questions. On the one hand, we exploit individual-level data proceeding from the Spanish Microcensus of 2001, a 5% random sample of the full sample for the year. This dataset is especially useful for our purposes for several reasons. First, it contains information about several sociodemographic characteristics such as gender, year of birth, place of birth, year of arrival in Catalonia for those who were born elsewhere, the municipality and residential dwelling of all individuals, completed education, marital status, and the number of children and adults in the household as well as self-reported knowledge of Catalan. The last variable is crucial in our analysis and is coded as a cumulative combination of different language competences. Specifically, respondents report whether they can 1) understand, 2) read, 3) speak, 4) read and speak, or 5) read, speak, and write in Catalan.[8] The Microcensus also contains information on employment status,

[7] It is worth noting that language courses are scheduled on the basis of the academic year, which means that in practice the starting year was September 1989. Moreover, we exclude from the analysis two small municipalities (Montgat and Tiana) located at the periphery of the metropolitan area, since they belong to a different county of the region and did not participate in the creation of the CLN during the period considered in this paper.

[8] Unfortunately, as already mentioned, there is no information regarding participation in local language courses.

weekly hours of work, and occupation and industry (both at two digits), which we use as labour market outcomes. Second, it refers to a period that enables a focus on the initial phase of deployment of local language training programmes provided by the CLN.[9] Third, it makes it possible to exploit information about the census tract of the place of residence, which corresponds to the geographical dimension considered in this paper (and the most spatially disaggregate unit considering the available data).

On the other hand, we also use aggregate data from two different sources, which are combined with our microdata using the census tract identifiers. One consists in aggregate variables defined at the census tract level, which we use as additional controls in our empirical analysis, to rule out the effect of local confounders that might correlate with the treatment variable of interest. Specifically, the aggregate information we exploit in this paper has been constructed from the Spanish censuses of 1991 and 2001 and the so-called "Population Statistics" of 1996, a census-based database that is available only for the region of Catalonia. We retrieved: 1) the fraction of individuals aged 16 years or over who had completed some post-compulsory education, 2) the fraction of individuals born in other Spanish regions, and 3) the fraction of employed individuals in the active population in the census tract. We also obtained information about the fraction of adults who were able to: 4) at least speak or 5) speak and write in Catalan, but we use these variables only for the heterogeneity analysis (they are excessively correlated with the share of individuals born in the rest of Spain). We generally prefer using the 1991 values of these three variables, since they are less likely to be directly related to the presence of Catalan courses or to any other aggregate endogenous change that might have taken place during the period of existence of the language consortium. However, we also show the results obtained using controls from 1996 and 2001 for robustness. It is worth commenting that, in order to impute aggregate variables from 1991 and 1996 to the Microcensus, we tracked changes in census tracts with respect to the 2001 definition. In the large majority of cases, the census tract's boundaries coincide over these three years (at least within the Metropolitan Area of Barcelona), but on other occasions we found either divisions or combinations of census tracts

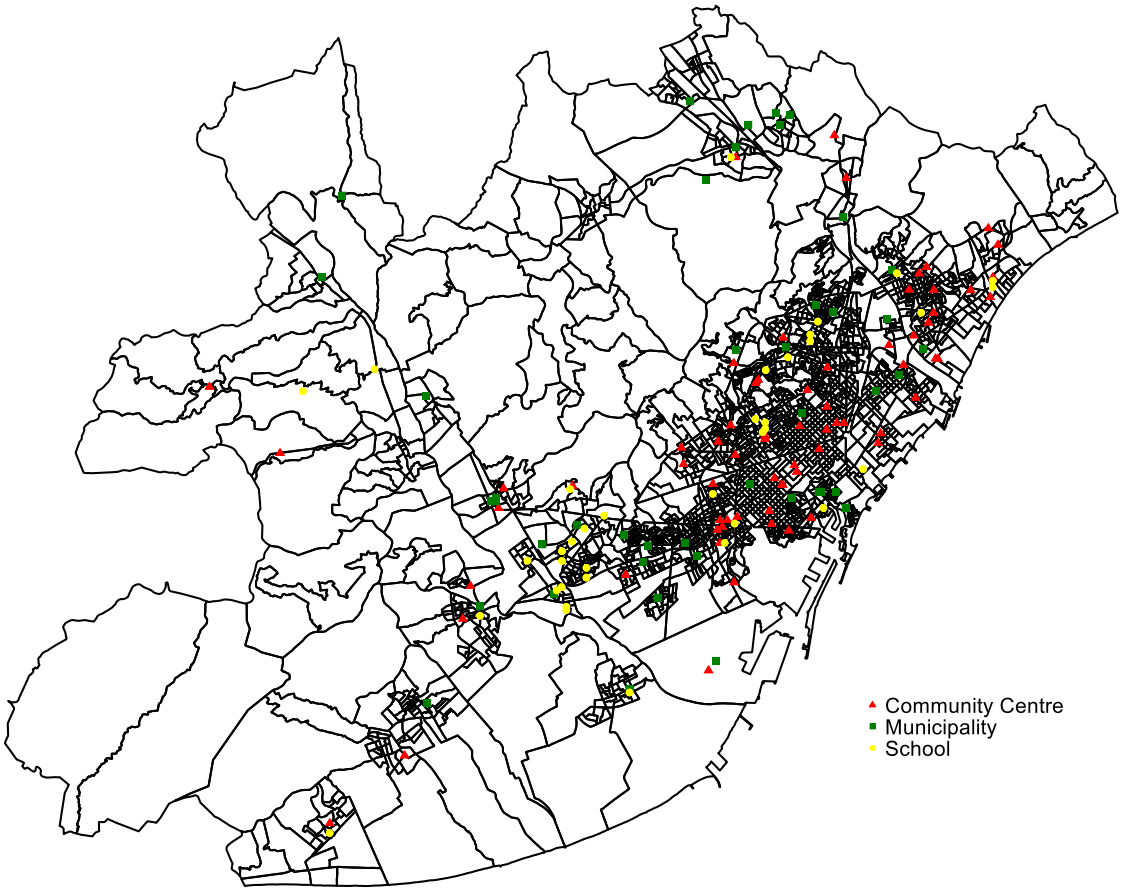
[9] Moreover, drawing on data from 2001 (and previous years) also enables us to circumvent the issues related to the subsequent large waves of arrival of foreign migrants (González and Ortega, 2011), not only from Spanish-speaking Latin American countries. We are therefore able to focus on Spaniards, born either inside or outside Catalonia, who account for the large majority of the population (as well as users of Catalan courses), who are also all proficient in Spanish.

over the years (243 mergers and 74 divisions in 1991, 94 mergers and 28 divisions in 1996). In these cases, we assumed a uniform distribution of the population within each census tract, which seems a reasonable and practical solution given the absence of information about the exact location of residential dwellings inside each area. Finally, we also consider the area (in square metres) of each census tract, the distance to the nearest school, and the distance from each census tract's centroid to Catalunya Square in Barcelona (which represents the "city centre").[10]

The second and most important aggregate data we exploit in this paper consists in historical information about language learning centres, provided by the Department of Language Policy of the Catalan Government. We obtained information about the exact geographical location (address) of each language school for each academic year since 1989/90, which made it possible to gauge the geographical coverage of language centres over the territory and its expansion over time. As we also knew the name of the centres in which the language courses took place, we were also able to classify them according to the public institution that was hosting language courses in a given year (i.e. schools, community centres, or other municipal public buildings). After cleaning the data and tracking changes in street names that occurred over time, we combined this information with the shape file at the census tract level (2001 definition) using GIS techniques. Figure 2.1 displays the geographical location of language centres in the Metropolitan Area of Barcelona, considering the information from the first year of existence of the CLN (we show the same map for 1994/95 and 2000/01 in Figures A1 and A2 of the Appendix).

[10] We decided to control for census tract area instead of population density because the census tracts, whose boundaries are defined for electoral purposes, generally contain the same size of population (between 500 and 2500 individuals with voting rights). Most importantly, the shape file from 1991 is not available and the conversion of the 1991 to 2001 census tracts was done by the Catalan Statistical Institute's technicians, using the original cartography. Therefore, as long as the information about the 1991 census tracts' surfaces is unavailable, we are unable to compute population density for this year. Moreover, we also consider the location of "incumbent" schools, constructed before 1990, in order to take into account the potential effect of school availability, which might have affected the individuals in our estimation sample and would not have been picked up by the fraction of highly educated individuals at the census tract level.

Figure 2.1: Location of language centres by hosting institution, 1989/90



The original individual-level dataset from selected municipalities of the Metropolitan Area of Barcelona contains 102,141 observations of individuals aged between 16 and 64 years. For the purposes of our estimations, we dropped observations for individuals who were still in education (16,837 observations) and those born after 1971 (24,356). The latter restriction is motivated by the fact that these individuals automatically achieved the certification of full proficiency in Catalan (level C1) at the end of compulsory education, as explained in section 2.2. After imposing these restrictions, we also eliminated from our estimation sample foreign migrants (8,096) in order to avoid introducing additional (and unnecessary) unobserved heterogeneity in our model. Moreover, in order to avoid ambiguity in the definition of exposure to language courses, we excluded individuals who moved to their current residential dwelling after 1989 and who were residing in other provinces of Catalonia (1,228). We obtained a final estimation sample of 51,624 observations

that satisfy the above criteria. Table 2.1 reports descriptive statistics of individual and aggregate covariates for the selected sample.

Table 2.1: descriptive statistics (individual and aggregate variables)

Variable	Obs.	Mean	Std. Dev.
skills in Catalan			
not understand Catalan	51,624	0.042	0.201
understand	51,624	0.217	0.412
read	51,624	0.139	0.346
speak	51,624	0.026	0.159
read and speak	51,624	0.236	0.425
read, speak and write	51,624	0.341	0.474
labour market outcomes			
employed (housework = missing)	34,931	0.898	0.303
employed extended (housework = 0)	45,483	0.689	0.463
hours of work	31,359	39.38	9.240
partime	31,359	0.083	0.276
overtime	31,359	0.175	0.380
public sector job (based on industry)	31,359	0.188	0.390
high skilled job	31,338	0.376	0.484
individual characteristics			
male	51,624	0.483	0.500
age	51,624	47.65	10.166
born in Catalonia	51,624	8.926	4.227
years of education (imputed from levels)	51,624	7.490	11.038
age of arrival to Catalonia (0 for natives)	51,624	0.542	0.50
adults in the household	51,624	2.843	1.146
children in the household	51,624	0.451	0.762
single	51,624	0.173	0.378
married	51,624	0.722	0.448
widowed	51,624	0.037	0.190
divorced	51,624	0.039	0.193
separated	51,624	0.029	0.168
local characteristics			
area	51,624	296327	1419234
distance to the nearest public school	51,624	0.202	0.259
distance to Catalunya Square	51,624	6.413	4.278
share of employed 1991	51,624	0.856	0.046
share of highly educated ind. 1991	51,624	0.238	0.150
share of individuals born in Spain 1991	51,624	0.426	0.145
share of employed 1996	51,624	0.780	0.062
share of highly educated ind. 1996	51,624	0.308	0.163
share of individuals born in Spain 1996	51,624	0.386	0.120
share of employed 2001	51,624	0.886	0.035
share of highly educated ind. 2001	51,624	0.348	0.160
share of individuals born in Spain 2001	51,624	0.346	0.111
exposure measures			
language centers within 0.5 km	51,624	0.978	0.866
language centers within 1 km	51,624	3.347	2.097
language centers within 1.5 km	51,624	6.770	3.958

Note: years of schooling = 0 for illiterate individuals, 3 for incomplete primary education, 6 for completed primary education, 8 for compulsory secondary education, 10 for post-compulsory general secondary education, 12 for vocational training, 14 for advances vocational training, 15 for short-term university degree, 17 for university degree and 20 for PhD. Public sector job is equal to one if the individual works in the following sectors: public administration, defence or social security; education; health, veterinary and social services.

As can be appreciated, around 63% of the sample declared that they were able to at least speak and read in Catalan and 37% that they were fully proficient (speaking, reading, and writing). However, another substantial part of the individuals in our data had limited skills in Catalan, especially considering the general tendency to overreport language proficiency. Moreover, the share of individuals born outside Catalonia is significantly high (44%), which is in line with the sociodemographic situation of this Spanish region described in section 3 (see Table A1 in the Appendix for more details about the province of birth). Indeed, the fraction of fully proficient individuals is substantially lower among them (only 14%, versus 54% of individuals born in Catalonia who are able to read, speak, and write in Catalan). We also report descriptive statistics for labour market outcomes, which are obtained considering only the available observations.[11] Specifically, the Microcensus contains information about the employment status of the respondent and, for those who are employed, the hours of work, and occupation and industry (both at two digits). Here we specifically consider, besides employment and working hours, an indicator for having a highly skilled job and another one for working in the public sector, which have been derived from the information about the employment sector.

2.4. Empirical Methodology

This section describes the approach that we follow to analyse the effect of the spatial deployment of language learning centres of the CLN across the Metropolitan Area of Barcelona on individuals' proficiency in Catalan and labour market outcomes. We specifically want to focus on the initial phase of the CLN, when the location of language learning centres across the territory was generally driven by the availability of public infrastructures rather than by the

[11] That is, the original variable capturing the employment status excludes individuals involved in housework, and other labour market outcomes refer only to regularly employed individuals.

concentration of potential students (e.g. internal and foreign migrants), as has happened in more recent years. For this reason, we rely on microdata from the 2001 Microcensus (and not on more recent data such as the 2011 Microcensus), plus aggregate census-tract-level variables from different years (2001, 1996, and 1991), which we use as local controls. We generally rely on reduced-form equations as follows:

$$Y_{ic} = \alpha + \beta Exp_{ic}^{r} + \gamma' X_i + \delta' Z_c + \mu_{yri} + \varepsilon_{ic} \qquad (1)$$

Here, Y_{ic} represents the outcome under investigation of individual i residing in census tract c. We initially focus on local language skills as the main outcome, considering the original ordinal scale from 1 (no knowledge) to 6 (speak, read and write), as well as a dummy that takes the value 1 if the individual i is fully proficient in Catalan. The vectors X_i and Z_c are, respectively, individual[12] and local characteristics. The term μ_{yri} represents fixed effects for the years of residence in the current dwelling from the creation of the CLN in 1989 until the census year, 2001. Exp_{ic}^{r} is our treatment variable of interest, capturing potential exposure to language learning centres. The definition of this last variable is crucial. Our goal is to construct a measure of geographical exposure to language learning opportunities that can be reasonably taken as an Intention-to-Treat (ITT) but only using the available information about 1) the census tract of residence and 2) the year of arrival in the current dwelling, plus 3) the location of language centres. Our focus on ITT effects is also driven by the fact that we do not have information on participation in language courses, which forces us to rely on reduced-form estimates that directly relate accessibility of language courses and the outcomes of interest.

Based on the available information, we define Exp_{ic}^{r} in a way that captures the availability of language learning centres in the neighbourhood of residence over the time span during which each individual could have actually been

[12] The vector of individual controls includes gender, a third-order age polynomial, province of birth fixed effects, a second-order polynomial of age at migration to Catalonia (equal to zero for those born in Catalonia), and years of schooling. We also tried to include possible "bad controls" to capture household composition (number of children and adults in the household and marital status).

exposed to them. We focus on the number of language training centres rather than on the minimum distance from the place of residence for two reasons: first, our data contains information about the census tract of residence, but not the exact location of the dwelling, which means that we are only able to retrieve the centroid of each census tract. Because census tracts' area is significantly larger in less populated areas, working with the minimum distance from the centroid would introduce substantial measurement error for individuals residing in specific neighbourhoods of the metropolitan area. Second, we construct a measure of accessibility based on the number of locations surrounding the place of residence, based on the hypothesis that the higher presence of language learning centres in the neighbourhood would positively affect the propensity to enrol in a language course.

Specifically, with the aim of capturing local accessibility of language centres, we consider buffers of different radii (r) from the centroid of the census tract of residence (specifically 0.5, 1, and 1.5 km[13]). From the geocoded information of language centres, we were able to count the number of sites located in each possible buffer across the Metropolitan Area of Barcelona. Moreover, to consider the fact that each individual could have been exposed during different years since 1989, we sum all the language centres located within each radius over the academic years 1989/90 to 2000/01, to capture potential exposure throughout the period of existence of the CLN.[14] Finally, since not all the individuals in our sample resided in the same place between 1989 and 2000, because some might have moved to their current dwelling during this period, we exploit the information about the year of migration to the place of residence to sum up the number of "nearby" language centres only over the relevant years. This actually introduces some, though limited, degree of individual-level variation in our measure of exposure Exp_{ic}^{r}, which also justifies the inclusion of fixed effects for the years of residence since 1989 in the equation to be estimated.[15] We cluster the standard errors at the census tract level, which is the

[13] We also checked the results obtained with smaller radii such as 0.25 and 0.1 km, but the estimates were quite imprecise and meaningless, due to the very limited amount of variation in the corresponding exposure variables (i.e. there are too few census tracts with at least one language centre within such small radii).

[14] We rescale the sum by 12 to facilitate the interpretation of the corresponding coefficient. We also tried separate use of the number of nearby centres in each year, which provided generally similar but less precise estimates (results are available upon request).

[15] The other side of the coin is that, as explained in section 4, we are forced to exclude the relatively few individuals who a) moved to the current dwelling between 1989 and 2001 and b) came from other locations in Catalonia, since they could have been exposed to Catalan

main source of variation in our regressor of interest (using two-way clusters by census tract and years of residence provides similar results).

2.4.1. Threats to identification and corresponding checks

The extent to which the estimate of the β parameter from Equation (1) can be plausibly interpreted as an ITT effect depends on whether the exposure variable we adopt in this paper can be considered "as good as randomly assigned". However, at least two general issues could undermine the exogeneity of Exp_{ic}^{r}, representing potential threats to the identification of the causal effect of interest. First, the variable we use to capture local exposure to language centres could pick up the effect of local factors that might be related to the spatial distribution of language centres as well as to the outcomes of interest. Second, residential decisions of individuals and families might not be completely exogenous with respect to the location of language centres or to other (directly or indirectly related) local characteristics. Given the absence of randomization or any exogenous shock in the availability of language learning centres, we carry out several checks to provide evidence that corroborates the causal interpretation of our findings. As mentioned in section 3, an interesting feature of the initial implementation of the CLN is that the placement of local language courses by public authorities that participated in the Consortium was mostly driven by the availability of public infrastructures. This somehow limits the degree of unobserved heterogeneity but does not exclude the possible existence of local confounders, which we try to rule out in different ways.

Therefore, we first proceed with the inclusion of a vector of local controls Z_c, containing the area of each census tract and the distance to the centre of Barcelona (Plaça Catalunya), to rule out the effect of residing in more isolated areas in the periphery of the metropolitan area, and the distance to the nearest public (or charter) school. This last control is especially important, because schools host a large fraction of language courses and, although we only consider adult individuals not in education, school availability in the neighbourhood of residence could correlate with their outcomes through a variety of channels.[16]

elsewhere. Therefore, we only retain individuals who migrated between 1989 and 2001 and who came from other Spanish regions.

[16] Moreover, the presence of schools could endogenously attract certain types of families who decide where to live based on their children's education opportunities (among other things), leading to endogenous residential sorting. We also carry out additional robustness checks to discard this potential issue.

We also control for other socioeconomic and demographic features of the census tracts that might be correlated with the location of language centres, specifically the local employment share (of the active labour force), the share of individuals over 15 with post-compulsory education, and the share of adult individuals born in other Spanish regions. Controlling for these aggregate variables is relevant, since they are likely to capture a substantial part of local unobservables from Equation (1) but, to some extent, they could have been affected by the location of language centres. The ideal setting would use pre-determined values (i.e. before the creation of the CLN), but such information is not available at the census tract level for periods before 1989. Therefore, we generally exploit the first available information, namely from the 1991 Census, but we also try to control for the value of these three local characteristics observed in 1996 and 2001 (separately and all together), to check for the stability of the coefficient of interest after partialling out the effect of neighbourhood composition. In addition, we saturate our model with district[17] fixed effects (π_d), which makes it possible to partial out time-invariant unobserved characteristics that vary at a slightly more aggregated level than our variable of interest (and thus exploit within-level variation in exposure), that is:

$$Y_{ic} = \alpha + \beta Exp_{ic}^{r} + \gamma' X_i + \delta' Z_c + \mu_{yri} + \pi_d + \varepsilon_{ic} \qquad (2)$$

Second, using our preferred specification, which includes district fixed effects [i.e. Equation (2) including local controls from 1991], we also carry out two different falsification exercises. On the one hand, we re-estimate the same equation using an additional sample of "young" individuals, born between 1972 and 1985 (who satisfy the other conditions imposed for the main estimation sample). This is because they should not be directly affected by the geographical availability of language courses since, as explained above, they were exposed to Catalan at school and, most importantly, they automatically received the proficiency certification required to enter (local and regional) public sector jobs. Therefore, they should not have any direct incentive to enrol on Catalan courses, except for the very unusual cases of intrinsically motivated individuals who want to achieve the higher qualification (level D) that is required to be a

[17] Notice that this corresponds to a municipality fixed effect when it has a single district.

language teacher. Moreover, we also perform a permutation test in which we randomly assign census tracts (and the corresponding exposure measure) to individuals and re-estimate Equation (2) using the placebo exposure measures. We consider two different variants of the same experiment: 1) using a uniform distribution of observation for each census tract and 2) maintaining the original proportion of observations within each census tract, and we run 10000 simulations, to obtain the corresponding distribution(s) of the fake β coefficients.

Third, we estimate Equation (2) considering alternative definitions of Exp_{ic}^{r}, based on different radii from the census tracts' centroids to calculate the number of surrounding language learning centres. Specifically, we start using a radius of 0.5 km ($Exp_{ic}^{0.5}$), which ends up being the preferred specification of the exposure variable, and we progressively increase the buffer radius to 1 km (Exp_{ic}^{1}) and 1.5 km ($Exp_{ic}^{1.5}$), which also allows checking for distance-decay effects of exposure to language learning opportunities. Moreover, we present the results from a variant of Equation (2) that includes the number of language centres (summed over time) located between 0 and 0.5 km ($Exp_{ic}^{0.5}$), between 0.51 and 1 km $Exp_{ic}^{0.5-1}$, and between 1.1 and 1.5 km ($Exp_{ic}^{1-1.5}$), which corresponds to the following equation:

$$Y_{ic} = \alpha + \beta_1 Exp_{ic}^{0.5} + \beta_2 Exp_{ic}^{0.5-1} + \beta_3 Exp_{ic}^{1-1.5} + \gamma' X_i + \delta' Z_c + \mu_{yri} + \pi_d + \varepsilon_{ic} \quad (3)$$

Fourth, we also deal with other potential confounding effects, which could stem from other services offered at the same institutions hosting language courses. As explained in section 2, Catalan courses of the CLN took place in different types of public infrastructures, which can be generally classified as public schools, community centres, and other municipal infrastructures. The issue comes from the fact that the last two types of hosting institutions were also likely to offer other courses, workshops, and training aimed at fostering different types of skills and employability of residents (especially unemployed and disadvantaged individuals). Moreover, other social activities that take place in community centres might also have a direct effect on the propensity to learn

and know Catalan, since they could foster community ties and social interactions in the neighbourhood (Domínguez and Montolio, 2021). Finally, school buildings were not used for language courses during the morning but only during the afternoon, because they were busy with regular educational activities during the morning. Therefore, there could be some difference in the composition of language students by type of hosting institution (for example, by employment status). This means that accessibility of language centres could be capturing not only the impacts of shaping language learning opportunities but also the direct and indirect effects induced by other activities offered by community centres and by the municipalities that participate in the CLN. For this reason, using the preferred specification from the previous check, we split the exposure variable according to the type of public infrastructure where language courses were held. Specifically, we separately estimate the effect of exposure to language courses held in public schools (Exp_{ic}^{sch}), community centres (Exp_{ic}^{cc}), and other municipality buildings (Exp_{ic}^{mun}), including these variables in the model one by one and all together, as in Equation (4):

$$Y_{ic} = \alpha + \beta_1 Exp_{ic}^{sch} + \beta_2 Exp_{ic}^{cc} + \beta_3 Exp_{ic}^{mun} + \gamma' X_i + \delta' Z_c + \mu_{yri} + \pi_d + \varepsilon_{ic} \quad (4)$$

Given that public schools do not provide additional services targeted to adult individuals besides Catalan courses (and considering that we retain only adults not involved in education[18]), a finding that the total effect of exposure is mostly driven by the number of language centres located in the two other types of institutions would be against the causal interpretation of our main estimates.

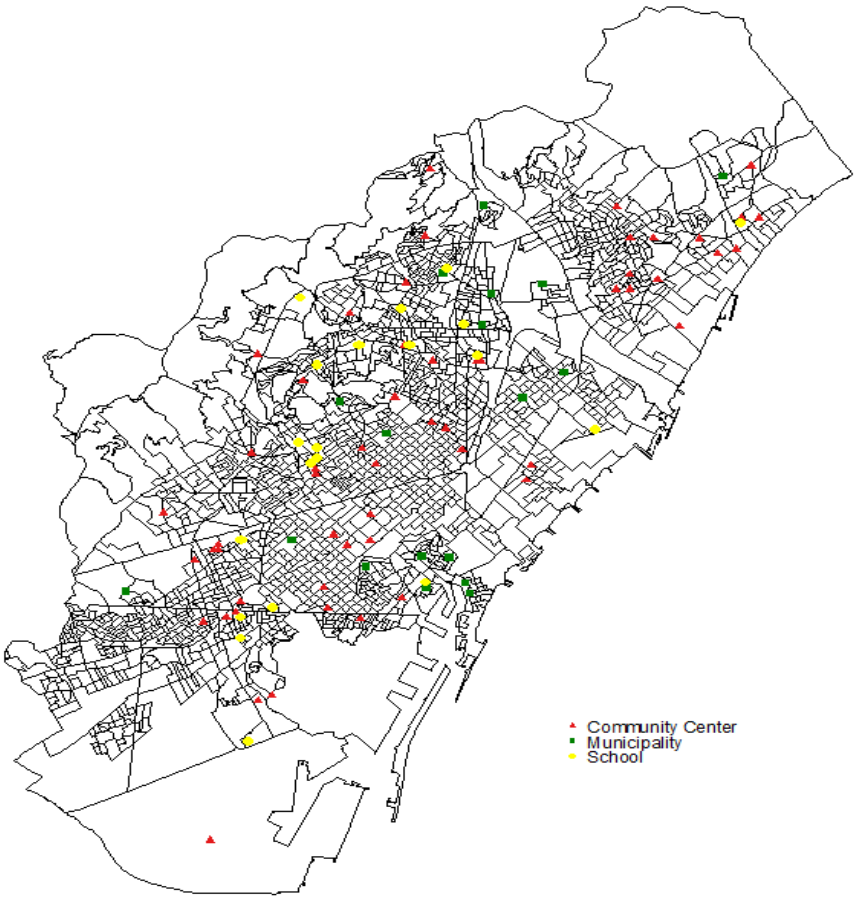
Fifth, we specifically deal with the issue of endogenous residential sorting. As we mentioned before, the location of language centres in the initial phase of the CLN was generally driven by the availability of public infrastructures (and the agreement between the corresponding institutions and the CLN). However, this does not prevent the possibility that residential decisions of individuals and families are in part based on observed and unobserved characteristics of the

[18] Notice that, as we also control for the distance to the nearest school (including adult's school), this mitigates the possibility of capturing the effect of some sorts of adult education programmes that take place in the schools hosting language courses.

neighbourhood. Although we already control for the distance to the nearest school and for several other local characteristics, besides the fact that the likelihood that families decide where to live on the basis of the availability of Catalan learning centres is actually very limited, we still consider that endogenous residential sorting could be somehow biasing the estimate of the parameter of interest. Therefore, we proceed by re-estimating Equation (2), our preferred specification (using a radius of 0.5 km) after excluding individuals who moved to the current dwelling after the creation of the CLN in 1989 (as well as those who moved after 1985 for completeness). Moreover, due to the ambiguity in the definition of exposure for individuals who arrived in the census year (2001) and considering that the interviews were done in November, we try to drop individuals who arrived at their place of residence during the same year.

Finally, our last robustness check consists in re-estimating the main equation only considering observations from "dense areas", which are the city of Barcelona and large surrounding municipalities of the metropolitan area (see Figure 2.2). This check is important not only because of the unevenness of census tracts in less dense areas of the Metropolitan Area of Barcelona in terms of surface and population size (which in turn would affect the chances of social interaction), but also because there could be spatial spillovers from municipalities outside the metropolitan area. That is, it is possible that some individuals residing in the periphery of the metropolitan area could also be affected by the presence of language centres in other municipalities located outside of its limits.

Figure 2.2: location of language centres by hosting institution, core of the metropolitan area (Barcelona and large and dense surrounding municipalities), 1989/1990



After presenting the evidence from all these sensitivity checks, which favour the interpretation of our exposure variable as "conditionally exogenous", we provide additional results regarding the presence of heterogeneous effects of exposure. Specifically, we consider possible heterogeneous effects with respect to several individual characteristics such as gender, age, education, place of birth, and age at migration, as well as according to different local characteristics (mostly the share of internal migrants and the share of individuals who are proficient in Catalan).

Finally, we also analyse the relevance of skills in Catalan and the potential effect of exposure to language training courses on labour market outcomes. Specifically, exploiting the information in the Microcensus, we first show the

conditional correlation between the variable capturing skills in Catalan and the probability of being employed (considering also labour market participation of individuals involved in housework tasks), weekly hours of work,the probability of working part-time or overtime, as well as the probability of working in the public sector (inferred from industry) and the chances of having a highly skilled job (inferred from occupation codes) among the employed. Second, we estimate the reduced-form equation that directly relates the exposure measure to the above-mentioned labour market outcomes.

2.5. Results

Our first set of results are obtained by estimating Equation (1) by OLS, whose estimates are reported in columns (1) to (6) of Table 2.2. We start by including only a parsimonious set of individual controls, namely gender, age (third-order polynomial), years of education, and age at migration (second-order polynomial, equal to 0 for natives), plus fixed effects for the province of birth and years of stay in the current dwelling (since 1989). The coefficient of interest (β), reflecting the impact of potential exposure to local language schools, indicates that each additional centre (per year) in a radius of 0.5 m from the centroid of the census tract of residence increases skills in Catalan by 0.046 points on the 1 to 6 ordinal scale of the dependent variable (mean 4.2, s.d. 1.71). This corresponds to approximatively 2.7% of the standard deviation of the language skills variable. In column (2) we include a first set of census tract characteristics in the vector Z_c (area, distance to the most central square of Barcelona, and distance to the closest public school), which produces a certain reduction in the estimate of interest. This is possibly due to the fact that in the previous specification we were picking up some local confounders associated with the location of different census tracts within the Metropolitan Area of Barcelona. The coefficient of potential exposure obtained from this specification is 0.026, around 1.5% of the standard deviation of the outcome. In columns (3) to (6), we augment the vector of local controls by adding time-varying characteristics of the census tract, namely the share of employed individuals, the fraction of adult individuals born in other Spanish regions, and the percentage of individuals aged above 15 years with post-compulsory education. As mentioned above, we use data from 1991, 1996, and 2001, separately and jointly [column (6)].

Table 2.2: OLS, Dependent Variable = self-reported knowledge of Catalan

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
constant	11.066***	11.455***	9.148***	10.497***	9.611***	8.610***	10.174***	9.795***
	(0.710)	(0.709)	(0.748)	(0.728)	(0.805)	(0.786)	(0.758)	(0.800)
exposure 0.5 km	0.046***	0.026**	0.028***	0.030***	0.036***	0.027***	0.030***	0.030***
	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)
male	-0.085***	-0.079***	-0.063***	-0.062***	-0.063***	-0.062***	-0.063***	-0.056***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
age	-0.505***	-0.513***	-0.512***	-0.525***	-0.518***	-0.514***	-0.514***	-0.481***
	(0.047)	(0.047)	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)	(0.049)
age2	0.011***	0.011***	0.010***	0.011***	0.011***	0.010***	0.011***	0.010***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
age3	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
years education	0.134***	0.129***	0.114***	0.113***	0.113***	0.113***	0.114***	0.114***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
age of arrival to Catalonia	-0.063***	-0.062***	-0.060***	-0.059***	-0.059***	-0.060***	-0.059***	-0.060***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
age of arrival to Catalonia 2	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
area		0.000***	0.000	0.000*	0.000***	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
distance to the nearest public school		0.084	0.052	0.078	0.012	0.055	0.024	0.019
		(0.068)	(0.051)	(0.048)	(0.054)	(0.047)	(0.049)	(0.049)
distance to Catalunya Square		-0.027***	-0.007***	-0.018***	-0.012***	-0.015***	-0.041***	-0.042***
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.011)	(0.011)
share of employed 1991			4.081***			2.170***	2.707***	2.651***
			(0.313)			(0.308)	(0.350)	(0.347)
share of highly educated ind. 1991			-0.965***			-0.950***	-0.254*	-0.238*
			(0.108)			(0.211)	(0.140)	(0.140)
share of individuals born in Spain 1991			-1.733***			-0.698***	-1.095***	-1.108***
			(0.098)			(0.208)	(0.120)	(0.119)
share of employed 1996				3.305***		1.654***		
				(0.266)		(0.273)		
share of highly educated ind. 1996				-1.046***		-0.598**		
				(0.110)		(0.282)		
share of individuals born in Spain 1996				-2.013***		-1.070***		
				(0.117)		(0.414)		
share of employed 2001					3.317***	1.370***		
					(0.433)	(0.379)		
share of highly educated ind. 2001					-0.217**	0.241		
					(0.104)	(0.234)		
share of individuals born in Spain 2001					-1.918***	-0.339		
					(0.130)	(0.376)		
adults in the household								-0.009
								(0.007)
children in the household								-0.049***
								(0.011)
single								ref. cat.
married								0.113***
								(0.020)
widowed								0.090**
								(0.035)
divorced								0.171***
								(0.034)
separated								0.140***
								(0.038)
province of birth fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
years of residence fixed effects (1989-2001)	yes	yes	yes	yes	yes	yes	yes	yes
municipality-district fixed effects	no	no	no	no	no	no	yes	yes
adjusted R-squared	0.428	0.431	0.448	0.449	0.446	0.451	0.454	0.454
number of observations	51624	51624	51624	51624	51624	51624	51624	51624

Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

The inclusion of these local controls slightly increases the estimate corresponding to exposure to local language courses, especially when using more recent census tract level variables, but the main conclusion remains the same (i.e. a modest but positive and significant effect of geographical exposure to language centres). However, the same local characteristics could themselves have been affected, directly or indirectly, by the deployment of Catalan courses after the creation of the CLN and could thus represent bad controls. The ideal setting would have used "pre-determined variables", observed before 1989, but such information is unavailable. However, we are much more confident about the exogeneity of local controls measures in 1991, just two years after the policy implementation, and therefore we retain this set of variables to control for neighbourhood composition. Subsequently, in column (7) we show the results obtained including district fixed effects, which represents our preferred specification from both a conceptual and a statistical point of view.[19] The coefficient of exposure obtained from this specification is 0.030 (s.e. 0.010), which is equivalent to 1.7% of the variation in self-reported language skills.

In addition to the main evidence, Table 2.2 also shows other interesting effects that are worth commenting on. At the individual level, males are less likely to be proficient than females and schooling is positively associated with language proficiency, as expected. This relationship does not capture exposure to Catalan at school, since the sample only includes individuals who received compulsory schooling in Spanish only. Moreover, still at the individual level, the age of arrival in Catalonia has a U-shaped and significant effect: the older the individual was, the lower the propensity to learn Catalan, although the marginal effect is decreasing. As for census tract level variables, the share of employed individuals is positively correlated with language skills, while the fraction of individuals born in other Spanish regions has the expected negative sign, reflecting the effect of living in Spanish-speaking enclaves of the Metropolitan Area of Barcelona. The

[19] In column (8), we also show the results obtained after including the number of children and adults in the household and marital status as additional controls, which provides virtually the same results. However, we decided to retain the results displayed in column (7) since these additional variables could represent bad control, because they might be directly or indirectly related to the presence of language schools and skills in Catalan.

coefficient of the variable capturing the share of highly educated individuals is unexpectedly negative, but it turns out to be positive without controlling for the composition of the neighbourhood in terms of province of birth.

In Table 2.3, we complement the previous analysis by displaying the average marginal effects obtained from an Ordered Probit, using the preferred specification [as in column (7) of Table 2.2], which facilitates the quantitative interpretation of the results. The geographical accessibility to the CLN courses has an effect that is quantitatively modest, but strongly robust, on the probability of being able to speak and write in Catalan. An increment of one school per year within a radius of 0.5 km from the place of residence increases the probability of being able to speak and write in Catalan (i.e. being fully proficient) by 0.8 percentage point (with the unconditional probability being equal to 0.34). Similar evidence is obtained by estimating a Linear Probability Model, in which the dependent variable is the dummy for being able to speak and write in Catalan (see Table 2.A2 of the Appendix), although the point estimate is slightly higher (1.1 percentage point).

Table 2.3: Average Marginal Effects from Ordered Probit Estimates

	ΔPr(Cat = 1)	ΔPr(Cat = 2)	ΔPr(Cat = 3)	ΔPr(Cat = 4)	ΔPr(Cat = 5)	ΔPr(Cat = 6)
exposure 0.5 km	-0.002***	-0.005***	-0.001***	-0.000***	0.000***	0.008***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.002)

Dependent variable: self-reported knowledge of Catalan (1 = not understand Catalan, 2 = understand, 3 = read, 4 = speak, 5 = read and speak read, 6 = speak and write). Control variables as in column (7) of table 2. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

In order to check that our results are not driven by other unobserved factors, we present the results obtained from two different falsification exercises. First, we re-estimate our model following the same specifications displayed in Table 2.2 but using an additional sample of "young" individuals born after 1971 (and thus initially excluded from our estimation sample; see Table 2.A3 in the Appendix for descriptive statistics). As mentioned above, these individuals were exposed to Catalan as a medium of instruction at school during compulsory education and automatically received the proficiency certificate. Therefore, they

are not expected to be users of the language courses provided by the CLN and, indeed, finding an effect of geographical exposure among them would indicate that our main estimates are likely to reflect some spurious correlation rather than a real ITT effect. Reassuringly, the results reported in Table 2.4 indicate that the point estimate of the placebo coefficient is generally very close to zero and not statistically significant in any specification, which speaks in favour of the validity of our approach.

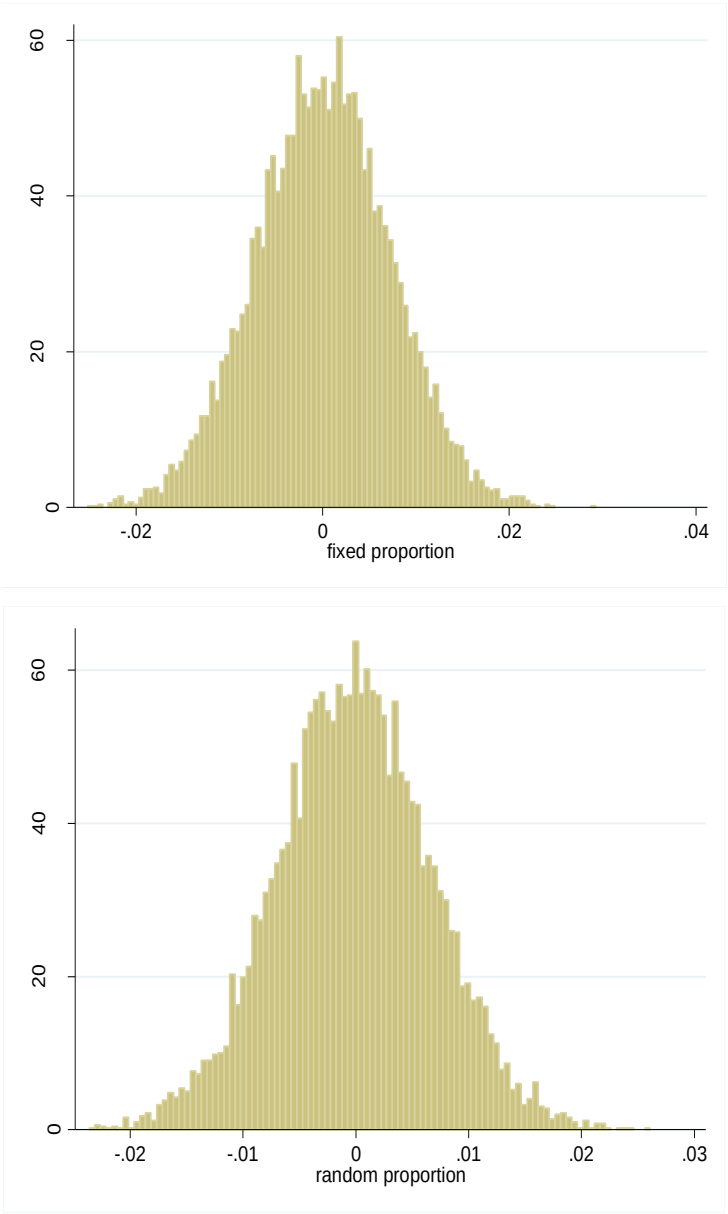
Table 2.4: Falsification analysis using young individuals (i.e. born after 1971)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
constant	8.479***	8.645***	5.523***	6.585***	5.667***	4.773***	6.823***	6.721***
	(1.564)	(1.564)	(1.556)	(1.559)	(1.590)	(1.589)	(1.546)	(1.543)
exposure 0.5 km	0.012	-0.008	-0.003	-0.001	0.003	-0.004	-0.003	-0.003
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)	(0.012)	(0.012)
male	-0.080***	-0.081***	-0.086***	-0.086***	-0.086***	-0.087***	-0.086***	-0.096***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
age	-0.396*	-0.406*	-0.397*	-0.398*	-0.407*	-0.394*	-0.406*	-0.350*
	(0.215)	(0.215)	(0.210)	(0.211)	(0.212)	(0.210)	(0.210)	(0.209)
age2	0.013	0.014	0.014	0.014	0.014	0.014	0.014	0.011
	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)
age3	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
years education	0.111***	0.108***	0.099***	0.099***	0.099***	0.098***	0.098***	0.095***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
age of arrival to Catalonia	-0.039**	-0.038**	-0.038**	-0.039**	-0.039**	-0.039**	-0.038**	-0.036**
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
age of arrival to Catalonia 2	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
area		0.000	-0.000	-0.000	0.000	-0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
distance to the nearest public school		-0.024	-0.015	0.013	-0.045	-0.012	-0.007	-0.008
		(0.063)	(0.053)	(0.054)	(0.058)	(0.052)	(0.059)	(0.058)
distance to Catalunya Square		-0.014***	-0.005*	-0.012***	-0.008***	-0.009***	-0.039***	-0.038***
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.012)	(0.012)
share of employed 1991			3.907***			2.297***	2.542***	2.427***
			(0.397)			(0.403)	(0.425)	(0.412)
share of highly educated ind. 1991			-0.580***			-0.913***	-0.125	-0.106
			(0.111)			(0.327)	(0.143)	(0.143)
share of individuals born in Spain 1991			-0.486***			-0.479	-0.338***	-0.340***
			(0.098)			(0.292)	(0.127)	(0.125)
share of employed 1996				2.838***		0.956***		
				(0.329)		(0.354)		
share of highly educated ind. 1996				-0.432***		-0.181		
				(0.124)		(0.411)		
share of individuals born in Spain 1996				-0.311**		-0.490		
				(0.123)		(0.533)		
share of employed 2001					3.301***	1.546***		
					(0.515)	(0.443)		
share of highly educated ind. 2001					0.165	0.383		
					(0.107)	(0.290)		
share of individuals born in Spain 2001					-0.190	0.563		
					(0.132)	(0.447)		
adults in the household								-0.015*
								(0.009)
children in the household								-0.109***
								(0.024)
single								ref. cat.
married								-0.075**
								(0.038)
widowed								-0.935
								(0.581)
divorced								-0.148
								(0.128)
separated								-0.195
								(0.203)
province of birth fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
years of residence fixed effects (1989-2001)	yes	yes	yes	yes	yes	yes	yes	yes
municipality-district fixed effects	no	no	no	no	no	no	yes	yes
adjusted R-squared	0.181	0.183	0.201	0.199	0.197	0.204	0.209	0.212
number of observations	24334	24334	24334	24334	24334	24334	24332	24332

OLS, Dependent Variable = self-reported knowledge of Catalan. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

Second, we randomly assigned each individual in the main sample to a census tract of the Metropolitan Area of Barcelona, with and without maintaining the proportion of observations in each census tract. We replicated this random assignment process 10,000 times and, for each permutation, we re-estimated our preferred specification [Equation (2)] and plotted the resulting distribution of fake beta coefficients. As can be seen in Figure 2.3, in both cases the placebo beta coefficients are distributed around zero and are very unlikely to be equal to the value estimated using the real exposure variable (0.030). This evidence again suggests that our estimates are not capturing unobserved local factors that correlate with language skills, but are rather capturing the real potential effect of being exposed to language learning opportunities.

Figure 2.3: fake beta coefficients from the permutation test



Note: estimated beta coefficients obtained after assigning at random the census tract to each individual, 10,000 replications. Left graph: permutation carried out maintaining the same proportion of individuals in each census tract as observed in the main sample (fixed proportion). Right graph: permutation obtained without keeping the proportion of observations in each census tract observed in the main sample (random proportion).

Furthermore, to have a better understanding of the effect of geographical exposure to language courses, we used different radii in the definition of the variable in our preferred regression, which also makes it possible to gauge the presence of distance-decay effects. The results are reported in Table 2.5. The first three columns show the β coefficient with radii of 0.5 km (the baseline), 1 km, and 1.5 km, respectively, again estimated using our preferred specification. We can see that as we enlarge the radius, the point estimate gets smaller and loses significance. To disentangle the effect of the geographic proximity, we included three variables that include information on the number of centres (by year) in each ring: closer to 0.5 km, between 0.5 km (not included) and 1 km, and between 1 km (not included) and 1.5 km. The results are shown in the last column of Table 5. The results point to a distance decay in the effect of the exposure, in both magnitude and significance. Actually, the number of centres in the outer ring seems to have no effect and the null hypothesis of equality of the exposure coefficients in column (4) is soundly rejected. This is clear evidence that proximity matters and indeed the number of centres located within 0.5 km from the centroid of the census tract is what matters more.

Table 2.5: Exposure at different radii

	(1)	(2)	(3)	(4)
exposure 0.5 km	0.030***			0.036***
	(0.010)			(0.011)
exposure 1 km		0.021***		
		(0.006)		
exposure 1.5 km			0.008**	
			(0.004)	
exposure 0.51-1 km				0.017***
				(0.006)
exposure 1.1-1.5 km				-0.000
				(0.005)
adjusted R-squared	0.454	0.454	0.453	0.454
number of observations	51624	51624	51624	51624

OLS, Dependent Variable = self-reported knowledge of Catalan. Control variables as in column (7) of tables 2.2 and 2.3. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

Additionally, to understand whether the estimated β coefficient is picking up direct or indirect effects of other activities that took place in the same centre in which the language courses were offered, as well as differences in student composition, we split the overall exposure variable by type of building (see Table 2.6). The results reported in Table 2.6 indicate that the point estimate of exposure to language courses undertaken in public school buildings is higher (and more statistically significant) than those of exposure to language centres located in community centres or other municipal infrastructures. This evidence goes against the idea that the baseline estimate is capturing the effect of other services provided in the same locations, but one could argue that it could be driven by differences in language student composition by infrastructure type (i.e. those who attend language courses located in public schools are those who would learn the language more easily anyway). However, the test for the equality of the coefficients of exposure to the three types of infrastructures does not allow the null hypothesis to be rejected (p-value = 0.68), which overall leads us to consider that our main estimates are capturing a "genuine" effect of geographical exposure to language schools (and thus we retain overall exposure in what follows).

Table 2.6: Exposure by type of building

	(1)	(2)	(3)	(4)	(5)
exposure 0.5 km	0.030***				
	(0.010)				
exposure 0.5 km - schools		0.043**			0.042**
		(0.017)			(0.017)
exposure 0.5 km - com. centres			0.021		0.022
			(0.015)		(0.015)
exposure 0.5 km - munic				0.026	0.026
				(0.022)	(0.021)
adjusted R-squared	0.454	0.453	0.453	0.453	0.453
number of observations	51624	51624	51624	51624	51624

OLS, dependent variable: self-reported knowledge of Catalan. Control variables as in column (7) of table 2.2. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

We also would like to discard the (residual) possibility that our estimates are still biased by the presence of endogenous residential sorting, even after controlling for a large set of census tract characteristics and district fixed effects. Hence, we replicated the analysis by excluding the individuals who might have chosen their place of residence considering the location of language centres. Specifically, Table 2.7 contains the results obtained after excluding individuals who had been living in the same place since before the creation of the CLN [i.e. since 1989 in column (2), since 1985 in column (3), and since 1980 in column (4)]. This exercise provides estimates slightly higher than the baseline (but not statistically different), increasing with the length of stay in the current dwelling of residence, which goes against the idea that endogenous residential choices are affecting the validity of our main results. Moreover, given the ambiguity of the definition of exposure for those who moved to the current place of residence in 2000, just before the 2001 census, we also try to exclude them [column (5)] and again the evidence remains stable. Finally, as a last sensitivity check, we replicate the estimations by excluding observations of individuals residing in the periphery of the metropolitan area, which is useful for two possible issues. On the one hand, census tracts located away from the city centres are larger in terms of area, that is, the population density[20] is much lower in these places and this could be related to omitted local unobservables. On the other hand, individuals residing in the periphery of the metropolitan area could have been exposed to other language centres located outside its limits (i.e. spatial spillovers from adjacent areas). The results are included in Table 2.A4 of the Appendix and are very similar to our main estimates.

Table 2.7: restrictions on year of arrival in the current dwelling

	(1)	(2)	(3)	(4)	(5)
exposure 0.5 km	0.030***	0.031**	0.038***	0.043***	0.033***
	(0.010)	(0.013)	(0.014)	(0.016)	(0.010)
adjusted R-squared	0.454	0.460	0.466	0.477	0.456
number of observations	51624	35572	30431	23002	50187

[20] Note that we do not have exact information about census tract area in 1991; we only observe the area in 2001. In alternative specifications, we tried to control for population density in 2001, which provides similar results (available upon request).

OLS, dependent variable: self-reported knowledge of Catalan. Control variables as in column (7) of table 2.2. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%. Column (1): baseline estimates (column (7) of table 2)). Column (2): only individuals living in the same place since 1989 or before. Column (3): only individuals living in the same place since 1985 or before. Column (4): only individuals living in the same place since 1980 or before. Column (5): only individuals living in the same place since 2000 or before.

2.5.1. Heterogeneous effects and local labour market outcomes

The next two tables show the heterogeneous effects of exposure. The former concerns individual characteristics and the latter considers local characteristics. In column (2) of Table 2.9, we consider differential effects of spatial accessibility of language centres according to gender. The point estimates are positive and significant for both genders, although they are somewhat higher for females. However, the test of equality of the coefficients does not allow its null hypothesis to be rejected. In column (3), we allow for heterogeneous coefficients according to age intervals. Interestingly, the effect decreases with age, being much stronger for individuals in the age range of 29 to 37 years. This is possibly due to the fact that these individuals (born between 1964 and 1971) were only partially exposed to Catalan at school and did not obtain the proficiency certificate that enables people to work in the (regional and local) public sector in Catalonia. Because they are likely to compete for the same jobs with individuals who are slightly younger, they had an incentive to enrol on courses offered by the CLN to obtain the language certificate.

We also analysed heterogeneous effects depending on completed education [see column (4)]. Although the coefficient of exposure is clearly much higher for individuals with a low education level and not statistically significant for those with post-compulsory schooling, the null hypothesis of equality of coefficients is not rejected. Columns (5) and (6) provide information about the heterogeneous effects depending on the individual's origin. First, it is possible to appreciate that geographical exposure to language training centres only benefited, in terms of proficiency in the local language, individuals born in Catalonia. The coefficient estimated for individuals born in the rest of Spain is virtually zero and not significant. In Column (6), we consider both the province of birth and age on arrival in Catalonia for those who were born abroad. Again, exposure to language schools does not affect language proficiency among

individuals born outside Catalonia even if they arrived during childhood. This evidence is indeed consistent with the results reported in Table 3, which indicate that accessibility of language centres mostly affects the probability of being able to speak and write in Catalan. Indeed, language courses provided by the CLN during its first decade of existence were very effective in shaping written proficiency among adult individuals born in Catalonia. Many of them were indeed native Catalan speakers who were orally fluent in Catalan thanks to intergenerational transmission of the language (Cappellari et al., 2021) but did not have written skills because they received education in Spanish only. Therefore, they took advantage of the language policy targeted to adults to acquire written skills in their native language.

Table 2.8: heterogeneous effects, individual characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
exposure 0.5 km	0.030***					
	(0.010)					
exposure 0.5 km - female		0.033***				
		(0.011)				
exposure 0.5 km - male		0.027*				
		(0.012)				
exposure 0.5 km - age 29-37			0.064***			
			(0.016)			
exposure 0.5 km - age 38-46			0.031*			
			(0.017)			
exposure 0.5 km - age 47-55			0.014			
			(0.016)			
exposure 0.5 km - age 56-64			0.015			
			(0.015)			
exposure 0.5 km - low education				0.034***		
				(0.012)		
exposure 0.5 km - high education				0.021		
				(0.013)		
exposure 0.5 km - born in Catalonia					0.053***	
					(0.011)	
exposure 0.5 km - born in the rest of Spain					-0.006	
					(0.014)	
exposure 0.5 km - born in Catalonia						0.046***
						(0.011)
exposure 0.5 - arrived with age 1-13						0.001
						(0.026)
exposure 0.5 - arrived with age 14-29						0.002
						(0.017)
exposure 0.5 - arrived with age 30 or more						-0.046
						(0.036)
p-value for coefficients equality test	--	0.605	0.030	0.418	0.000	0.013
adjusted R-squared	0.456	0.446	0.453	0.424	0.456	0.452
number of observations	51624	51624	51624	51624	51624	51624

OLS, dependent variable: self-reported knowledge of Catalan. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%. Control variables as in column (7) of table 2.2, except: column (3), control for age groups dummies rather than cubic age. Column (4), control for high-low education (post-compulsory vs lower levels) rather than years of schooling. Column (6), control for age of arrival dummies rather than quadratic age at arrival.

Finally, we explore the existence of heterogeneous effects according to local characteristics. We analyse whether the effect of exposure to language centres

varies according to a) the fraction of individuals born in the rest of Spain in the census tract, b) the fraction of adult individuals who are able to speak in Catalan, and c) the fraction of individuals who are able to speak and write in Catalan. For the three contextual variables, we consider whether the share observed in the census tract is higher than the overall proportion or not. The results are reported in Table 2.9 and suggest that exposure matters somewhat more in neighbourhoods with a higher share of individuals born outside Catalonia and (symmetrically) where there is a lower proportion of individuals who are proficient in Catalan. This seems to indicate that the public provision of local language training could mitigate the impact of neighbourhood composition, that is, the negative effect of living in a Spanish-speaking enclave on proficiency in Catalan. However, for any of the three variables considered here, the test for the equality of the estimated coefficients does not allow the corresponding null hypothesis to be rejected. Nevertheless, we consider that this issue deserves special attention and it will represent the focus of our future research.

Table 2.9: heterogeneous effects, local characteristics

	(1)	(2)	(3)	(4)
exposure 0.5 km	0.030***			
	(0.010)			
exposure 0.5 km - low % born in the rest of Spain		0.032***		
		(0.012)		
exposure 0.5 km - high % born in the rest of Spain		0.041**		
		(0.018)		
exposure 0.5 km - low % at least speak in Catalan			0.040**	
			(0.018)	
exposure 0.5 km - high % at least speak in Catalan			0.031***	
			(0.012)	
exposure 0.5 km - low % at least speak in Catalan				0.038**
				(0.018)
exposure 0.5 km - high % at least speak in Catalan				0.027**
				(0.011)
p-value for coefficients equality test	--	0.681	0.636	0.578
adjusted R-squared	0.454	0.453	0.453	0.453
number of observations	51624	51624	51624	51624

OLS, dependent variable: self-reported knowledge of Catalan. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, ***

significant at 1%. Control variables as in column (7) of table 2, except: column (3), control for an indicator of high/low fraction of individuals born in the rest of Spain rather than the share. Column (4), control for an indicator of high/low fraction of individuals who are able to speak or speak and write in Catalan rather than the share of individuals born in the rest of Spain. Column (5), control for an indicator of high/low fraction of individuals who are able to speak write in Catalan rather than the share of individuals born in the rest of Spain.

As the last evidence, we looked into possible labour market effects induced by geographical exposure to local language schools. As mentioned before, we consider as outcomes the probability of being employed (including among those who are not employed, mostly women, who are involved in household tasks) and, among employed individuals, weekly hours of work and indicators of having a part-time job (below 30 hours per week), working in the public sector, and having a highly skilled job. The first evidence we provide consists in the conditional correlation between skills in Catalan and these labour market outcomes (see Table 2.5A in the Appendix). The results indicate that skills in Catalan are conditionally correlated with labour market outcomes and both the size and sign of the estimated coefficients are as expected. Specifically, the probability of working is positively associated with the level of proficiency in Catalan. The effect is larger when we keep in the sample individuals involved in housekeeping tasks, suggesting that language skills are also related to labour market participation. Only the highest levels of proficiency are positively related to the probability of working in the public sector and having a high-skilled job. The only exception is for hours of work (and the derived dummy for part-time jobs), for which any of the coefficient of skills in Catalan are statistically different from zero. Nevertheless, when considering the reduced form estimates of the direct effect of exposure to language training centres, all the estimated coefficients are very low in size and not significantly different from zero (as can be appreciated in Table 2.10). This result is possibly due to the fact that, although skills in Catalan are rewarded in the labour market and actually correlate positively with several labour market outcomes considered here, the impact of exposure to Catalan learning centres on language skills is possibly too modest in size to induce any positive effect on labour market outcomes.[21]

[21] We also analysed possible heterogeneous effects on labour market outcomes, considering the same variables that we used in Tables 2.8 and 2.9. However, in this case too we were unable to detect any significant effect of exposure to language schools on labour market outcomes (results are available upon request).

Table 2.10: labour market outcomes

outcome:	employed	employed	ext. hours of work	partime	public sector	high skilled
exposure 0.5 km	0.001	0.003	0.11	-0.002	0.000	-0.001
	(0.002)	(0.003)	(0.075)	(0.002)	(0.003)	(0.003)
adjusted R-squared	0.025	0.237	0.054	0.044	0.156	0.339
number of observations	34931	45483	31358	31358	31358	31337

OLS, control variables as in column (7) of table 2.2. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

2.6. Conclusions

In this paper, we investigated the effect of the geographical accessibility of local language learning centres on the formation of language skills and labour market outcomes in a bilingual urban labour market, the Metropolitan Area of Barcelona. We exploited the creation of the so-called Consortium for Language Normalization (CLN), a public institution of the Spanish region of Catalonia that provided publicly subsidized courses of Catalan (the local language of Catalonia) targeted to adult individuals who were schooled in Spanish only. The empirical analysis considers the first decade of existence of the CLN (1989/90 to 2000/01) and is based on a measure of spatial availability of language courses that captures potential exposure in an Intention-to-Treat (ITT) framework.

Our results show a modest but robust effect of exposure to language centres on language proficiency, especially regarding the probability of being able to speak and write in Catalan. All the evidence from falsification exercises and sensitivity checks points towards the validity of the causal interpretation of our findings. The impact of the local supply of language courses is more pronounced for younger and less educated individuals and, most strikingly, is relevant only for those who were born in Catalonia: they are indeed those who are more sensitive to an increase of spatial accessibility (i.e. having more language centres near to the place of residence). Actually, this result is consistent with the fact that the availability of language learning centres only affects the probability of speaking and writing in Catalan. In fact, the latter domain of language proficiency is precisely what people born in Catalonia (many of them being native Catalan

speakers) wanted to achieve, because they were not exposed to language at school and had limited written skills. However, although competences in Catalan are positively related to labour market performance (especially employment probability and occupation), the effects induced by the geographical accessibility are too modest to lead to any improvements in any of the labour market outcomes considered in this work.

The evidence reported in this paper indicates that the public provision of local language training and its spatial accessibility matters for the acquisition and improvement of language competences in urban labour markets. Therefore, a first policy implication that can be derived from our findings is that geographic proximity helps to increase the effectiveness of language training programmes targeted to adults, which is also in line with the results regarding the spatial decay effect we have detected. This means that decentralizing the location of language courses and trying to reach target individuals is a sensible route to follow. However, the results regarding the heterogeneous effects, especially the fact that only individuals born in Catalonia are affected by the presence of language schools near to their place of residence, should also be carefully considered by policymakers. In fact, this indicates that additional effort is needed to stimulate the acquisition of language skills among individuals who are in a less advantaged position, who are (in this case) individuals born in other parts of Spain. Indeed, especially during the period of analysis, internal migrants and their descendants were likely to be spatially segregated in specific areas of the city (Garcia-Lopez et al., 2020). Accordingly, the geography of the supply of language centres should be more directly based on the residential location of the target population, in order to increase their availability in areas with a higher concentration of inhabitants with more need to improve their skills in the local language. This will possibly contribute to stimulating participation in language courses among individuals who have much more to gain from training in local languages. Moreover, it is possible that information about the availability of local language courses, as well as about the potential benefits related to local language acquisition, could also play some role in explaining the absence of any reaction to local supply of language courses among individuals born outside Catalonia. Although this possible channel cannot be investigated with available data, policymakers should more carefully consider implementing active information policies to reach their target groups and increase their language skills endowment. Finally, the absence of effects on labour market outcomes should be better analysed in the future, but one tentative policy recommendation in this

line is to review the content of language courses and their organization not only to foster their impact on language skills acquisition but also to orientate language courses towards the competences that are relevant in the bilingual labour market.

Appendix

Figure 2.A1: location of language centres by hosting institution, 1994/1995

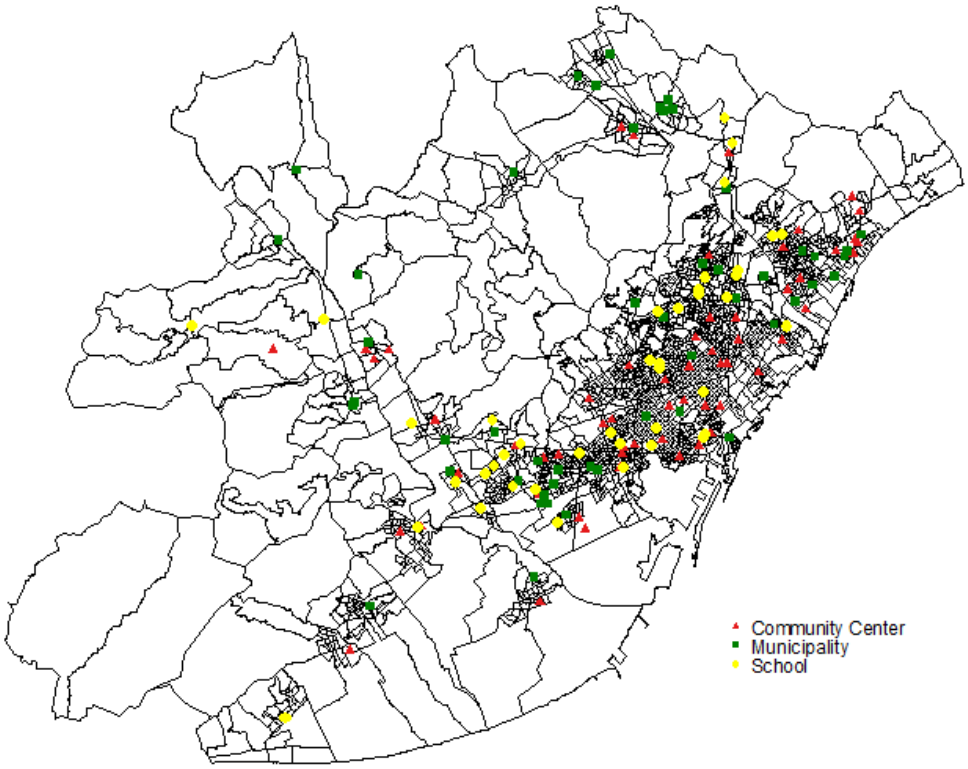


Figure 2.A2: location of language centres by hosting institution, 2000/2001

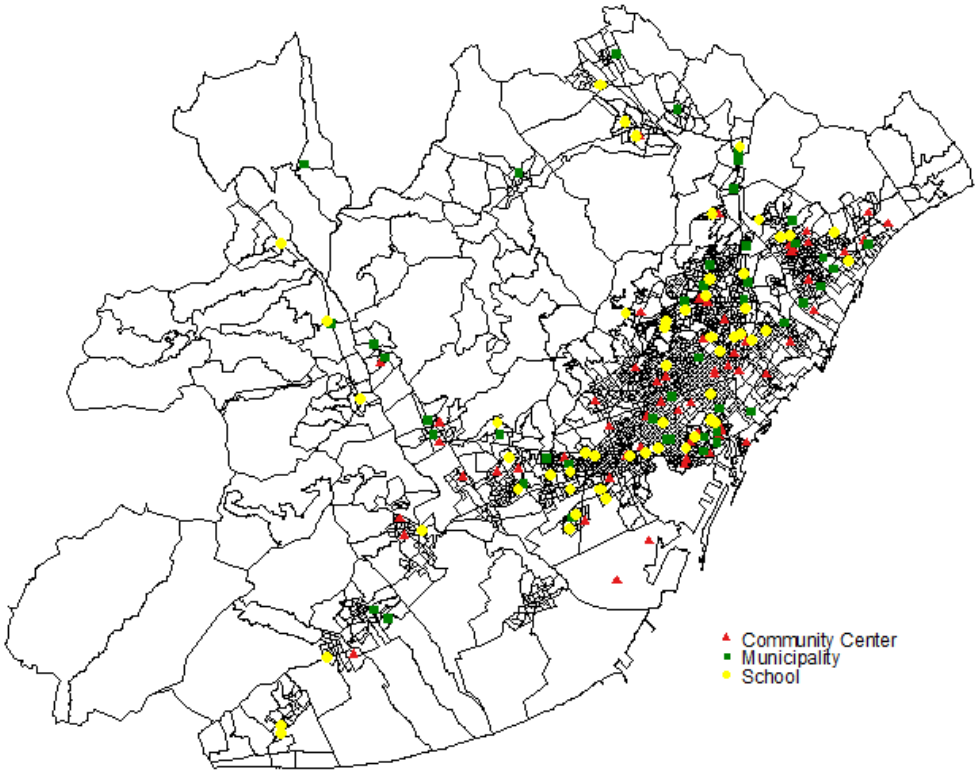


Table 2.A1: Province of Birth

Province of birth	Freq.	%
Born in Catalonia	**27967**	**54.17**
Barcelona	26,583	51.49
Girona	251	0.49
Lleida	689	1.33
Tarragona	444	0.86
Born in the rest of Spain	**23657**	**45.83**
Araba/Álava	23	0.04
Albacete	378	0.73
Alicante/Alacant	139	0.27
Almería	734	1.42
Ávila	102	0.2
Badajoz	1,616	3.13
Balears, Illes	104	0.2
Burgos	352	0.68
Cáceres	679	1.32
Cádiz	672	1.3
Castellón/Castelló	224	0.43
Ciudad Real	853	1.65
Córdoba	2,170	4.2
Coruña, A	276	0.53
Cuenca	454	0.88
Granada	1,968	3.81
Guadalajara	196	0.38
Gipuzkoa	84	0.16
Huelva	524	1.02
Huesca	443	0.86
Jaén	1,719	3.33
León	578	1.12
Rioja, La	100	0.19
Lugo	808	1.57
Madrid	500	0.97
Málaga	792	1.53
Murcia	611	1.18
Navarra	159	0.31
Ourense	496	0.96
Asturias	211	0.41
Palencia	149	0.29
Palmas, Las	48	0.09
Pontevedra	165	0.32
Salamanca	424	0.82
Santa Cruz de Tenerife	41	0.08
Cantabria	99	0.19
Segovia	79	0.15
Sevilla	1,682	3.26
Soria	335	0.65
Teruel	584	1.13
Toledo	161	0.31
Valencia/València	403	0.78
Valladolid	243	0.47
Bizkaia	118	0.23
Zamora	329	0.64
Zaragoza	588	1.14
Ceuta	90	0.17
Melilla	154	0.3
Total	51,624	100.00

Table 2.A2: OLS, Dependent Variable = indicator for speaking and writing in Catalan

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
constant	4.022***	4.083***	3.844***	4.058***	3.864***	3.743***	4.068***	3.796***
	(0.216)	(0.216)	(0.225)	(0.221)	(0.231)	(0.232)	(0.228)	(0.241)
exposure 0.5 km	0.014***	0.011***	0.012***	0.013***	0.014***	0.012***	0.011***	0.010***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
male	-0.044***	-0.043***	-0.040***	-0.040***	-0.040***	-0.040***	-0.040***	-0.041***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
age	-0.235***	-0.236***	-0.237***	-0.238***	-0.238***	-0.237***	-0.237***	-0.219***
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)
age2	0.005***	0.005***	0.005***	0.005***	0.005***	0.005***	0.005***	0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
age3	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
years education	0.036***	0.036***	0.033***	0.033***	0.033***	0.033***	0.033***	0.033***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
age of arrival to Catalonia	-0.008***	-0.008***	-0.008***	-0.008***	-0.008***	-0.008***	-0.008***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
age of arrival to Catalonia 2	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
area		0.000***	0.000	0.000	0.000*	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
distance to the nearest public school		0.012	0.004	0.007	-0.002	0.003	-0.004	-0.004
		(0.017)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
distance to Catalunya Square		-0.004***	-0.000	-0.002**	-0.001	-0.001*	-0.010***	-0.010***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
share of employed 1991			0.536***			0.290***	0.272***	0.287***
			(0.071)			(0.086)	(0.086)	(0.086)
share of highly educated ind. 1991			-0.130***			-0.137**	-0.001	0.001
			(0.033)			(0.068)	(0.044)	(0.044)
share of individuals born in Spain 1991			-0.328***			-0.101*	-0.232***	-0.227***
			(0.027)			(0.057)	(0.034)	(0.034)
share of employed 1996				0.400***		0.137*		
				(0.063)		(0.076)		
share of highly educated ind. 1996				-0.140***		-0.128		
				(0.035)		(0.092)		
share of individuals born in Spain 1996				-0.401***		-0.291***		
				(0.033)		(0.111)		
share of employed 2001					0.495***	0.277***		
					(0.092)	(0.097)		
share of highly educated ind. 2001					-0.038	0.096		
					(0.031)	(0.070)		
share of individuals born in Spain 2001					-0.385***	-0.014		
					(0.035)	(0.096)		
adults in the household								0.001
								(0.002)
children in the household								-0.008**
								(0.003)
single								ref. cat.
married								-0.024***
								(0.006)
widowed								-0.023**
								(0.010)
divorced								-0.001
								(0.011)
separated								-0.017
								(0.012)
province of birth fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
years of residence fixed effects (1989-2001)	yes	yes	yes	yes	yes	yes	yes	yes
municipality-district fixed effects	no	no	no	no	no	no	yes	yes
adjusted R-squared	0.309	0.311	0.317	0.318	0.317	0.318	0.320	0.321
number of observations	51624	51624	51624	51624	51624	51624	51624	51624

Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

Table 2.A3: descriptive statistics for the sample of young individuals used for falsification (1972-1985)

Variable	Obs.	Mean	Std. Dev.
skills in Catalan			
not understand Catalan	24,334	0.010	0.101
understand	24,334	0.053	0.225
read	24,334	0.040	0.197
speak	24,334	0.010	0.102
read and speak	24,334	0.054	0.227
read, speak and write	24,334	0.831	0.375
labour market outcomes			
employed (housework = missing)	12,958	0.847	0.360
employed extended (housework = 0)	13,374	0.820	0.384
hours of work	14,199	37.88	9.522
partime	14,199	0.122	0.328
overtime	14,199	0.124	0.329
public sector job (based on industry)	14,199	0.132	0.339
high skilled job	14,178	0.313	0.464
individual characteristics			
male	24,334	0.520	0.500
age	24,334	23.04	3.849
born in Catalonia	24,334	10.656	3.498
years of education (imputed from levels)	24,334	0.513	3.174
age of arrival to Catalonia (0 for natives)	24,334	0.946	0.23
adults in the household	24,334	3.503	1.196
children in the household	24,334	0.239	0.560
single	24,334	0.910	0.287
married	24,334	0.083	0.276
widowed	24,334	0.000	0.019
divorced	24,334	0.005	0.069
separated	24,334	0.002	0.042
local characteristics			
area	24,334	330111	1483119
distance to the nearest public school	24,334	0.209	0.265
distance to Catalunya Square	24,334	6.801	4.323
share of employed 1991	24,334	0.854	0.048
share of highly educated ind. 1991	24,334	0.229	0.151
share of individuals born in Spain 1991	24,334	0.438	0.145
share of employed 1996	24,334	0.778	0.064
share of highly educated ind. 1996	24,334	0.300	0.165
share of individuals born in Spain 1996	24,334	0.395	0.120
share of employed 2001	24,334	0.886	0.036
share of highly educated ind. 2001	24,334	0.340	0.161
share of individuals born in Spain 2001	24,334	0.354	0.111
exposure measures			
language centers within 0.5 km	24,334	0.928	0.858
language centers within 1 km	24,334	3.170	2.126
language centers within 1.5 km	24,334	6.394	4.026

Table 2.A4: estimation with observations from dense municipalities

	(1)	(2)	(3)
exposure 0.5 km	0.030***	0.022**	0.023**
	(0.010)	(0.011)	(0.011)
adjusted R-squared	0.454	0.445	0.452
number of observations	51624	37887	40412

OLS, dependent variable: self-reported knowledge of Catalan. Control variables as in column (7) of table 2. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%. Column (1): baseline estimation; column (2): estimation obtained using a subset of 7 large and dense municipalities around Barcelona (included); column (3): estimation obtained using a subset of 12 large and dense municipalities around Barcelona (included).

Table 2.A5: knowledge of Catalan on labour market outcomes

	employed	empl. ext.	hours of work	partime	public sector	high skilled
not understand			reference category			
understand	0.030**	0.074***	-0.229	0.014	-0.008	-0.001
	(0.013)	(0.012)	(0.356)	(0.011)	(0.010)	(0.011)
read	0.021	0.077***	-0.078	0.005	-0.001	-0.009
	(0.014)	(0.012)	(0.373)	(0.011)	(0.011)	(0.012)
speak	0.033*	0.100***	-0.028	-0.003	0.015	-0.003
	(0.017)	(0.017)	(0.482)	(0.015)	(0.015)	(0.018)
read and speak	0.036***	0.115***	0.335	-0.003	-0.014	0.023**
	(0.013)	(0.012)	(0.368)	(0.011)	(0.011)	(0.012)
read, speak and write	0.053***	0.171***	0.127	-0.015	0.079***	0.094***
	(0.014)	(0.012)	(0.372)	(0.011)	(0.011)	(0.012)
adjusted R-squared	0.026	0.243	0.054	0.045	0.164	0.344
number of observations	34931	45483	31358	31358	31358	31337

OLS, control variables as in column (7) of table 2. Standard errors clustered at the census tract level; * significant at 10%, ** significant at 5%, *** significant at 1%.

3. The Effect of Competition on Language Diversity in the Movie-Theatre Industry

3.1. Introduction

Cultural goods (such as books, films or theatre) and media products (newspapers, TV and radio) deliver their content to the consumers through a particular language. The existence of multilingual local markets makes the choice of language a difficult and sensitive decision for firms (Caminal, 2010). In fact, it has been estimated that more than one-half of the world's population speaks more than one language (Tucker, 2001). The language of these cultural goods will be a determinant factor for the decision of the consumer. For monolingual consumers, the decision will be straightforward: they will consume the cultural good which is in their language. For bilingual consumers, the decision will be more complex.

Economists have changed their view on language. Classically, it was viewed as a mere tool of communication. In this case, the bilingual consumer's decision would depend on the price and the quality of the cultural good; for identical cultural goods with different language options (e.g. a film offered in different versions), the bilingual consumer would choose the one with the lowest price. Nonetheless, nowadays language is no longer regarded by economists as a neutral tool of communication (Caminal & Di Paolo, 2019; Ginsburg & Weber, 2011). The efforts to promote minority languages, the resilience of linguistic groups, the multilingualism in supranational organizations (e.g. EU Parliament) are all examples of the existence and relevance of the subjective dimension of language, which goes beyond the communicative benefits. Such dimension shall not be neglected and actually plays a role in the decision of the bilingual consumer, who has to balance the preference for their own language[22] with the other characteristics of the cultural goods. Some kinds of linguistic preferences have already been introduced in a variety of economic frameworks. See, for example, Grin (1992), Wickström (2005), Mèlitz (2012) and Caminal (2010); the latter is the closest to this paper from a theoretical perspective. It extends the spoke framework (Chen & Riordan, 2007) by adding an additional dimension of product differentiation: language. In Caminal's model, consumers may trade

[22] The language to which she or he feels emotional attachment.

off a good match in terms of content against a good linguistic match for these cultural goods. This paper aims to empirically explore an actual cultural market, the cinema, in a bilingual local market. To the best of my knowledge, it is the first attempt to empirically analyse a cultural market focusing on the linguistic preference and its intensity among consumers.

The movies market presents the perfect set-up to explore the choice of the product's language by firms. As we can see in many countries, offering another language version rather than the original (that is, dubbing) is a common practice, especially when the knowledge of the original language is not generalized among all potential local consumers. Distributors have the rights of film distribution in the country, and they can decide to offer different versions. If one version does not exist, they need to incur the costs of dubbing so as to have this version available. In a second stage, local exhibitors, who can be independent cinemas or chains, negotiate with the distributor of the film they want to screen in the movie theatre. If the distributor has more than one version, that is, another version in addition to the original one, they will be able to choose their preferred version. Exhibitors typically choose only one version, although they could offer different ones, depending on the time, the day, or whether they can offer a different screen if it is a multiplex cinema. For instance, one cinema could offer the original version with subtitles in the country's language and the dubbed version of the same film to capture different kinds of demand. Also, in bilingual local markets the versions in the two languages could also be offered (e.g. Québec in Canada or Catalonia in Spain). Therefore, in the first step the distributor with the rights of distribution of the film decides to incur the cost of dubbing in a certain language if it wants to have an additional language version, as well as the original. In multilingual markets, the distributor can decide to incur the fixed cost of dubbing for each language version it wants to offer (for instance, a distributor in Spain of a film whose original language is English has to decide whether to incur the cost of dubbing for each language version: Spanish, Catalan, Basque and Galician[23]). In the second stage, exhibitors negotiate (non-exclusive) contracts with the distributor and decide which versions will be screened.

[23] Spanish is the official language of the country. Catalan is the co-official language in three regions which represent 29.5% of the population, Basque in two regions, representing 6.1%, and Galician in one region representing 5.7% of the total population (source: *Instituto Nacional de Estadística*, 2021).

The aim of this paper is to analyse whether the structure of the industry affects the language diversity in the movies market, a cultural market in which language matters. More precisely, I study how the concentration at the distribution and the exhibition levels affects the supply of films in the local language in the bilingual region of Catalonia (Spain). By using web-scraping techniques, I built a unique data set of all the screenings in the cinemas of Catalonia during one year, which I merged with other sources of data: census data, information on the distributors of the *Institut Català d'Empreses Culturals* (Catalan Institute of Cultural Enterprises; ICEC) and geographical information of the *Institut Cartogràfic i Geològic de Catalunya* (Cartographic and Geological Institute of Catalonia, ICGC). Next, I conducted an empirical analysis to explore the effect of the three relevant factors: concentration at the distribution and the exhibition levels, and the demand for films in Catalan language. I found that the concentration at the distribution level reduces the percentage of films in Catalan by 4.04 percentage points compared with the counterfactual of perfect competition. The effect of the concentration at the exhibition level is not significant. This implies that without such market failure the total supply of films in Catalan would be 96% greater (8.24% instead of the actual 4.2%). Then, the effect was disentangled depending on the target audience of the film: children audience or general adult public. The results from the heterogeneous analysis by type of audience indicate that children have higher preference intensity over the language because the market is more responsive to them; the concentration at the exhibition level matters when it comes to this type of consumer. Moreover, the effects are robust to many alternative specifications.

Thus, this paper contributes to the scarce literature on language and competition (Doh-Shin et al., 2021), but it also relates to the literature on the movie-theatre industry (Leung et al., 2020; Orbach & Einav, 2007). In a broader scope, it also contributes to the literature on competition and vertical relations (Allain et al., 2016; Fauli-Oller & Sandonis, 2016; Gans, 2007).

The rest of the paper is organized as follows: in the next section I will provide some background to the market and the case study. In Section 3, the theoretical framework will be formulated. Afterwards, I will show the data and descriptive statistics in Section 4. The empirical methodology will be presented in Section 5 and will be followed, in Section 6, by the results, which provide empirical evidence and validate the main hypothesis. Finally, I will conclude with a brief summary of the results, as well as some of their policy implications.

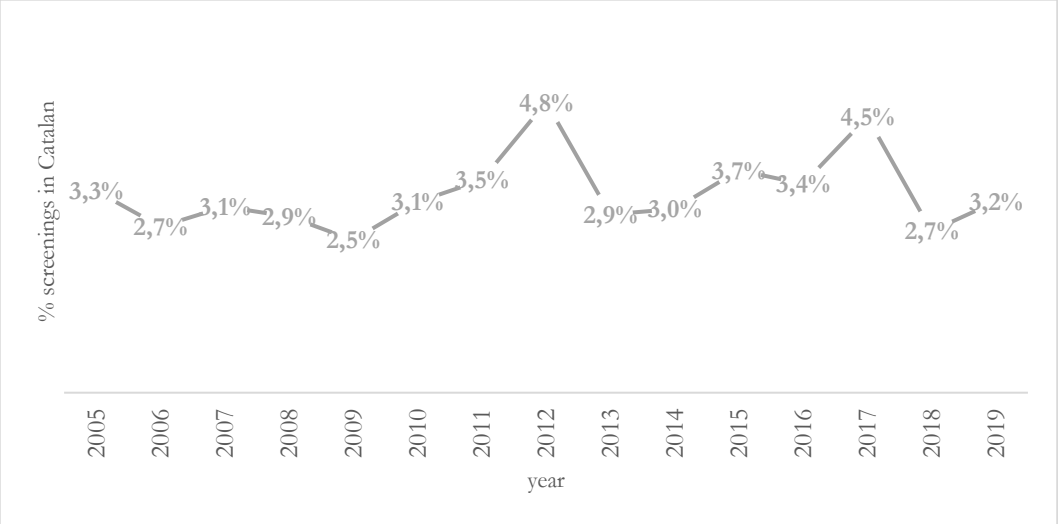
3.2. Background

In this section I will present some stylized facts to better understand the movies market in the Spanish region of Catalonia.

The population of this Autonomous Community[24] is above 7.5 million inhabitants[25] (source: IDESCAT). According to the Survey on Language Uses of the Population of 2018, Catalan is the most used language for 36.1% of the population, 48.6% use mostly Spanish and 7.4% both (and the remainder use other languages). With regard to knowledge of the languages, 81.2% of the population are able to speak Catalan and 99.5% (virtually everyone) are able to speak Spanish. This indeed shows an asymmetry in knowledge, which corresponds with the framework that will be presented in the next section.

Thus, the case analysed is a good example of a bilingual market. In such markets we can find the two versions (of the two official languages) sometimes offered, either as dubbed films of foreign productions and the original version. In the last 15 years, there has been a low but stable supply of films in Catalan (see Figure 3.1).

Figure 3.1: Share of films in Catalan

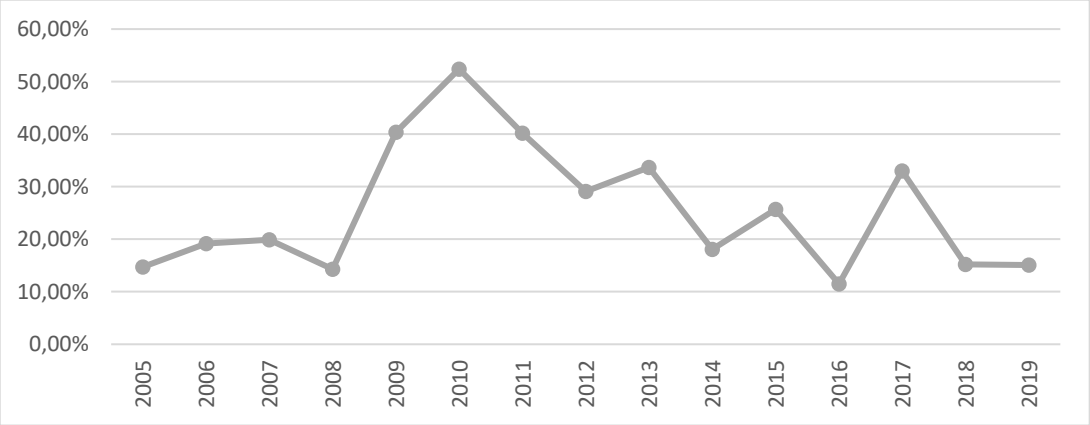


[24] Administrative division of Spanish regions.
[25] 7,716,760 according to IDESCAT in 2021. Provisional figure, from the Estimates of Population (IDESCAT) and "Cifras de población" (INE).

Note: films in Catalan over the total number of screenings. Source: IDESCAT.

Moreover, among these films, except in 2010, the majority of them did not have Catalan as the original language. This means that most of these screenings are in Catalan due to the decision of the distributors and the exhibitors that chose to offer the Catalan version (see Figure 3.2).

Figure 3.2: Share of films with original in Catalan out of the total films in Catalan



Note: The percentage is with respect to all the screenings in Catalan (dubbed, subtitled or original version). Source: IDESCAT.

The supply of films in the Catalan version might seem disproportionately low if we compare it with other cultural goods such as TV, radio or theatre. The share of annual time watching TV in Catalan was around 20% from 2017 to 2020 (source: IDESCAT). If we look at the radio listeners, Catalan language broadcasters have more than 50% of the market (source: General Study of the Media, EGM; *Estudi General dels Mitjans*). Theatre performances are also mostly in Catalan, as it can be seen in the Table 3.1:

Table 3.1: Language in Theatres

Theatre:	2015	2016	2017	2018	2019
Catalan	58.4%	64.4%	58.0%	59.9%	56.2%
Spanish	24.3%	21.0%	24.7%	23.7%	25.6%
Bilingual	8.2%	5.2%	7.0%	7.8%	7.7%
Non-spoken	7.1%	6.7%	7.3%	5.5%	7.8%

Source: IDESCAT

Indeed, these cultural markets are not perfectly comparable to the movies market in many aspects and, specifically, they can have a different demand (e.g. the average consumer of theatre is wealthier and more educated than the average consumer of cinema). However, such sharp differences cannot only be explained by the public that each cultural product attracts. Two things may happen: linguistic preferences change depending on the cultural good to be consumed, and/or the markets differ on the capability to match the demand for each language version.

The *Òmnibus* survey of the Centre of Opinion Studies (in Catalan, *Centre d'Estudis d'Opinió*) of the Catalan government asked about the language version preference over different cultural goods in several waves (2014, 2015, 2018 and 2019). If we look at the most recent wave (2019), we see that for the same random sample of individuals, the preference regarding the language version changes depending on the cultural good (see Table 3.2).

Table 3.2: Preferences over language for cultural goods

Version	Books	Cinema	Theatre	Music
Spanish	47.42	49.08	36.92	17.42
Catalan	23	13.08	25.17	5.42
Original version	5.92	17.17	8.67	41.67
Indifferent	19.75	17.58	23.0	32.58
Other	2.42	1.08	0.42	2.33
Does not consume it	0.75	0.92	1.08	0.08
No answer	0.17	0.58	1.58	0.17
Does not know	0.58	0.5	3.17	0.33
Total	100	100	100	100

Source: Òmnibus 2019, *Centre d'Estudis d'Opinió*

This table does not represent the language demand for each good, but the preferences of this random sample of 1,200 individuals. Thus, the variability between products is the result of the pure change of the revealed linguistic preference. For instance, if we restricted the sample to those who answered that they actually consume theatre periodically, the Catalan version is preferred by a 35.6% and the Spanish version is preferred by a 29.2%; that would explain the predominance of performances in Catalan in the theatre market.

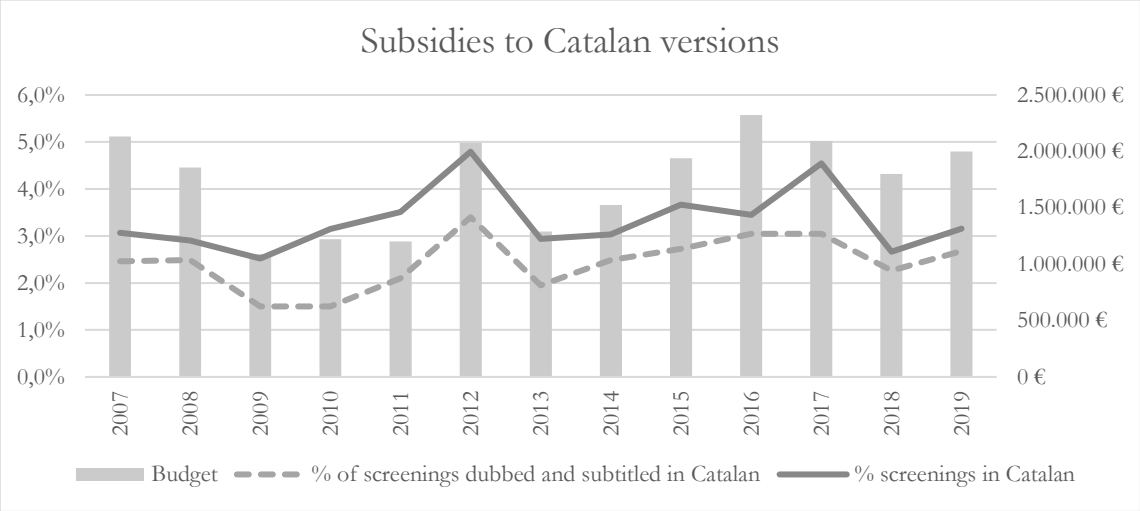
Even though there is a disparity in the linguistic preferences over different cultural goods, that cannot explain by itself the differences in supply that were presented before. As we could see in Figure 1, the percentage of screenings in Catalan has been wavering around 3–4%. Throughout the waves of this *Òmnibus* survey, the preference for cinema in Catalan has always been much higher than that: 21.7%, 24.2%, 20.9% (2014 waves), 22% (2015), 15.9% (2018) and 13.1% (2019). Hence, the supply of films in Catalan has always been between 5 and 7 times lower than these revealed preferences.

One concern regarding the responses of the survey might be that they are somehow politically motivated. If there was some political bias that, for instance, made indifferent people answer 'Catalan', such bias should be similar in the four cultural goods that are shown in the table. Even in this case, the disproportion between the actual supply and these declared (and perhaps biased) preferences would still be greater in the case of Cinema. On the other hand, the drivers of the language preference are irrelevant to the market: better knowledge of the language, emotional attachment or national identity, they all have the same effect in terms of demand behaviour: to be more inclined to one version rather than another.

The debate about the undersupply of cinema in Catalan is not new. On 30th June in 2010, the Catalan Parliament approved by 117 votes out of 135 (87%) the so-called Law of the Cinema (in Catalan, *Llei del Cinema*), which established that 50% of the productions should be either dubbed or subtitled in Catalan. Later, the Constitutional Court declared that such a percentage was unconstitutional and lowered it to 25%. Then, with the quota of 25% approved by the Constitutional Court, it was time to apply the law. Nonetheless, the Catalan government never applied it, perhaps due to the political pressure from the exhibitors and distributors, who threatened to stop supplying cinemas in Catalonia.

Since the regulation was not effective, the Catalan government has tried to influence the supply of films in Catalan through the dubbing costs, by subsidizing them.[26] The budget efforts for such item have varied over time, depending on the budget constraints of the moment, as well as the political preferences of the party in power. On 27th July (2021), a major increase in this budget item was announced, reaching the peak of 2,705,000€. The historical series is presented in Figure 3.3:

Figure 3.3: Subsidies for Catalan versions and films in Catalan



Source: Language policy reports of the Language Policy Secretariat (Government of Catalonia).

Apparently, the subsidies do have an effect on the supply of screenings in Catalan. Nonetheless, precisely because the costs of dubbing are much lower with the subsidies, the gap between the declared demand (in the surveys) and the supply have to be explained by other factors.

The question is, therefore, if it is true that there is an undersupply of films in Catalan language, why does this happen? Is there any market failure that leads to that outcome?

[26] Distributors can apply for the subsidies whenever they want to dub a film. These subsidies are non-competitive.

3.3. Theoretical Framework

In this section I present an informal, standard, theoretical framework that will provide a better understanding of the empirical results. In particular, the impact of concentration on language diversity will be interpreted as a measure of the inefficiency caused by market power.

In this representation of the movie-theatre industry, consumers have preferences regarding the content of the film as well as the language. I consider two languages: strong and weak. Native speakers of the strong language are monolingual (they cannot watch the film in the weak language). In contrast, native speakers of the weak language are bilingual (they can watch films in either language) but they prefer to watch movies in the weak language. Thus, in such a context a film is always released in the strong language and the question is whether the weak-language version is also provided.

Upstream Market

Distributors own the distribution rights of a number of films and choose whether or not to produce the weak-language version.

Downstream Markets

These are geographically segmented. Consumer prices are exogenously fixed (Orbach & Einav, 2007). Downstream firms, the exhibitors, supply films in one or several local markets. If the weak-language version exists, then the distributor bargains with each of the exhibitors, among other things, about its possible exhibition in the local market (adoption).

In the absence of informational and other types of frictions, we should expect the contract to maximize the joint payoffs of the distributor–exhibitor pairing (efficient bargaining). If the pair chooses to exhibit the film in both languages, then:

1. They incur a fixed cost
2. They steal some consumers from rivals (other pairs, distributors–exhibitors). That is, they attract consumers with strong preferences over language version that like the content of other films better (business-stealing effect).

3. They attract consumers who in the absence of the weak-language version would stay at home (market expansion effect).

The business-stealing effect increases with the degree of competition (fragmentation of the supply) in both the upstream and the downstream markets. In one of the extremes, if the exhibitor is a local monopolist, then the business-stealing effect is zero. The lower the market share of the exhibitor in the local market then the higher the temptation to steal business from rivals. In addition, both the business-stealing effect and the market expansion effects increase with the fraction of consumers with a preference for the weak language (this can be thought of as the demand for the weak language).

Prediction in the Downstream Market (Adoption)

The rate of adoption of weak-language versions increases with (1) the degree of competition in the local market, and (2) the fraction of consumers who prefer the weak language.

For the weak-language production decisions, the analysis is analogous. If the distributor produces the weak-language version then: it incurs a fixed cost, and generates business stealing and market expansion effects. All these effects will be filtered by negotiation with the exhibitors. Once again, the higher the degree of competition in distribution the stronger the business stealing effect and hence the higher the provision of weak-language versions.

In addition, competition in the upstream market causes a countervailing effect: higher competition in the upstream market reduces the bargaining power of the distributor, vis-à-vis exhibitors. Hence, if the distributor appropriates a lower share of the rents, then it will be less willing to produce more weak-language versions. Similarly, more competition in the downstream markets also increases the distributors' bargaining power (smaller and/or more fragmented exhibitors will have lower bargaining power with distributors), which enhances the incentives to produce weak-language versions.

Prediction in the Upstream Market (Production)

The rate of production of weak-language versions balances two effects of different signs: (1) it increases with the business-stealing effect and (2) it decreases with the bargaining power effect. As a result, the effect of market power in the upstream market on language diversity is, in principle, ambiguous.

Thus, the expected results are the following:

- **Higher competition (more fragmented market) at the distribution level will have an ambiguous effect on language diversity;**
- **Higher competition (more fragmented market) at the exhibition level will have a positive effect on language diversity;**
- **Higher demand (greater share of consumers that prefer the weak-language version) will have a positive effect on language diversity.**

Regarding the welfare analysis, let us ignore for the moment the market expansion effect. Will the equilibrium fraction of films in the weak language be excessive or insufficient from a total surplus point of view? According to our theoretical framework, the answer is ambiguous. However, it is important to emphasize that in the realistic case that fixed costs are relatively high, so that the equilibrium fraction of films with two linguistic versions is low, then the equilibrium level of linguistic version is inefficiently low.

In order to simplify the presentation of the welfare analysis, let us focus on the case in which upstream and downstream firms are integrated. Note that when firms are not integrated and the exhibitors share some of the surplus generated by the dubbed version, distributors' incentives to invest in linguistic diversity are further diminished. When the firm chooses whether or not to produce a second linguistic version, it compares the additional audience stolen from rivals with the fixed cost. Instead, the social planner internalizes the fact that the extra profits of the firm and the lost profits of the rivals cancel each other out. Thus, it only compares the increase in consumer surplus (speakers of the weak language get a better match) with the fixed cost. Hence, private incentives are insufficient or excessive, depending on the size of the business-stealing effect relative to the enhanced consumer surplus effect. The business-stealing effect positively depends on the intensity of language preferences as well as on the degree of competition. Thus, a higher preference intensity raises both the extra consumer surplus and the business-stealing effect. However, the business-

stealing effect increases with the degree of competition, and goes to zero as we approach monopoly. Thus, in markets with a low degree of competition, the equilibrium level of linguistic diversity will tend to be inefficiently low.

To deal with the case of markets with intense competition we can refer to some previous research. In a similar model of monopolistic competition, Caminal (2010) showed that the extra consumer surplus is lower (higher) than the business-stealing effect when the fixed cost of producing an additional linguistic version is relatively low (high), whereas the business-stealing effect is similar in both cases. The reason is simple. If the fixed cost is low, then most rivals will also offer the weak-language version. In this case, a firm that does not offer the weak-language version only attracts a small fraction of consumers with a preference for such a linguistic version. Thus, when the firm chooses to introduce the weak-language version, only a small fraction of consumers benefits from it. As a result, the private incentives exceed the public incentives and the equilibrium level of language diversity is excessive. On the contrary, if the fixed cost is relatively high and rivals only rarely offer the two linguistic versions, the extra consumer surplus generated by the additional version is large, private incentives are lower than social incentives, and therefore the equilibrium level of language diversity is inefficiently low. In the market we study in this paper, the fraction of versions in the weak language is small, which suggests that the latter case is the one that is empirically relevant.

Finally, note that when enhanced language diversity generates additional consumption (market expansion effect), private incentives are insufficient, as the firm cannot appropriate all the rents generated by the additional consumption.

Summarizing, if the market expansion effect is strong enough or if the intensity of competition is low, or if the fixed cost of producing a linguistic version is relatively high, then the equilibrium level of linguistic diversity is insufficiently low from a welfare point of view.

3.4. Data and Descriptive Statistics

For the empirical analysis of this paper, I combine several databases to answer the research question. First, I use a data set of screenings in all the movie

theatres in Catalonia from July 2020 to May 2021,[27] which contains 181,978 observations. This data set includes information about the weekday, the time of the screening, the week, the name of the movie, the cinema and, our relevant variable, the version, which can be in Catalan, in Spanish, Original version with Catalan subtitles and Original version with Spanish subtitles (when the original language is neither Catalan nor Spanish). The source of the database is the newspaper *El Periódico*, one of the most read newspapers in Catalonia, which publishes the billboard every week.[28] To gather all this information, I have used 'web-scraping' techniques, which enabled collecting in a systematic manner all the information from the billboard during the gathering period. This makes my data set quite unique, since I constructed it.

I also used web scraping to obtain the film's country of production, as well as its genre. In this case, the information was taken from FilmAffinity, a movie recommendation website that works as a huge movie database with more than 125,000 films listed. This information was matched with the main data set using the name of the film.

Similarly, I web-scraped International Movie Database (IMDb), the largest movie database, to obtain information on the original language of the film.

In addition, I also included cinema characteristics, such as the property (private, public or semi-public) and the geolocation through its address. I included information about the chain it belongs to in the case where it is not an independent cinema. This is crucial because in order to compute the competition measure of the exhibitors (cinemas), I had to take into account that they do not compete against cinemas of the same chain, but rather coordinate, avoiding cannibalization.

Another source of data was the Census 2011, which gave me information about the demand for Catalan versions for each cinema. Although the census does not ask about the linguistic preference in films, it asks about the ability to speak Catalan; therefore, the percentage of bilingual people (able to speak Catalan) can be a good proxy of the demand for the Catalan version. Since the information

[27] Although during this period movie theatres were not working at full capacity due to the COVID-19 restrictions (especially in 2020), that should only affect the number of observations, which is still very large, but not the rationale explained in Section 3.
[28] Each Friday the billboard is updated, so 'cinema weeks' run from Friday to Thursday.

is given by census tract, which is also geolocated, I can impute the proxy for demand to each cinema, as explained below.

Finally, I obtained the information of the distributor of each film through the Catalan Institute of Cultural Enterprises (in Catalan: *Institut Català d'Empreses Culturals*) and ComScore. That information was matched to the main data set through the name of the film, and allowed me to compute the market share for each distributor in my data set.

3.4.1. Measures of Competition and Market Power

In order to analyse how market power and competition can affect the likelihood that a screening is in the Catalan version, I had to compute a proxy for the two levels of competition: the distribution and the exhibition.

The distribution is characterized by a business-to-business market. The marketplace is the whole country in which the distributor has the rights of the film, so in this case it would be Spain. The distributor can decide to offer the Catalan version or not, in the case where the original version is not in Catalan, and if so, it would then be up to the exhibitors to offer this version in some of their screenings. The distributor, hence, incurs the fixed cost of dubbing, although this is often subsidized by the Catalan government. The bargaining power of the distributor will depend on their market share, and that is why I compute the share of the market for each distributor in the data set (market as number of screenings) and use this variable as proxy for market power.

There are 89 distributors in the sample. It is important to note that this is the market share by screenings, not by tickets sold. I could get information on the distributors of 180,032 of the 181,978 film screenings. Table 3.3 shows the screenings of each distributor in the sample as well as their share.

On the other hand, the exhibition is a business-to-consumer market, in which movie theatres can belong to a chain or can be an independent cinema. The marketplace, unlike the distributors, is not the whole country but its area of influence. We cannot consider that two cinemas, which are, for instance, more than 20 km[29] from each other, compete between them, since they might share few or no potential consumers. We try several radii in order to check whether a

[29] Davis (2006) finds that 'geographic markets consist of at most 15-mile circles around theatres, probably less'.

10 km radius is in fact an appropriate baseline. Moreover, it is also unlikely that two neighbouring cinemas that offer different films will compete on the language version: the consumer that wants to watch film A and prefers the Catalan version over the Spanish one, will decide to go to another cinema if it offers the same film A in their preferred language version, but will not change their preference for watching film A because another film B is offered in Catalan.[30] The assumption is that the preference for film goes before the preference for the language version, and so each film has its own demand. Furthermore, I also consider that film screenings compete in a concrete point in time: screenings in different weeks do not compete with each other. If *Star Wars: Episode 1* was available in one cinema, we cannot consider that it competes in language version with all the cinemas that offered the same film many years ago. The same applies for weeks for a simple reason: the billboard is updated every week so the consumer only knows the films offered in the following 6 days. The language version of a film screening in a certain week cannot be a factor of differentiation with another cinema offering the same film the following week because the consumer cannot know about that. Thus, the proxy of competitive pressure will be the number of same film screenings in the surrounding area that specific week, excluding those of cinemas in the same chain. This allows for great variability, unlike the proxy of market power of the distributor, which only depends on the distributor.

[30] The same applies for Original Version with Subtitles. By default, the different versions are considered to play a role for a given film. This assumption is later relaxed.

Table 3.3: Distributors

Name	Screenings	Share	Name	Screenings	Share
WBI	39,623	22.01%	EURODF	479	0.27%
UPI	20,871	11.59%	FESTIVAL	476	0.26%
ACONTRA	13,975	7.76%	BOSCO	457	0.25%
DIAMOND	12,858	7.14%	INDP	439	0.24%
SONY	10,697	5.94%	PAYCOM	385	0.21%
FILMAX	9,321	5.18%	RITA&LUCA	355	0.20%
DISNEY	7,104	3.95%	SYLDAVIA	329	0.18%
FLINS	6,426	3.57%	BEGIN AGAIN	327	0.18%
BTEAM	6,315	3.51%	PACKMAGIC	325	0.18%
DEAPLANETA	6,181	3.43%	AVENTURA	305	0.17%
eOne	5,022	2.79%	BENECE	300	0.17%
SELECTAVISION	3,869	2.15%	PPI	246	0.14%
AVALON	3,797	2.11%	39 ESC	185	0.10%
CARAMEL	3,586	1.99%	SEGARRA	173	0.10%
TRI	3,537	1.96%	SHER	164	0.09%
ALFAPICT	3,437	1.91%	SPLENDOR	134	0.07%
VERCINE	2,770	1.54%	ATALANTE C	101	0.06%
ADSO	1,846	1.03%	BARTON	99	0.05%
VERTIGO	1,819	1.01%	CAPRICCI	99	0.05%
VERTICE	1,782	0.99%	FOX	70	0.04%
KARMA	1,594	0.89%	PARK CIRCUS	69	0.04%
SURTSEY	1,477	0.82%	REVERSO	62	0.03%
CINEMARAN	1,337	0.74%	SURFILMS	60	0.03%
WAND - AVAL	905	0.50%	EMON	55	0.03%
Wanda	794	0.44%	FLAMINGO	52	0.03%
FILMIN	772	0.43%	PREMIUM	37	0.02%
GOLEM	699	0.39%	PACK	36	0.02%
VERDIG	517	0.29%	MEDIA SOLUTIONS	33	0.02%
BIGPICTURE	501	0.28%	BARLOVENTO D.	25	0.01%
ELAMEDIA	492	0.27%	Others (30 distributors)	231	0.13%

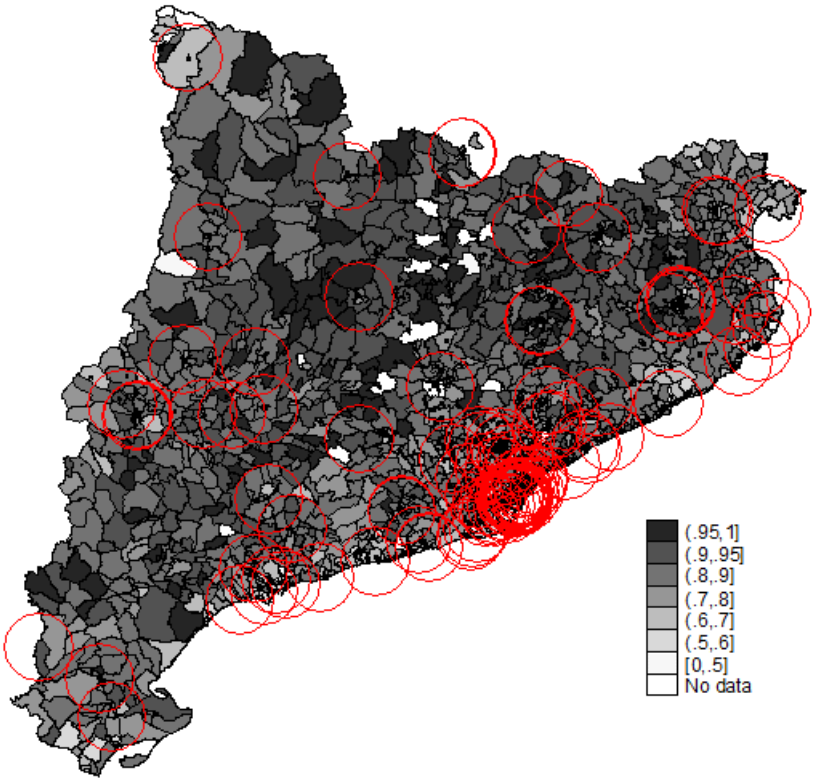
3.4.2. Demand

It is impossible to know the precise demand[31] for each language version since one individual could have different preferences for different films if they were bilingual: a spin-off film from a series that they used to watch in Catalan would lead to a preference for the Catalan version, even if, for the rest of the films, they had a preference for the Spanish version. However, since most of the films are not a spin-off, I consider that the main determinant will be the own language. The census data do not include information on the mother tongue or the identification language, but they give information on competences in Catalan

[31] Henceforth, 'demand' will mean 'the demand for the Catalan version'.

(ability to understand, speak and write). In order to assign the demand to each cinema, I apply a similar method to the one I used when constructing the proxy of competitive pressure for exhibitors: I set a 10 km radius and assign all the census tracts whose centroids fell within the area of influence to the cinema. Hence, I can calculate the fraction of the population assigned to the cinema that is able to speak Catalan (see Figure 3.4).

Figure 3.4: Cinemas and Bilingual Population



The average percentage of bilingual population in the data set (i.e. the 181,978 screenings) is 72.8%. This is lower than the percentage of the bilingual population (able to speak Catalan[32]) of the Survey on Language Uses of the Population of 2018 (recall, 81.2%) because cinemas tend to be concentrated in urban areas, in which the percentage of bilingual population is below the mean.

[32] As explained in Section 2, virtually all individuals are able to speak Spanish.

3.4.3. Dependent Variable

Table 3.4 reports the language versions that we can find in the data set:

Table 3.4: Language versions in the dataset

Version	Freq.	Percent
3D Spanish Version	355	0.2
Catalan Version	7,666	4.21
Spanish Version	152,820	83.98
Original Version with Subtitles in Catalan	459	0.25
Original Version with Subtitles in Spanish	20,628	11.34
No information	50	0.03
Total	181,978	100

Only 50 screenings had no information on the language version (0.03%). Original versions with subtitles represent 11.59% of the total. Original versions with subtitles are not comparable to Spanish and Catalan versions, since the preference for the language of the subtitles might not be relevant for the individual, but rather the language of the original version (for which I do not have information). Putting together OVSC+CV and OVSS+SV[33] could be misleading. This is why I decided to not consider subtitled versions for the computation of the dependent variable. Therefore, the dependent variable *Catalan* will be a dummy that takes value 1 if CV and 0 if SV (or 3D SV).

3.4.4. Genres

From the FilmAffinity database I could web-scrape information about 169,778 screenings. In Table 3.5, I present the distribution of these screenings among the genres.

[33] OVSC: Original Version with Subtitles in Catalan. CV: Catalan Version; OVSS: Original Versions with Subtitles in Spanish; SV: Spanish Version.

Table 3.5: Film genres

Film Genre	Freq.	Percent	Cum.
Action	9,719	5.72	5.72
Animation	26,138	15.4	21.12
Adventures	142	0.08	21.2
War	30	0.02	21.22
Science fiction	3,103	1.83	23.05
Comedy	27,409	16.14	39.19
Documental	7,239	4.26	43.46
Drama	36,162	21.3	64.76
Fantastic	6,264	3.69	68.45
Intrigue	3,619	2.13	70.58
Musical	4,113	2.42	73
Romance	8,828	5.2	78.2
TV Series	1,602	0.94	79.14
Terror	9,400	5.54	84.68
Thriller	25,572	15.06	99.74
Western	438	0.26	100
Total	169,778	100	

Note that 15.4% of the films are in the category Animation'. This distinction will be used later on in the analysis of the heterogeneous effects.

3.4.5. Summary Statistics

Table 3.6 reports the summary statistics of the relevant variables as well as the other control variables for the sample of the baseline model, that is, screenings which are either in Catalan or Spanish (not subtitled versions) and do have information on film genre.

Table 3.6: Summary Statistics

Variable	Mean	Std. Dev.	Obs
Catalan	0.042	0.201	152,707
ShareDist	0.095	0.077	152,707
CompExhib (10km)	42.78	71.923	152,707
CompExhib (20min drive)	48.839	79.017	152,707
Demand (10km)	0.732	0.051	152,707
Extension 1: films are substitutes within each genre			
ShareDist	0.269	0.262	152,707
CompExhib (10km)	109.008	165.505	152,707
Extension 2: upstream and downstream markets separately			
Catalan-Version	0.233	0.423	152,707
First stage, observation units are films			
Catalan-Version	0.149	0.356	626
ShareDist	0.044	0.060	626
Second stage, observation units are screenings but restricted to Catalan-Version=1			
Catalan	0.181	0.385	35,539
CompExhib (10km)	28.678	54.49	35,539
Demand (10km)	0.735	0.053	35,539
Control variables			
Public	0.005	0.072	152,707
Semi-public	0.016	0.124	152,707
Highly-educated (10km)	0.214	0.040	152,707
Unemployed (10km)	0.278	0.034	152,707
Born in Catalonia (10km)	0.643	0.047	152,707
USA origin	0.439	0.496	152,707
Spain origin	0.259	0.438	152,707
Monday	0.123	0.328	152,707
Tuesday	0.115	0.319	152,707
Wednesday	0.126	0.332	152,707
Thursday	0.127	0.333	152,707
Friday	0.148	0.355	152,707
Saturday	0.175	0.380	152,707
Sunday	0.187	0.390	152,707

The final sample consists of 152,707 screenings (the screenings for which I can merge all the variables and that do not have missing information). Only 4.2% are in Catalan, which is in the range of the supply of films in Catalan in the last 15 years (see Figure 1). We can see that the proportion of screenings of films that have the two versions is just 23.3% (the dummy Catalan-Version for the final sample). Among those films that do have the Catalan version available,

18.1% of their screenings are in fact exhibited in Catalan. This shows that there is room for an increase in language diversity both at the distribution and the exhibition level.

3.5. Empirical Methodology

The final purpose of this work involves analysing whether the market imperfections lead to a lower provision of film screenings in the Catalan version. The aim of the empirical analysis is to explain the probability of offering a screening in Catalan (vs Spanish) as a function of proxies for the demand and market power of the distribution and exhibition. The econometric specification, thus, relies on the following reduced form, Equation (1):

$$Catalan_{s,f,m,t} = \beta_0 + \beta_1 ShareDist_{d(f)} + \beta_2 CompExhib_{s,f,m,t} + \beta_3 Demand_m + \delta' W_f + \gamma' W_m + \theta_{c(m)} + \theta_t + \varepsilon_{s,f,m,t} \quad (1)$$

Here, $Catalan_{s,f,m,t}$ is a binary variable that takes 1 if the screening *s* of film *f* in the movie theatre *m* at time *t* is in the Catalan version, and 0 otherwise; we do not include subtitled screenings. There are three parameters of interest. The first one (β_1) corresponds to the variable $ShareDist_{d(f)}$, which is the proxy of the distributor's market power and it is at the distributor *d* level, which depends on film *f*. The second relevant coefficient (β_2) corresponds to $CompExhib_{s,f,m,t}$, which is the proxy of exhibitor competitive pressure, as explained in the previous section, which depends on screening *s* of film *f* in the movie theatre *m* at time *t*. Finally, the third relevant coefficient (β_3) corresponds to the $Demand_m$, which is the percentage of the Catalan-speaking population within the specified radius (10 km) around the movie theatre *m*.

I also include several control variables and fixed effects in the model. Specifically, W_f and W_m are vectors of control variables at the film level and movie theatre level, respectively. As for film level controls, I include country of production as well as genre, and by movie theatre level we include the property type (private, public or semi-public). Moreover, with the aim of not

confounding the demand effect with the impact of other sociodemographic characteristics of the local consumers, I also control for the percentage of highly educated individuals, unemployment rate and percentage of individuals born in Catalonia at the local level (within 10 km), which are contextual factors that are likely to be correlated to the share of individuals who are able to speak Catalan. $\theta_{c(m)}$ are fixed effects at the chain *c* level, which depends on the movie theatre. The term and θ_t include time fixed effects: weekday and week. By including these fixed effects, I take into account the effect of price, which varies between chains and days, but it is fixed for a given cinema and day (all films at the same price).[34] This econometric specification is based on the demand model of Davis (2006), which defines the utility of a consumer who watches film *f* in the movie theatre *m*,[35] plus the supply side proxies and controls.

This reduced form includes in the same equation the market power of the distributor (upstream market) and the competition at the exhibition level as well as the local demand for Catalan versions (downstream market). We also use two separate equations for the two stages as an alternative specification, as will be explained in Section 5.3.

Due to the nature of my dependent variable (dummy), I use a Linear Probability Model in order to get a direct interpretation of the coefficients.[36] If the supply of films in the Catalan version was only driven by the demand, β_1 and β_2 should not be statistically different from 0, while β_3 should be positive and statistically significant. If market imperfection determined to some extent the supply of films in the Catalan version, as explained in Section 3, we should see that both β_1 and β_2 would be statistically different from 0. β_1 is in principle ambiguous, since the business-stealing effect might be offset by the bargaining effect. Nonetheless, we expect the former to be predominant, making the coefficient negative (the higher the market share, the lower the incentives to offer the Catalan version). β_2, in turn, should be positive (the higher the number of competing screenings, the higher the incentive to compete on the language version).

[34] I also tried a time-specific trend by chain and obtained very similar results. Available upon request.

[35] In Davis (2006) this is referred to as *h*.

[36] I also tried to estimate the same equation using non-linear binary choice models (probit and logit) and compute the corresponding average marginal effects. The results are reported in the Appendix.

The threats to the identification of β_1 and β_2 are due to unobserved (or not included) variables that can affect the probability that a screening is in the Catalan version and are correlated to $ShareDist_{d(f)}$ and $CompExhib_{s,f,m,t}$, respectively. The main threat to the identification of β_1 was that the original language of the movie was related to the distributor share of the market. Indeed, the so-called majors (big distributors) are often American and they distribute Hollywood movies in Spain, which tend to be blockbusters. We could potentially see that local distributors with a small share of the market tend to distribute films in which the original language is Catalan, while big distributors, instead, distribute foreign productions in which the original language is English. That would lead to an upward bias and β_1 could capture other things not directly related to the market power of the distributor. For this reason, I include the W_f, which specifically consists of two dummies for Spanish origin and US origin; hence, we rule out the possibility that β_1 captures other effects.

Regarding β_2, one of the main threats to its identification was the demand itself; indeed, without controlling for the proxy for the demand, there would be an omitted variable leading to a downward bias, since the demand and $CompExhib_{s,f,m,t}$ are spuriously correlated: the percentage of Catalan-speaking population is higher in rural areas, in which cinemas do not have competitors nearby. However, even controlling for demand-side effects and other sociodemographic controls at the local level, there could be other variables that, if not included, would lead to a downward bias, such as W_m, the property type of the cinema. Public and semi-public tend to offer a higher percentage of films in the Catalan version and they have a greater presence in less populated areas, in which there are no private cinemas even close by, and therefore the City Council tries to supply movies on its own (e.g. summer cinema outdoors).

Even if there could be some other unobserved factors that undermine the validity of the estimation, by including the time and chain fixed effects θ I make sure to control by all the unobserved heterogeneity related to time and chains. Moreover, since the price is fixed for a given cinema and day (it does not vary between films), the fixed effects will also capture the effect of the price.

Last but not least, standard errors are clustered at the cinema level since it is the level of variation of most of the aggregated variables of interest.

3.5.1. Heterogeneous Effects

After the analysis of the effect of the distribution and exhibition concentration on the supply of films in the Catalan version, using the data on the film genre, I can analyse the existing heterogeneous effects, that is, the effects of two different types of consumers: the children and the adults.

These two groups of consumers might differ on the linguistic preference for two reasons: there are no monolingual children since they learn both languages in school; and the intensit of the preferences are likely to be greater. Although children do not choose by themselves the films they watch, we can consider parents' decisions to be a good representation of their child's preferences; when a parent chooses to go to the cinema with her/his child, he or she tries to maximize the utility of the child. If the intensity is greater, cinemas will have a higher incentive to compete on the language version to attract more consumers for those movies targeted at children, that is, both the market expansion and the business-stealing effects will be greater. That would be further evidence of the market failure explained in Section 3.

In order to test for this heterogeneous effect, I use the dummy $Animation_f$, that takes the value 1 if the film's genre is 'Animation' and 0 otherwise. I assume that the 'Animation' genre is for films targeted at children and the others target the general adult audience. $Animation_f$ is interacted with *ShareDist*, *CompExhib* and *Demand* to disentangle the effect of these variables for adults and children separately.

3.5.2. Robustness Checks

In order to test the robustness of the results, several sensitivity checks are performed, both to the baseline model and to the model allowing for heterogeneous effects by the genre of the film.

As explained in Section 4.1., to check whether 10 km is an appropriate radius, we try several radii, both together and separately (5 km, the default 10 km, 15 km and 20 km). According to Davis (2006), 'geographic markets consist of at most 15-mile circles around theatres, probably less'. Also, we should keep in mind that the analysis was made in the US, which is more car-based than European countries, and thus such local markets might likely be smaller.

In addition to that, I also use a different definition of the relevant variables *CompExhib* and *Demand*. Recall that the boundaries of the local market were defined as a radius of 10 km from the location of the cinema, thus I could count the number of competing films in the same week (*CompExhib*) and the proportion of bilingual speakers among the population older than 2 years (*Demand*). Alternatively, I set the boundaries using a measure of accessibility, the driving time. More specifically, I set the boundaries of the local market to a 20-minute driving distance from the cinema. I computed this with the command 'georoutei' in Stata, which uses the Web mapping HERE WeGo. This new measure implicitly takes into account the geography, connections, as well as the rural/urban difference. Thus, *CompExhib* and *Demand* are recalculated.

Furthermore, since a 'preference towards the original' exists, and although it would be possible from a theoretical point of view, it is not a common practice to offer another version rather than the original for those films whose original language is Spanish or Catalan. Except for animation films, because they are not real actors and therefore do not have an original voice, I restrict the sample to only those films whose original language is a foreign language, and therefore in which no 'preference towards the original' might bias the results.

3.5.3. Extensions

The model relies on a strong, although realistic, assumption: each film has its own demand, they are not substitutes for each other (the language version cannot shift demand from one film to another). The extension I propose, which represents an additional robustness check, consists of assuming that similar films are substitutes. More precisely, films of the same genre are considered to be perfect substitutes. Hence, animation films will be competing in the same market, as well as drama, comedy and so on (see Table 6, all the genres in the data set).

The two competition measures at the two steps – distribution and competition – are redefined accordingly. That is, the share of the market of the distributor (*ShareDist*) is (re)computed as the market share within each 'genre', rather than for the whole market. If most of the screenings of animated films belong to a certain distributor, such a distributor will have a very high share even if it is non-present in other genres. Moreover, the number of competing films (*CompExhib*) is obtained considering all the films of the same genre offered by nearby competing exhibitors during that week.

The second extension, which also works as an additional robustness check, consists of separating the two stages in two equations. Thus, in the first step I look at the probability that a distributor offers the Catalan version of a film, therefore films are the observation units; and in the second step, I look at the probability that a screening is offered in Catalan in a movie theatre, restricting the sample to those films that do have a dubbed version (so at the first stage the distributor incurred the cost of dubbing in Catalan).

First step[37]:

$$Catalan_{Version_f} =$$

$$\beta_0 + \beta_1 ShareDist_{d(f)} + \delta' W_f + \theta_t + \varepsilon_{s,f,m,t} \quad (2)$$

Second step:

$$Catalan_{s,f,m,t} = \beta_0 + \beta_2 CompExhib_{s,f,m,t} + \beta_3 Demand_m + \delta' W_f + \gamma' W_m + \theta_{c(m)} + \theta_t + \varepsilon_{s,f,m,t} \quad if\ Catalan\_Version_f = 1 \quad (3)$$

3.6. Results

The results of the baseline model are reported in Table 3.7. In column (1) we observe that indeed there is a negative and significant effect of the market power of the distributor in the language diversity, but the coefficient of *CompExhib*, the β_2, is negative and significant. This is reversed once I include the *Demand* in column (2). Local markets with a higher degree of competition (urban areas) also have a lower percentage of bilingual speakers. By including *Demand*, such spurious correlation is driven out and the coefficient β_2 becomes positive, although not significant. In column (3) the controls at the film level, W_f (dummies of genre and country of origin), are included. β_2 leaps up in

37 θ_t are dummies for weeks since the initial period, not days. A film can be on screen for more than one week.

magnitude and significance, although remains non-significant even at 10%; β_3 remains stable, and β_1 increases its magnitude and significance. In column (4) the controls at the movie-theatre level, W_m (property type and sociodemographic characteristics of the local market), are included. The main change here is the increase of β_3, since previously it was downward biased by the correlation with local sociodemographic characteristics. Finally, the full specification is used in column (5) by including the fixed effects $\theta_{c(m)}$ (ownership dummies) and θ_t (weekday and week dummies); the coefficients remain quite stable.

Table 3.7: LPM, Baseline Model

	(1) Catalan	(2) Catalan	(3) Catalan	(4) Catalan	(5) Catalan
ShareDist	-0.323*** (-4.88)	-0.338*** (-4.99)	-0.476*** (-6.15)	-0.466*** (-5.97)	-0.425*** (-5.53)
CompExhib	-0.0121** (-3.15)	0.00259 (0.75)	0.00534 (1.53)	0.00575 (1.32)	0.00686 (1.49)
Demand		0.506*** (3.38)	0.506*** (3.74)	0.632** (2.91)	0.613** (2.90)
US origin			-0.0502*** (-4.57)	-0.0499*** (-4.50)	-0.0493*** (-4.43)
Spain origin			-0.00102 (-0.12)	-0.000275 (-0.03)	-0.00439 (-0.54)
Public				0.0416 (1.65)	0.0424 (1.54)
Semi-public				0.0914 (1.39)	0.0984 (1.58)
High-educ				-0.499** (-2.97)	-0.305 (-1.50)
Unemployed				-0.664*** (-3.99)	-0.336 (-1.54)
Cat-born				-0.501* (-2.17)	-0.441+ (-1.88)
Genre dummies	No	No	Yes	Yes	Yes
$\theta_{c(m)}$	No	No	No	No	Yes
θ_t	No	No	No	No	Yes
Const	0.0779*** (5.83)	-0.297** (-2.83)	-0.341*** (-3.40)	0.171 (1.22)	-0.0078 (-0.06)
N	152707	152707	152707	152707	152707

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib rescaled by 100.

Considering the full specification displayed in column (5) of Table 8, it appears that the market power of the distributor has a strong and significant negative effect on the supply of films in Catalan. For a 1 percentage point (p.p.) increase in the share of the distributor, the probability of supplying a film in Catalan will decrease by a 0.425 p.p.

In this specification, the competition at the exhibition level does not seem to have a significant effect. However, the results presented in the next subsection indicate that it indeed matters, depending on the target audience.

In order to get a better idea of the scope of the effect of the competition on the language version, we can compute a counterfactual. The monopolistic behaviour of the distributor is due to its market power. In fact, small distributors, with a share of the market close to zero, do not have such monopolistic behaviour. Thus, we can compute the total effect of the monopolistic behaviour at the distribution level by assuming a mean of *ShareDist* equal to 0 (hypothetical perfect competition):

$$\beta_1 * Mean\left(ShareDist_{d(f)}\right) - \beta_1 * 0 = 0.425 * 0.095 = 4.04 \text{ p. p.}$$

This means that the concentration at the distribution level leads to an undersupply of 4.04 p.p. of films in the Catalan version. Such a figure might be low at first sight, but considering that the percentage of films in the Catalan version in our sample is 4.2%, that implies that without such monopolistic behaviour the share of films in Catalan would be 96% greater (8.24% instead of the actual 4.2%).

With regard to the *Demand*, the interpretation of the coefficient β_3 is also simple: for a 1 p.p. increase of bilingual speakers in the local market, the propensity to offer a film in the Catalan version would be 0.613 p.p. higher.

3.6.1. Heterogeneous Effects

Table 3.8 reports the results of the analysis of heterogeneous effects. In the first column, the aggregated effects are displayed; they are the coefficients of the three relevant variables of the full specification of the baseline model (column (5) of Table 8). In column (2) we can see the coefficients for adults and children, respectively. As explained in Section 5.1., we interact *ShareDist*, *CompExhib* and *Demand* with the dummy *Animation,* which indicates whether it was an animated film or not, assuming that animated films are targeted at children. In column (3)

the equality tests are shown; they are all significant, meaning that $\beta^{children} \neq \beta^{adults}$ for none of the three relevant variables. As can be seen, the magnitudes are greater for the $\beta^{children}$. This should be interpreted as a clear sign of differentiated behaviour of firms (distributors and exhibitors), depending on the type of audience.

It is also important to note that $\beta_2^{children}$ is positive and significant. An increase in the competition at the exhibition level leads to an increase in language diversity but only for animated films. More precisely, for an increase of 100 competing screenings, the supply of movies in Catalan will increase by 4.06 p.p.

Table 3.8: Heterogeneous Effects

	(1) Aggregated effect	(2) Heterogeneous	(3) Equality test (p-value)
ShareDist	-0.425***		
	(-5.53)		
CompExhib	0.00686		
	(1.49)		
Demand	0.613**		
	(2.90)		
ShareDist - adults		-0.1835***	
		(-4.29)	
			0.0000
ShareDist - children		-1.157***	
		(-6.69)	
CompExhib - adults		-0.0012	
		(-0.69)	
			0.0001
CompExhib - children		0.0406***	
		(4.14)	
Demand - adults		0.341+	
		(1.73)	
			0.0000
Demand - children		1.864***	
		(5.65)	
N	152707	152707	

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib rescaled by 100. The aggregated effect refers to the full specification of the baseline model, column (5) of Table 8.

The market is clearly more sensitive to the child audience than the adult, suggesting that indeed children have higher intensity of preferences.[38] This, in turn, means that the undersupply of a Catalan version is greater for child-targeted movies.

[38] An alternative explanation could be that the parents have a higher intensity of preferences when choosing a film for their children rather than for themselves.

3.6.2. Robustness Checks

The decision on which radius should be applied to define the local market is based on Davis (2006), although his applies a shorter distance. To check whether the decision of 10 km was appropriate, I used different radii for the definition of the variable *CompExhib*: 5 km, 10 km (the default), 15 km and 20 km: first separately, and then together in two variables $CompExhib_{10km}$ and $CompExhib_{10km\text{-}20km}$ (the number of competing screenings from 10 km to 20 km away). Since the aggregate effect of *CompExhib* is not significant for none of the radii, Table 3.9 reports only the coefficients of children. Since the variable *Demand* is also affected by the change in radius, it is also reported.

Table 3.9: Different Radii

	(1)	(2)	(3)	(4)	(5)
CompExhib – children 5km	0. 0596**				
	(2.89)				
CompExhib – children 10km		0. 0406***			0. 0382**
		(4.14)			(2.73)
CompExhib – children 15km			0. 0293***		
			(3.77)		
CompExhib – children 20km				0. 0256***	
				(4.47)	
CompExhib – children 10km - 20km					0. 0126
					(0.90)
Demand – children 5km	1.3923***				
	(5.89)				
Demand – children 10km		1.8644***			2.1157***
		(5.65)			(4.89)
Demand – children 15km			1.7743***		
			(4.41)		
Demand – children 20km				1.7669***	
				(4.28)	
Demand – children 10km - 20km					0. 5951
					(1.10)
N	152707	152707	152707	152707	152707

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib rescaled by 100. The radii of the control variables *High-educ*, *Unemployed* and *Cat-born* are also adapted in each equation. In equation (5) they are included with radius 10km and from 10 to 20km, altogether.

In column (1) to (4) we can see that there is a decay of β_2 as we increase the size of the local market, while the *Demand* coefficient β_3 remains substantially stable. Once we look separately at the effect of the competition within 10 km distance and between 10 km and 20 km distance, we see that the competition in the outer circle does not have any effect, and the same applies for the *Demand*. This result is relevant for two main reasons. First, because it supports the decision of choosing 10 km as the baseline radius for the local market. Second, it shows that, unlike Davis (2006), the competition does not matter beyond 10 km; according to Davis (2006), 'geographic markets consist of at most 15-mile circles

[24 km approx..] around theaters, probably less'. We find an even stricter limit of such geographic markets. This might be due to the fact that the case study of Davis was conducted in the US, a more car-based country in which individuals are used to driving longer distances compared to European countries.

In Table 3.10, I present the two additional robustness checks for the aggregated effects, column (1) and (3), and for the heterogeneous effects, column (2) and (4). The restriction of the sample to films whose original language is neither Catalan nor Spanish (the two official languages in Catalonia) aims to avoid a 'preference towards the original' bias. In this case, hence, we make sure that the preference over the language version has nothing to do with other characteristics that might be correlated with the relevant variables. The results are very robust; β_2 is actually much higher and significant for the aggregated effect in column (1). The results of the heterogeneous effects in column (2) show that the coefficients are also very robust here. The equality tests are highly significant, meaning that $\beta^{children} \neq \beta^{adults}$ for the three relevant variables.

The redefinition of the boundaries, from 10 km to 20-minute driving distance, does not lead to different conclusions. The direction and significance of the coefficients in both column (3), for the aggregated effects, and column (4) for the heterogeneous effects, are the same as those we found using the 10 km radius. The magnitudes of the coefficients do not vary considerably. This is strong and robust evidence in favour of the results of Table 3.7 and Table 3.8.

Table 3.10: Robustness Checks

	(1) Foreign Languages	(2) Foreign Languages	(3) Redefined boundaries (20min drive)	(4) Redefined boundaries (20min drive)
ShareDist	-0.462***		-0.425***	
	(-5.14)		(-5.51)	
CompExhib	0.01686**		0.00634	
	(2.64)		(1.54)	
Demand	0.702**		0.452**	
	(3.05)		(2.60)	
ShareDist - adults		-0.0999+		-0.185***
		(-1.83)		(-4.29)
ShareDist - children		-1.119***		-1.168***
		(-7.25)		(-6.75)
CompExhib - adults		0.0064		-0.00051
		(1.15)		(-0.15)
CompExhib - children		0.0432***		0.0392***
		(3.90)		(4.13)
Demand - adults		0.3010		0.2008
		(1.39)		(1.33)
Demand - children		1.8408***		1.7376***
		(5.54)		(5.05)
N	95514	95514	152707	152707
Equality tests (p-values)				
ShareDist		0.000		0.000
CompExhib		0.001		0.000
Demand		0.000		0.000

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib rescaled by 100.

3.6.3. Extensions

In this subsection the results of the two extensions explained in Section 3.5.3. are presented.

Table 3.11: Extension 1. Films Are Substitutes if Same Genre

Extension 1	(1) Aggregated effect	(2) Heterogeneous	(3) Equality test (p-value)
ShareDist	-0.168***		
	(-5.79)		
CompExhib	0.0071***		
	(4.28)		
Demand	0.640**		
	(3.07)		
ShareDist - adults		-0.0886***	
		(-4.44)	
			0.0000
ShareDist - children		-0.8877***	
		(-6.37)	
CompExhib - adults		0.0043**	
		(2.96)	
			0.0181
CompExhib - children		0.0250**	
		(2.84)	
Demand - adults		0.368+	
		(1.90)	
			0.0000
Demand - children		1.818***	
		(5.64)	
N	152707	152707	

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib rescaled by 100.

Table 3.11 reports the results of the first extension. Here, *ShareDist* and *CompExhib* are recalculated, taking into account that films compete within each genre. The magnitude of β_1 and β_2 is indeed different, but the direction and significance of the coefficient support the hypothesis of the paper and prove the robustness of the previous results. The main difference in this new extension is that the competition at the exhibition level does not only affect the language diversity in animated films, but in all kinds of films. Thus, the previous results (in Tables 3.7 and 3.8) might provide lower bound estimates, since in this formulation adult-targeted movies are also affected. The equality tests are

significant for the three relevant variables, indicating that indeed the market is more sensitive to children than adults: $|\beta^{children}| > |\beta^{adults}|$.

Finally, I present in Table 3.12 the results of the two stages separately. As explained in Section 5.3., in the upstream market the distributor has to decide whether to incur the cost of dubbing to offer the Catalan version or not, for each film (not screening, this is why the film is the observation unit). In the downstream market, the exhibitors (the cinemas) have to choose whether to offer the version in Catalan or Spanish, for those films in which a Catalan version is available (and therefore the sample is restricted to screenings of films with a Catalan version available).

Let us first focus on the aggregated effects, in column (1) and (3). We can note that the coefficients are equal in significance and direction to the baseline model: a negative and significant effect of *ShareDist*, a positive and non-significant effect of *CompExhib*, and a positive and significant effect of *Demand*. Nonetheless, the effect of the competition at the exhibition level is again significant when we separate animated films from the rest, as we observed in Table 3.8. Both $\beta_2^{children}$ and $\beta_3^{children}$ are significantly higher than their adult counterparts, according to the equality tests. The lack of significance of $\beta_1^{children}$ and its equality tests is mainly due to the small size of the sample used (recall: films, not screenings). Among the 626 films of this sample, only 77 are animated films. These are too few observations to lead to significant results both in the $\beta_1^{children}$ and the corresponding equality test. However, if we look at the point estimate, we can note that $\beta_1^{children}$ is indeed higher than the β_1^{adults}. Overall, these second extension's results, reported in Table 3.12, provide evidence that the results found in Table 3.7 and Table 3.8 are consistent.

Table 3.12: Extension 2. Two Stages Separately

Extension 2	**Upstream market (1st stage)**		**Downstream market (2nd stage)**	
	(1) Catalan Version	(2) Catalan Version	(3) Catalan	(4) Catalan
ShareDist	-0.605**			
	(-2.94)			
CompExhib			0.0145	
			(1.03)	
Demand			2.043**	
			(2.57)	
ShareDist - adults		-0.5225**		
		(-2.66)		
ShareDist - children		-1.552		
		(-1.24)		
CompExhib - adults				-0.0136
				(-0.65)
CompExhib - children				0.0283*
				(2.08)
Demand - adults				1.554+
				(1.92)
Demand - children				2.463**
				(3.09)
N	626	626	35539	35539
Equality tests (p-values)				
ShareDist		0.416		
CompExhib				0.022
Demand				0.000

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib rescaled by 100. In the upstream market we consider films as the observation units. In the second stage, we restrict to those screenings which have a Catalan version.

3.7. Conclusions

This paper represents, to the best of my knowledge, the first attempt to empirically analyse the language supply of a cultural market. The aim of this work was to find whether the concentration of the movie-theatre industry, both at the distribution and the exhibition levels, led to an undersupply of language diversity in a bilingual context.

To do so, I built a unique data set, using advanced web-scraping techniques and combining different sources, of all the cinema screenings in the bilingual region of Catalonia to explore a potential market failure. Such failure would consist in an undersupply of films in the Catalan version due to the concentration in two stages of the market. Applying a Linear Probability Model controlling by relevant factors and adding several fixed effects, I found that the concentration of the market indeed leads to an undersupply of 4.04 percentage points of films in Catalan, which means that if those market failures did not exist there would be 96% more screenings in Catalan. This undersupply can be entirely attributed to the monopolistic behaviour of the distributors.

By using the genre of the film, I disentangle two types of consumers, children and adults. The market response differs if the film is targeted at children or at the general adult audience. Due to the higher intensity of the linguistic preference of the children compared to the adults, I found that the competitive pressure of the exhibitors only changes the supply of films in Catalan when it deals with the child demand (animation films), and the effect is negligible for the rest. These results were robust to several checks.

According to the standard theoretical framework I provide, the empirical evidence indicates that the movie-theatre industry underprovides language diversity. More specifically, the fraction of movies dubbed into Catalan is insufficiently low from a welfare point of view. Most of the inefficiency can be attributed to the upstream firms, the distributors, who act as a bottleneck. Such underprovision is higher in local markets with low competition when it comes to animated films, that is, films targeted at children; this additional effect can be attributed to the downstream firms. Therefore, a first policy implication that can be derived from these findings is that the focus should be put on the distributors, who in the first stage decide whether to incur the fixed cost of dubbing in order to offer an additional version in Catalan. Thus, subsidizing the costs of dubbing or establishing quotas of films distributed in Catalan would be

policies that can be justified from an efficiency point of view. Also, another policy implication that can be derived from the findings on the heterogeneous effects is that the welfare loss due to the low provision of films in Catalan is greater for animated films, so policies could be more tailored and focus on child-targeted films. For instance, exhibitors could be legally bound to offer the Catalan version at least in one of the screenings of an animated film.

The finding on the heterogeneous effects sheds light on another important issue: the endogeneity of the preferences. Indeed, if Catalan-speaking adults accommodate their preference more than children (lower intensity of preference), it indicates that the linguistic preference changes over time, at least in intensity. The drivers of this change are not the focus of the research of this paper, but it would be an interesting topic for further research on linguistic preference in cultural markets.

Appendix

Table 3.A1: Logit and Probit (Average Marginal Effects)

	(1) Logit	(2) Logit	(3) Probit	(4) Probit
ShareDist	-0.474*** (-7.33)		-0.4593*** (-7.45)	
CompExhib	-0.00052 (-0.13)		-0.00121 (-0.29)	
Demand	0.4247+ (1.95)		0.4615* (2.27)	
ShareDist - adults		-0.3885*** (-6.20)		-0.3924*** (-6.44)
ShareDist - children		-0.9115*** (-4.96)		-0.8853*** (-5.26)
CompExhib - adults		-0.0079** (2.66)		-0.00896** (-2.73)
CompExhib - children		0.0301*** (2.39)		0.0298** (2.63)
Demand - adults		0.220 (1.61)		0.2444+ (1.83)
Demand - children		1.425** (2.52)		1.5292** (3.10)
N	143042	143042	143042	143042
Equality tests (p-values)				
ShareDist		0.004		0.009
CompExhib		0.000		0.000
Demand		0.028		0.002

t statistics in parentheses + p<0.1, * p<0.05, ** p<0.01, *** p<0.001
Clustered standard errors at the cinema level. CompExhib escaled by 100. Aggregated effects in columns (1) and (3). Heterogeneous effects in columns (2) and (4).

4. How Have Video-on-Demand Platforms Shaped Our Preferences? Endogenous Preferences in a Cultural Market

4.1. Introduction

This work analyses whether the introduction of video-on-demand (VoD) platforms generated an endogenous change in users' preferences in terms of the different language versions of movies. The empirical setup exploits the launch of Netflix in Spain in late 2015, which was the starting point of the VoD expansion, by using repeated cross-section data about the preferred language version when watching a film, as well as the use of VoD.

There is a vast body of literature on endogenous preferences and habit formation. Bowles (1998) asserted that policies and constitutions do affect preferences and that this is crucial to welfare analysis; indeed, the welfare analysis should take into account that the policy itself can change the preferences of the individuals and thus the surpluses that were considered a priori. Rational choice theory defines how an individual can rationally behave, given his or her [exogenous] preferences, but says little about where those preferences come from (Dietrich & List, 2013). However, most of the studies on endogenous preferences are theoretical and focus on welfare analysis or how economic institutions determine the formation of the preferences. To the best of my knowledge, there has been little research that empirically analyses how preferences change and adapt to new products and paradigms.

This paper also contributes to another emerging strand of the literature on the effects of television on social phenomena. Television influences violent crime (Dahl & DellaVigna, 2009), voting turnout (Gentzkow, 2006), political patterns in the US (DellaVigna & Kaplan, 2007) and international policy (Eisensee & Strömberg, 2007). The study by Micola et al. (2019) is in this regard the closest work to this paper, since it considers the effect of the type of version on English skills attainment. Specifically, the authors examine the influence of television translation techniques on the worldwide distribution of English-speaking skills.

In this paper, the case of the new paradigm in the movie market, VoD platforms, is analysed. VoD platforms have become the main source of movie

consumption in the recent years, and experienced a boost due to the pandemic lockdown. Unlike cinemas, VoD platforms allow users to choose from different language versions, as well as subtitles. Cultural goods, such as movies, deliver their content to the consumers through a particular language; the language of these cultural goods will be a determinant factor in the decision of the consumer (Caminal, 2010). Hence, in this context, the language should be regarded as a relevant product characteristic.

The technological differences between VoD and movie theatres define the provision of each language version, including the original version with subtitles. If we focus on dubbing, VoD platforms face the same problem of movie theatres: the need to incur the dubbing cost to offer a film in more than one language. However, if a one-screen movie theatre wants to offer a movie in a minority language, then it needs to separate the audience by alternating different linguistic versions, and hence the local linguistic minority has to be sizeable. Similarly, in a multiscreen movie complex a minimum demand for a second language is also required in order to allocate an entire screen to a version in that language. If the linguistic minority is very small at each location, then no movie theatre may want to show the movie in the minority language. In contrast, because of its centralized nature, the VoD platform can still profit from serving the linguistic minority. Such a phenomenon was theorized by Chris Anderson in his book *Long Tail* (2013). According to the long tail theory:

For many product categories, smart technology is transforming mass markets into millions of small niche markets. Although each of these niche markets may be small, when all the various niches are combined, the volume of business is actually greater than the traditional mass market successes. Thus, the great commercial opportunity of the future isn't catering to the "short head" of the demand curve where multiple copies of the same product are sold. Instead, the real opportunity to move forward lies in serving the "long tail" of the same curve – the millions of various niche markets. (Anderson, C. 2013: p. 2)

Quan and Williams (2018) provided a recent empirical example of enhanced product variety thanks to the online market. Hence, in our case the "smart" technology of VoD could allow for increasing the linguistic diversity, especially for those languages that did not reach critical mass when movies were more often consumed in cinemas. Regarding subtitled original versions, the cost of them is negligible compared to dubbing, so technological change should imply a great increase in this type of version.

In reality, what we observe is not precisely that. On the one hand, the availability of original versions has indeed rocketed and virtually all the movies on VoD platforms can be watched in their original versions with many options for subtitles. On the other hand, minority languages are even less present in the VoD compared to the movie theatres. The reason for that might be that although the long tail effect works in favour of the provision of minority languages in VoD, the bundling effect makes the local markets and small niches irrelevant: since the user does not buy each film but the whole subscription (bundling), the globalized VoD platforms can overlook the long tail. Nonetheless, the reasons behind the actual supply of different versions in VoD is beyond the scope of this paper, which treats their provision of language versions as exogenous.

The sudden change in the availability of language versions in the movie market represents the ideal setup for exploring the endogeneity of users' preferences. In monolingual markets in which dubbing was (and is) the norm, such as Spain, France, Italy or Germany, VoD platforms have dramatically increased the exposure to original versions with subtitles (in their own language). In bilingual markets with minority languages, in turn, the increase in diversity is not comparable; while the availability of original versions with subtitles have clearly increased as well, access to local-language versions has been reduced. Moreover, as usually happens with innovations, there was no regulation on this matter until very recently in many countries.

In the European Union (EU), in late 2018 the European Parliament approved the first directive on the VoD platforms, which should be steadily applied by each member state in their respective parliaments. In the case of Spain, for instance, the EU directive was implemented in 2022 through the "Ley General de Comunicación Audiovisual" (General Law of Audiovisual Communication) that established quotas for production in the different official languages of Spain,[39] among other obligations. Overall, VoD platforms have allowed viewers to consume movies in foreign-language and subtitled versions, but may have reduced the diversity (at least in the first unregulated years) regarding local languages.

[39] Spanish is the official language of Spain among other regionally official languages: Catalan, Basque and Galician.

This case study considers the bilingual region of Catalonia in Spain. As explained, in bilingual regions the impact of VoD platforms have been twofold. Hence, this makes the study of this particular case more complex and richer. The main hypothesis to be tested in this paper is that a greater exposure to a language version increases the preference for that version. Consumers adapt their tastes to what is offered in the market. More specifically, I expect to find an increase in the preference for the original version and a decrease in that for the Catalan version. The effect on the preference for the Spanish versions is unclear, since the Spanish versions are still always present but, unlike the movie theatres, they are non-exclusive: the consumer cannot choose the language version in movie theatres, as it offered in one version or another. Instead, VoD can offer the full availability of two versions for all the films in the catalogue.

Furthermore, the competition between versions must be understood in two categories: the horizontal and vertical differentiations. The vertical differentiation refers to the quality difference between the two dimensions: dubbing vs subtitled original versions. If the exposure to VoD increases the preference for original versions, it should do so at the expense of the preference for dubbing, either Catalan or Spanish, or both. The horizontal differentiation refers to the language difference within the same dimension, the dubbing. Hence, if the exposure to VoD directly decreases the preference for Catalan versions, it should do so in favour of the Spanish version as long as the user does not switch dimension. Therefore, the preference for Spanish might go either way, depending on which effect is stronger.

To analyse the effect of VoD platforms, I use survey data from 2014 to 2019 containing information about language version preferences and the use of VoD platforms, among other relevant characteristics. By using probability models with different specifications and checks, I find a strong positive effect of VoD on the preference for original versions. The effect on both Spanish and Catalan is negative but only significant for some specifications; the heterogeneous analysis shows how the effect greatly depends on individual characteristics.

The rest of the paper is structured as follows. In the next section, I provide some background relating to the audiovisual market in Catalonia. After this, in Section 3, the data used for the empirical analysis are described. In Section 4, the empirical method is presented. In Section 5 I display the results and discuss them, and finally, in Section 6, I conclude this paper with some final comments.

4.2. Background

Before the introduction of VoD platforms in Spain, there were two main sources of cinema consumption by most of the population: TV and movie theatres. In TV, following the definitive change from analogue to digital terrestrial television in 2010, although by default the language movies are shown in is the language of the broadcaster, the original version (if different) can be chosen, as well as adding subtitles. The reality is, though, that this is not a common practice. Hence, we can take the percentages of TV audiences by language as a good indicator of the market size of each language in television. In Catalonia, the share of annual time watching TV in Catalan was approximately 20% (and thus approximately 80% for Spanish) from 2017 to 2020 (Statistical Institute of Catalonia [IDESCAT]).

On the other hand, in the movie theatres of Catalonia the Catalan language was less present, only representing approximately 3%–4% of the total number of screenings (IDESCAT). Since in Spain dubbing is the norm, most of the rest of the screenings were in Spanish, either original or dubbed, although in the last years the presence of original versions has increased. There are no official data on the percentage of screenings in the original versions (with subtitles) in past years, but they represented approximately 11.5% of the total number of screenings in Catalonia in 2020 and 2021 (Mallén, 2023).[40]

As explained in the introduction, the Catalan language has a much lower presence on the globalized VoD platforms than in the other sources of audiovisual consumptions previously described. In May 2020, the Pro-Language Platform (in Catalan, *Plataforma per la Llengua*), a non-governmental organization concerned with the situation of the Catalan language, issued the first report analysing the availability of the Catalan versions of movies on Netflix (the most consumed VoD platform). In this report, they showed that only 0.02% of the movies in the Netflix catalogue had a Catalan-language version available, even though 16% already had a version in Catalan (so Netflix should not need to incur the cost of dubbing or subtitling[41]). One year later, the Audiovisual Council of Catalonia (*Consell Audiovisual de Catalunya*), an independent regulatory authority, published a similar study showing that the percentage of movies with

[40] This comes from own-collected data using webscraping techniques. I tracked all the screenings in movie theatres in Catalonia from June 2020 to May 2021. See chapter 3.
[41] The dubbed versions are freely available upon request to Catalan public television, which is the owner.

a Catalan version available on Netflix was only 0.5%. It is important to note that the percentage of screenings in movie theatres and the percentage of films with the version available on VoD platforms are not the same indicator. For instance, during 2020–2021, while screenings in Catalan made up 4.7% of the total, this meant that almost 18% of the movies were also available in Catalan, although this version was exclusive, unlike VoD.

While the availability of Catalan-language versions has decreased in this new source of audiovisual consumption, VoD platforms, the availability of original versions has sharply increased. Actually, it is possible to watch all the movies in their original version (with subtitles, if needed).

To sum up, the introduction of Netflix decreased the exposure to the Catalan language from 20% in TV and 3%–4% of the screenings (or 18% of the films) in movie theatres to 0.2%–0.5% in the online catalogue, while it increased the availability of original versions from 0% in TV and approximately 10% in movie theatres to 100% in the online catalogue. The exposure to Spanish did not experience any considerable change: this language version is still fully available in the online catalogue.

4.3. Data and Descriptive Statistics

The data used in this paper come from the Òmnibus surveys that are periodically carried out by the Opinion Studies Center (in Catalan, *Centre d'Estudis d'Opinió*), for the years 2014, 2015, 2018 and 2019. These surveys asked individuals a wide set of questions including their language preferences for different cultural goods (books, cinema, theatre and music). Moreover, they include information on several sociodemographic and economic characteristics such as age, gender, language (mother tongue, language of identification and language of use), education, country of birth, self-reported income, social class, labour status and the size of the municipality of residence, among others.

In addition to that, the waves after the introduction of Netflix (2018 and 2019[42]) asked about the use of VoD platforms. The number of observations per year can be seen in Table 4.1.

[42] Netflix was introduced in October 2015 and the 2015 survey was carried out in March.

Table 4.1: Observations Per Year

Years	Observations
2014	4,800
2015	1,050
2018	1,200
2019	1,200
Total	**8,250**

The sample used for the analysis is restricted to those individuals who did answer all the relevant questions: preference for language in cinema, mother tongue, labour status and education level. Thus, "do not know" or "do not answer" are treated as missing values, and therefore excluded (205 observations dropped). The final sample consists of 8,045 respondents. Table 4.2 displays the summary statistics of the estimation sample.

Table 4.2: Summary Statistics Control Variables

Variable	Obs	Mean	Std. dev.
VoD	8,045	0.060	0.238
Mother tongue			
Catalan	8,045	0.412	0.492
Spanish	8,045	0.499	0.500
Catalan and Spanish	8,045	0.040	0.197
Income			
Up to 1000€	8,045	0.156	0.363
Between 1,000€ and 2,000€	8,045	0.354	0.478
Between 2,000€ and 3,000€	8,045	0.219	0.413
Between 3,000€ and 4,000€	8,045	0.091	0.288
Between 4,000€ and 5,000€	8,045	0.0286	0.167
More than 5,000€	8,045	0.023	0.151
Does not know/does not answer	8,045	0.128	0.334
Class			
Low	8,045	0.106	0.308
Middle-low	8,045	0.288	0.453
Middle	8,045	0.517	0.500
Middle-high	8,045	0.062	0.241
High	8,045	0.005	0.069
Does not know/does not answer	8,045	0.022	0.146

Age	8,045	49.45	17.105
Sex	8,045	0.513	0.500
Place of birth			
Catalonia	8,045	0.680	0.467
Rest of Spain	8,045	0.234	0.424
EU	8,045	0.042	0.201
Non-EU	8,045	0.044	0.205
Municipality size			
Below 50,000 inhabitants	8,045	0.549	0.498
Between 50,000 and 150,000	8,045	0.214	0.410
Between 150,000 and 1M	8,045	0.084	0.278
More than 1M	8,045	0.153	0.360
Labour status			
Works	8,045	0.479	0.500
Does not work	8,045	0.500	0.500
Temporarily on leave	8,045	0.021	0.143
Education			
Primary education or less	8,045	0.118	0.323
Secondary education	8,045	0.227	0.419
Professional course	8,045	0.378	0.485
Tertiary education (university degree, master's or PhD)	8,045	0.277	0.448

It is important to note that the Òmnibus surveys only include people with Spanish nationality; that is why we observe low percentages of individuals born outside of Spain (both EU and non-EU), relative to official statistics. In addition, due to this restriction the percentages of Catalan-speaking respondents are higher than the average when non-nationals are included. According to the Survey on Language Uses of the Population (*Enquesta d'Usos Lingüístics de la Població*) of 2018, for 31.5% of the population of Catalonia (including non-nationals), Catalan is their mother tongue, while for 52.7% it is Spanish; this is due to non-nationals from Spanish-speaking countries.

Only 6% of individuals in the sample are VoD users. This is normal since all the observations for 2014 and 2015 are necessarily non-users. In order to obtain a better idea of the exposure to VoD, Table 4.3 reports the proportion of users in 2018 and 2019. Moreover, the dependent variables relating to preferences are reported. In Table 4.4 the preferences on movies are reported by year. We can observe some changes during this period: a decrease in the preference for Catalan versions and in increase in the preference for Spanish versions. The

empirical strategy, explained in the next section, allows for a ceteris paribus estimation of these effects.

Table 4.3: Summary Statistics Dependent and Independent Variables

Variable	Obs	Mean	Std. Dev.
Preference for movies			
Catalan	8,045	0.236	0.425
Spanish	8,045	0.419	0.493
Original version	8,045	0.153	0.360
Indifferent	8,045	0.187	0.390
Preference for books			
Catalan	8,045	0.338	0.473
Spanish	8,045	0.384	0.486
Original version	8,045	0.071	0.257
Indifferent	8,045	0.194	0.395
VoD			
2018	1,175	0.191	0.393
2019	1,175	0.221	0.415

Table 4.4: Movies Preference by Year

	2014	2015	2018	2019
Catalan	0.283	0.224	0.162	0.134
Spanish	0.376	0.454	0.476	0.501
Original Version	0.134	0.170	0.193	0.174
Indifferent	0.203	0.152	0.163	0.180
VoD	0	0	0.191	0.221
N	4,664	1,031	1,175	1,175

Means reported for the selected sample.

4.4. Empirical Methodology

This section describes the empirical approach followed to analyse the effect of the exposure to VoD platforms on the changes in the language version preferences. The identification strategy exploits the irruption of Netflix in the late 2015 in Spain as a game changer in the movie market; from this point on, consumers could be exposed to another growing source of movies consumption

that had a different provision of versions than the previous ones, which were the cinema and the TV. Thus, while the individuals of the sample for the 2014 and 2015 surveys did not have the chance to be exposed to VoD, those in 2018 and 2019 did.

The main econometric specification is the following:

$$P(CP_i = j) = \alpha + \beta VoD_i + \gamma' X_i + \phi_t + \varepsilon_i \qquad (1)$$

Here, cinema preference (CP_i) represents the outcome under investigation of individual i for version j. I use a dummy variable for "Catalan", "Spanish", "Original version" and "Indifferent", respectively, in separate models (one regression for each dependent variable). VoD_i is the independent variable that takes a value of 1 if the individual i is a user, and 0 otherwise (note that this dummy takes a value of 0 for all respondents in the waves of 2014 and 2015). X_i is a vector of control variables: dummies for mother tongue or first language, the polynomial of age, sex, place of birth (Catalonia, rest of Spain, rest of EU and non-EU), size of the municipality, labour status and education level. The term ϕ_t contains year dummies, capturing the evolution of preferences towards the cinema over time. I start with a baseline model with only the exogenous variables and then the other variables are added.

There are two potential issues with the exogeneity of the VoD_i dummy. The first is the omitted variables that were correlated with both VoD and the preference. The year dummies are crucial to avoid a bias due to changes in preferences over time, that could cause the beta to capture some spurious correlation. On the other hand, the vector of control variables contains all the relevant socioeconomic characteristics, as well as cultural traits, so as to avoid this kind of bias. Hence, the reliability of the beta is based on the assumption that both X_i and ϕ_t make the ε_i uncorrelated with the VoD dummy.

The second potential issue is the reverse causality: that is, the preference in relation to the language version of movies could somehow be driving the decision to be a VoD user. There is no straight way to know whether this occurs, but it is likely to be the case (at least for some individuals).

In order to make sure that the coefficient is capturing the causal effect of the VoD exposure on language preferences, I perform a placebo test with another cultural good that is not related to the cinema: books. If the only relevant causal

chain is that the exposure to VoD affects language preferences for that specific good (and it does not capture unobserved factors or reverse causality), exposure to VoD should be unrelated to language preferences towards other goods that have nothing to do with VoD. Conceptually, the placebo test works when the sample used as the placebo is not affected by the intervention, in this case the VoD exposure. Here, I am running a placebo test with the same sample but a different outcome. Hence, the crucial point is that if there was some sort of language preference change that was spuriously correlated with VoD exposure, it should also affect the language preference for books. Therefore, finding a significant relationship between VoD exposure and the language preferences in terms of books would be suggestive that the β coefficient from the main regression is likely to be picking up unobserved factors or reverse causality, invalidating the causal interpretation of the results.

Indeed, the main reason why I pick the language preference for books as the placebo, and not theatre or music preference (for which I also have information), is that while theatre and music are cultural goods that share features with movies, books are a sufficiently different cultural good. At least, it is not evident that VoD exposure could affect such preferences. Moreover, there is an additional reason why I prefer to use books as the placebo. All the surveys first asked about the language preference for books, followed by cinema, theatre and finally music. Hence, the answer relating to books is less likely to have been conditioned by the answer for cinema, which was asked afterwards. In the case of theatre and music, since these questions were asked immediately after the cinema preference, they could have been conditioned by the previous answer.

As an additional check to understand whether the coefficient of interest can be plausibly interpreted as the causal effect of VoD use on the endogenous change in language preferences, I also perform an instrumental variable (IV) approach exploiting the information about self-reported internet quality at home, on a scale from 0 (no internet) to 10 (excellent connection). Such a question was not asked in the surveys for 2014 and 2015. Therefore, this variable is equal to 0 for observations from the first two waves. However, this is in line with the fact that the VoD dummy varies only for individuals included in the 2018 and 2019 waves, for which the endogeneity issue could be relevant. In order to be a valid instrument, the variable capturing internet quality must satisfy two conditions. On the one hand, it should be a strong determinant of the use of VoD platforms

(the relevance condition), which is a testable hypothesis. On the other, it should not directly affect the preferences towards the languages of movies. In order to test this hypothesis, I use dummies for internet quality to run the overidentification test.

The first stage is:

$$VoD_i = \omega + \mu IntQ_i + \delta' X_i + \theta_t + u_i \quad (2)$$

In which *IntQ* is the internet quality measured from 0 to 10. I expect μ to be positive and significant, since in order to be a VoD user having a good internet connection at home is a must. In many cases, connection quality is determined by the infrastructure, which does not depend on individual decisions or characteristics. If this is the case, the instrument will not only be relevant but also exogenous, satisfying the exogeneity condition.

However, one may argue that the quality of internet at home could reflect other individual characteristics that might directly affect language preferences, such as social status or cultural level. Nonetheless, we control for several demographic and socioeconomic variables that are likely to pick up these unobservable factors, meaning that the variable capturing internet quality can be plausibly considered conditionally exogenous with respect to language preferences, thus satisfying the exogeneity condition.

Additionally, in order to test for the validity of the instrument, I estimate reduced-form equations for the four cultural goods (cinema, books, theatre and music), replacing VoD with the internet quality. Finding no reduced-form effects of internet quality on language preferences for other cultural goods than moves would be a strong argument in favour of the validity of the instrument.

In addition, I estimate a model in which the VoD effect is interacted with time dummies. This model is repeated for each cultural good and preference and shown graphically. The main insight of this alternative way of presenting the results is to show that the trends persist in 2018 and 2019 only for the cinema preference.

4.4.1. Robustness Checks

In order to further test the stability of the results, I perform several robustness checks. First, I exclude those with a native language that is neither Catalan, nor

Spanish (nor both). These individuals are grouped in the category "others", which is a very mixed category of all other languages spoken in Catalonia, from Occitan to Arabic. This is not a homogeneous group, which makes the interpretation of the coefficient unclear. For this reason, I repeat the main estimation after excluding them from the sample. I also applied another restriction of the sample similar to this one: the inclusion of only those born in Spain.

Another robustness check consists of excluding from the sample the two last waves of surveys made in 2014. There is a concern that 2014 is overweighed in the estimation sample since the survey was carried out three times in that year and therefore, I take more observations (see Table 1). This is due to the fact that for the subsequent years, 2015, 2018 and 2019, I am using only one wave for each one, while for 2014 there were three waves, 2014-1, 2014-2 and 2014-3; hence, I exclude 2014-2 and 2014-3, obtaining a more balanced sample in terms of years. Finally, I add some subjective controls: the dummified self-reported class and income.[43] As an additional robustness check, I include in the Appendix the average marginal effects of the main results using non-linear Logit and Probit models.

4.4.2. Heterogeneous Effects

After presenting the results of the placebo tests and the robustness checks, which provide evidence in favour of the interpretation of the exposure variable as "conditionally exogeneous", I provide additional results regarding the heterogeneous effects of the VoD exposure. Since this paper analyses the endogeneity of the language preferences, the effects may differ depending on the linguistic group (by mother tongue). As reported in Table 4.2, there are two main linguistic groups in Catalonia, Catalan speakers and Spanish speakers, as well as a third mixed group of individuals who have both languages as their first one, which is smaller in size. There are other minorities such as Occitan speakers or migrants who speak foreign languages, but due to the restriction of the survey to individuals with Spanish nationality, they are quantitatively too few in the sample (and too diverse) to consider them an additional separate group.

Thus, in this last part of the econometric analysis I will look at the effect of the VoD exposure to the language preferences of the three linguistic groups

[43] Including the "Do not know" category as a dummy.

according to their mother tongue or first language. Finally, I also analyse whether the effect of VoD exposure is heterogeneous according to the other sociodemographic characteristics of the respondents, namely gender, education level and age group.

In order to do so, I use interactions of the independent variable *VoD* with the classifying variables first language (four categories), sex (two categories), education level (four categories) and age group (three categories). To assess the significance of the difference between the coefficients, an equality test is undertaken for each one.

4.5. Results

The main results are displayed in Table 4.5, which reports the estimates of the main coefficient of interest from Equation (1) for each of the possible outcomes obtained from different specifications (see the full results in Table 4.A1.2 of the Appendix). In the first row, we can observe the basic regression with no controls (except year dummies). The relationship between the independent variable *VoD* and the version preferences are significant in all cases: positive for the original version and negative for the others. However, once the individual exogenous characteristics are taken into account, the effect on the preference for Catalan and on the probability of being indifferent with respect to the language versions of movies vanishes, while the effect on the Spanish and original versions remain stable in terms of sign and are highly significant. The introduction of the rest of the controls does not change the significance of the estimates, although they decrease in magnitude, especially in terms of preference for original versions. The exposure to VoD platforms reduces the probability of preferring Spanish versions by 6.26 percentage points (p.p.), while it increases the likelihood of a preference for the original version by 7.86 p.p. The result regarding preferences for original versions is in line with the initial hypothesis. Moreover, the a priori ambiguous effect on the preference for Spanish versions appears to be negative: the increase in variety has led to a switch from the Spanish to the original versions for some users. Nonetheless, the lower provision of Catalan versions has not led to a significant decrease of such a preference. I further discuss this specific result in Section 4.5.2.

In addition to the main evidence, Table 4.A1.2 in the Appendix shows the estimated coefficients for control variables, which are worth commenting on. The correlation of the first language and the language preferences in columns (1) and (2) is in line with the expectations; since the reference category is Catalan speakers, we observe that Spanish speakers prefer Spanish versions to Catalan ones to a great extent, and the same happens to a lower extent for individuals who have both Catalan and Spanish as their first languages (mixed use in the family). What is somehow unexpected is that Spanish speakers are less inclined to prefer the original versions than Catalan speakers. One potential reason could be that, since the availability of Catalan versions is low in cinemas and TVs compared to Spanish versions (not as low as on Netflix and VoD platforms), Catalan speakers have become less accustomed to versions shown in their own language and switch more easily to original versions. It is also worth commenting that individuals who have both Catalan and Spanish as their first language are significantly more indifferent in terms of preferences, which is an expected result since the emotional attachment does not incline the balance towards any of the languages.

The effect of age[44] is quadratic, and, surprisingly, women are less inclined to Catalan versions while they prefer Spanish versions; women are also more indifferent than men. As expected, being born outside Catalonia is negatively correlated with the preference for Catalan versions, but only those born in the rest of Spain have a greater preference for Spanish versions. Foreigners from other EU countries or outside the EU show a greater predilection for original versions. In order to understand the conditional correlation of the municipality size with the dependent variables, it is important to distinguish three main levels with different demolinguistic landscapes: small municipalities are present in inner Catalonia, in which most of the inhabitants are Catalan speakers; medium-sized municipalities, with more than 50,001, are mostly present in the metropolitan area of Barcelona, and most of the inhabitants come from other Spanish-speaking regions, or are second- or third- generation internal migrants; lastly, Barcelona, the capital and largest city of Catalonia (1.6M inhabitants), is a cosmopolitan metropolis in which Catalan and Spanish are present along with other languages. Hence, the results indicate that being in a medium-sized municipality decreases the preference for Catalan while increasing the preference for Spanish, with little or no effect on the preferences for original

[44] I also attempted the model with a polynomial of three, but the cubic term was not significant.

versions and, if anything, makes an individual less indifferent; living in Barcelona decreases both the predilection for Catalan and Spanish versions while strongly increasing the preference for the original version. People who do not work have a greater preference for Spanish versions at the expense of Catalan ones, while they do not show a difference in terms of the original version.

Finally, as the education level increases, the predilection for the Catalan versions and especially for the original versions increases at the expense of Spanish versions, whose preference is strongly reduced.

Table 4.5: Main Results

Controls	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
No	-0.0520***	-0.0779***	0.176***	-0.0425**
	(0.016)	(0.025)	(0.023)	(0.018)
Only exogenous	-0.0228	-0.0905***	0.113***	0.00426
	(0.016)	(0.024)	(0.023)	(0.018)
All	-0.0204	-0.0626***	0.0786***	0.00754
	(0.016)	(0.024)	(0.021)	(0.018)
N	8045			

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Year dummies included. Coefficient beta of the variable VoD

As explained in Section 4.4, a placebo test if performed by looking at the effect of the VoD exposure to the language preferences in books. The results reported in Table 4.6 indicate that the point estimates of the placebo coefficients are generally very close to zero and are not statistically significant for any outcome, which speaks in favour of the validity of the approach followed.

Table 4.6: Placebo Test. Preferences for VoD and Books

Books	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
VoD	0.000635	-0.0144	0.0164	0.00439
	(0.019)	(0.023)	(0.015)	(0.020)
N	8045			

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Dependent variable is the language preference in books

As also explained in Section 4.4, in order to further investigate whether the main results are potentially affected by any endogeneity bias, I use the connection quality at home to instrument the independent variable *VoD*. Table 4.7 reports the coefficients from the first stage (lower panel) and for the structural equation for each possible outcome. Note that the N is now 8,043 because I had to exclude two individuals who did not answer the question regarding "internet quality".

Table 4.7: IV Using Connection Quality

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
VoD	-0.128	-0.271**	0.215**	0.194
	(0.098)	(0.137)	(0.091)	(0.120)
First Stage				
IntQ			0.0208***	
			(0.002)	
N			8043	

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01
IV with full specification

The first stage proves that internet quality is indeed a relevant instrument, since it is strongly correlated with *VoD*. The coefficients in columns (1) to (4) lead us to the same conclusion, that the effect of *VoD* is negative and significant for the preference for Spanish versions, but positive and significant for the preference for original versions. In column (1) we can see that although negative, the coefficient is still non-significant. Nonetheless, the magnitude of the coefficients is much higher in the IV estimation. This tells us that the coefficients found in Table 5 might be biased towards zero and are thus a lower bound of the actual effect. Another possible explanation for the higher coefficient is that the IV approach captures the effect among those who are induced to use VoD because they have a better quality of internet (the compliers), not among all the treated.

Table 4.8 reports the estimates from the reduced forms with all the cultural goods, to assess the validity of the instrument. Note that the coefficients are

only significant for the cultural good under investigation, Cinema, in the second and third column (which is in line with the main results and with the IV estimates). However, no clear relationship is detected between internet quality and preferences towards other cultural goods.[45] This evidence is again suggestive that the main results are capturing a causal relationship between VoD exposure and preferences towards movies' language.

Table 4.8: Test of the Validity of the Instrument "Internet Quality"

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
IntQ - Cinema	-0.003	-0.006**	0.004**	0.004
	(0.002)	(0.003)	(0.002)	(0.003)
IntQ - Books	0.001	-0.002	0.001	0.002
	(0.002)	(0.003)	(0.001)	(0.002)
IntQ - Theatre	-0.000	-0.002	0.002	0.003
	(0.002)	(0.003)	(0.002)	(0.003)
IntQ - Music	0.003*	-0.004	0.003	-0.002
	(0.002)	(0.003)	(0.003)	(0.003)
N	8043			

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Reduced form IV

In Table 4.9 I also show an additional estimation in which I use dummies for internet quality as instruments, which provides an overidentified model that enables testing for overidentification. The corresponding null hypothesis is not rejected, which is indicative that the hypothesis of excludability of the instrument is satisfied.

[45] The coefficient on the preference for music in Catalan, first column and last row, is only weakly significant at the 10% level.

Table 4.9: Over-Identification test with Dummified IntQ

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
VoD	-0.103	-0.252**	0.207***	0.160
	(0.088)	(0.120)	(0.076)	(0.107)
Over idenfitication test (p-value)	0.609	0.091	0.455	0.847
N	8043			

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Over-indentified IV with full specification. The instruments are the dummified variable *IntQ* . Wooldridge's score test of overidentifying restrictions, which is robust to heteroscedasticity.

In Figures 4.A1, 4.A2 and 4.A3 in the Appendix, the predictive margins during the period analysed are shown for the four cultural goods (cinema, books, theatre and music), as explained in the empirical methodology section. The graphs are not very conclusive; we only observe a significant difference in both 2018 and 2019 in the case of the preference for the original version in cinema, which is still strong evidence in favour of this particular outcome.

4.5.1. Robustness Checks

The results of the robustness checks explained in Section 4.1 are reported in Table 4.10.

Table 4.10: Robustness Checks

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
Excluding L1 Others				
VoD	-0.0293*	-0.0378	0.0662***	-0.00375
	(0.017)	(0.025)	(0.022)	(0.019)
N			7658	
Excluding 2014-2 and 2014-3				
VoD	-0.0151	-0.0495**	0.0674***	0.00225
	(0.016)	(0.024)	(0.021)	(0.018)
N			5130	
Excluding Born outside Spain				
VoD	-0.0320*	-0.0292	0.0493**	0.0060
	(0.0188)	(0.0267)	(0.0233)	(0.0212)
N			7353	
Additional controls: self-reported class and income				
VoD	-0.0217	-0.0618***	0.0783***	0.00804
	(0.0156)	(0.0237)	(0.0210)	(0.0183)
N			8045	

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01
Full specification applied.

The exclusion of the speakers of non-official languages slightly changed the coefficients. On the one hand, the effect of *VoD* on the preference for Catalan versions is now weakly significant and negative. On the other hand, the coefficient of the second column is no longer significant, meaning that the exposure to VoD might not have any effect on the preference for Spanish versions. However, the coefficient in column (3), the effect of *VoD* on the preference for original versions, is still very significant and positive. This stresses the importance that the main expected effect on original version is very stable. I provide further explanations and interpretations of these differences in Section 4.5.2, when analysing the heterogeneous effects by first language.

The second robustness check consisted on an even stricter restriction. As we can see, excluding the speakers of non-official languages led to a sample of 7,658 individuals, and the exclusion of the two last waves of surveys of 2014 (2014-2 and 2014-3) leads to a sample of 5,130 individuals. Nonetheless, the results do not vary much from Table 4.5 and the coefficients in columns (2) and (3) are

still significant. Hence, the main results are not driven by an overrepresentation of individuals interviewed in 2014.

To further check the results, those born outside Spain were excluded. As explained in Section 4.4.1, this sample restriction works similarly to the first robustness check and thus, the outcomes are also very similar to the first robustness check. This is not surprising since the individuals excluded in both restrictions might be the same in a large proportion: those with a foreign native language are likely to be born outside of Spain. In addition, we do not observe significant changes compared to the main results in Table 4.5 when we add additional controls such as self-reported class and income.

Finally, in the Appendix the average marginal effects of the main results using non-linear Logit and Probit models are reported in Table 4.A2. The effects are robust. As in two of the previous checks, the effect of VoD on the Catalan preference is weakly significant. Still, we cannot find a consistent negative effect on Catalan, as predicted in the hypothesis, but this points out that indeed VoD exposure might be affecting such preference somehow. This is clearly seen in the next section.

4.5.2. Heterogeneous Effects

The next table (Table 4.11) reports the results of the heterogeneous effects of the VoD exposure by linguistic group, gender, education level and age group. As explained in Section 4.4.2, I use interactions of the explanatory variable *VoD* with the classifying variables and report the p-value of the equality tests.

Table 4.11 Heterogeneous Effects by Linguistic Group, Gender, Education Level and Age Group

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
VoD. L1 Catalan	-0.100***	-0.0171	0.0796**	0.0330
	(0.036)	(0.037)	(0.037)	(0.031)
VoD. L1 Spanish	0.0166	-0.0453	0.0713***	-0.0393*
	(0.014)	(0.033)	(0.027)	(0.022)
VoD. L1 Catalan-Spanish	-0.0727**	-0.0161	-0.000618	0.0440
	(0.030)	(0.108)	(0.101)	(0.094)
VoD. L1 Others	0.0511	-0.264***	0.132**	0.118**
	(0.036)	(0.061)	(0.060)	(0.054)
Equality test (p-value)	0.000	0.004	0.6856	0.0176
VoD. Male	-0.0175	-0.0469	0.0806***	-0.0130
	(0.020)	(0.032)	(0.028)	(0.022)
VoD. Female	-0.0240	-0.0813**	0.0762**	0.0320
	(0.021)	(0.032)	(0.030)	(0.027)
Equality test (p-value)	0.814	0.421	0.909	0.170
VoD. Primary education or less	0.0297	-0.00811	-0.0173	0.0129
	(0.038)	(0.070)	(0.043)	(0.057)
VoD. Secondary education	-0.0416	-0.00191	0.0884*	-0.0366
	(0.030)	(0.057)	(0.050)	(0.034)
VoD. Professional course	-0.00962	-0.0491	0.0370	0.0285
	(0.025)	(0.036)	(0.031)	(0.029)
VoD. Tertiary education	-0.0388	-0.117***	0.150***	-0.00104
	(0.024)	(0.037)	(0.038)	(0.028)
Equality test (p-value)	0.371	0.232	0.018	0.497
VoD. 18-39 years old	-0.00797	-0.102***	0.0994***	0.0142
	(0.018)	(0.032)	(0.029)	(0.021)
VoD. 40-64 yearls old	-0.0354	-0.0115	0.0652**	-0.0112
	(0.024)	(0.034)	(0.030)	(0.028)
VoD. 65 or older	-0.0333	-0.0608	-0.0703	0.117
	(0.100)	(0.097)	(0.055)	(0.118)
Equality test (p-value)	0.621	0.118	0.022	0.477
N		8045		

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Full specification applied

The p-value of the equality tests show which coefficients are significantly different depending on the classifying variable, that is, that there are heterogeneous effects on this regard. The first language seems to matter when it comes to the effect of the VoD exposure on the preference for Catalan and Spanish versions, columns (1) and (2), as well as the indifference in column (3) to a lesser extent (the p-value is not as low). The high p-values of the equality tests performed on the impact of VoD by gender prove that there are no differential effects depending on whether the individuals are males or females. On the other side, we observe low p-values and hence significant differences between the coefficients in column (3) for the two last classifying variables: education level and age group.

These results are quite interesting, since they enable distinguishing the two dimensions of differentiation: the horizontal differentiation, between languages (Catalan or Spanish), and the vertical differentiation, between types of versions (dubbed, including both Catalan or Spanish, and original versions). Thus, the exposure to VoD generates a switch from dubbed versions to original ones; such an effect is stronger (or existent) for younger and more educated individuals. Regarding the horizontal dimension, characterized in columns (1) and (2), we can observe how it greatly depends on the first language. VoD does not decrease the preference for Catalan for Spanish speakers and foreign-language speakers, probably because they did not prefer these versions previously. The decrease is greater for Catalan speakers than for individuals with both Catalan and Spanish as first tongues, since they might have a greater preference for these versions previously. However, we do not observe the reverse in the second column; the coefficient of Spanish speakers is negative but non-significant. One potential reason is that, as explained in the hypothesis, Spanish versions are still fully available on VoD platforms, unlike Catalan, so the previous users can maintain their audiovisual consumption habits. However, the effect is very large and significant for foreign-language speakers, who switch largely to the original versions or are indifferent. The interpretation of the fourth column is unclear since we cannot know to which options the individual is indifferent. Nonetheless, we can suppose that the indifference increases in the vertical dimension in that case, as an intermediate step between the dubbed and original versions.

4.6. Conclusions

In this paper, I investigated the effect of the irruption of VoD platforms in Spain, a major change in the movie market, on the preferences relating to different language versions. The spread of VoD platforms and especially Netflix, which was launched in late 2015, totally changed the landscape in terms of the language versions available for movies. In the specific case of Catalonia, it meant a new source of movie consumption with fewer films in Catalan versions than the other ways of consuming movies: cinema and TV. In turn, the VoD expansion dramatically increased the availability of films in original version (with subtitles), while the availability of films in the state language (in the case of Catalonia, Spanish) remained the same: virtually all were available in that language. Hence, the main hypothesis tested in this paper was that the increased (decreased) exposure to original (Catalan) versions should lead to an increase (reduction) of the preference for such a version. In the case of Spanish versions' preference, there was no clear prediction, since their availability remained the same, but there was an increase in the variety of other versions available, due to their non-exclusivity.

The results clearly confirmed the positive effect of VoD exposure on the preference for original versions. The estimates indicate that being exposed to a VoD platform increased at least by 7.86 percentage points the probability of preferring films in original versions. The general results also showed a negative but non-significant effect on the preference for Catalan versions. The effect on the preference for Spanish versions, that was subject to two counterposed effects, turned out to be negative as well; being a VoD user, ceteris paribus, decreases the probability of preferring Spanish versions by 6.26 p.p. The empirical analysis shows that the vertical effect (the acceptance of original versions with subtitles) is intense, while the horizontal effect (Spanish versus Catalan in dubbing) is negligible. Finally, no significant effect was found on the indifference towards language preferences.

In this work I also looked at the heterogeneity of the impact of VoD exposure on language preferences. The results show that it only reduced the preference for Catalan versions for the Catalan speakers, most likely due to the fact that such preference was already low for non-Catalan speakers. Such an effect was actually very high, with a reduction of 10 p.p. for Catalan speakers and 7.27 p.p. for individuals with both Catalan and Spanish as their first language. The other heterogeneous effects demonstrate that more educated and younger individuals

have a higher propensity to be affected by the exposure to VoD, indicating that they are more adaptable and/or they have the ability to understand other languages or read fast (to understand original versions with subtitles). In addition, these results highlight the two dimensions of differentiation: the horizontal (between languages) and the vertical (between the dubbed versions and the original versions with subtitles). The VoD expansion greatly affected the vertical dimension, with more individuals switching to original versions from dubbed ones, especially young, educated individuals. Whether VoD reduced the preference for Catalan or Spanish (or both) depends on the linguistic characteristics of the individuals; thus, VoD did not make individuals switch within the horizontal dimension, that is, between languages.

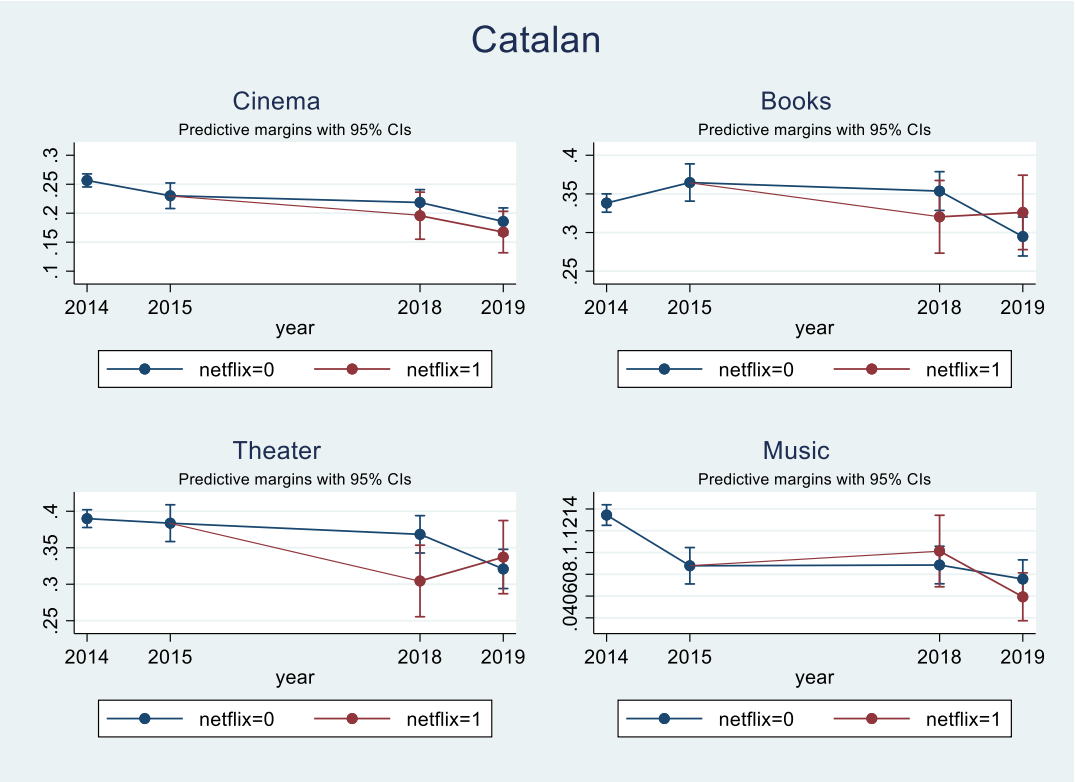
The evidence reported in this paper indicates that individual preferences are endogenous: they adapt to previous experiences. Individuals' tastes accustom or accommodate to market changes. Specifically, the case of VoD shows how the introduction of a new technology in a market (in the case of this paper, the movie market) can rapidly shape the preferences of consumers.

In terms of policy implications, the results show the importance of taking into account the endogeneity of the language preferences for policy design. A specific regulation that could alter the linguistic landscape of versions for films will likely create an adaptative change in consumers habits, since they will adapt their preferences. Thus, the welfare analysis of the implementation of the General Law of Audiovisual Communication of 2022, explained in the introduction, that established quotes for production in the different official languages of Spain, should take into account that in the mid-run it will generate a change in the preferences for different language versions for some consumers.

The expansion of the VoD that continued after the period of analysis of this paper and was boosted by the Covid lockdown will have predictably increased even more the preference for original versions among the population. This could potentially increase individuals' English language skills (Micola et al., 2019) and therefore lead to a subsequent rise in international trade (Ku & Zussman, 2010) and economic growth (Reksulak et al., 2004).

Appendix

Figure 4.A1: Catalan. Margins with Full Specification for Cinema, Books, Theatre and Music



Note: CIs refers to Confidence Intervals.

Figure 4.A2: Spanish. Margins with Full Specification for Cinema, Books, Theatre and Music

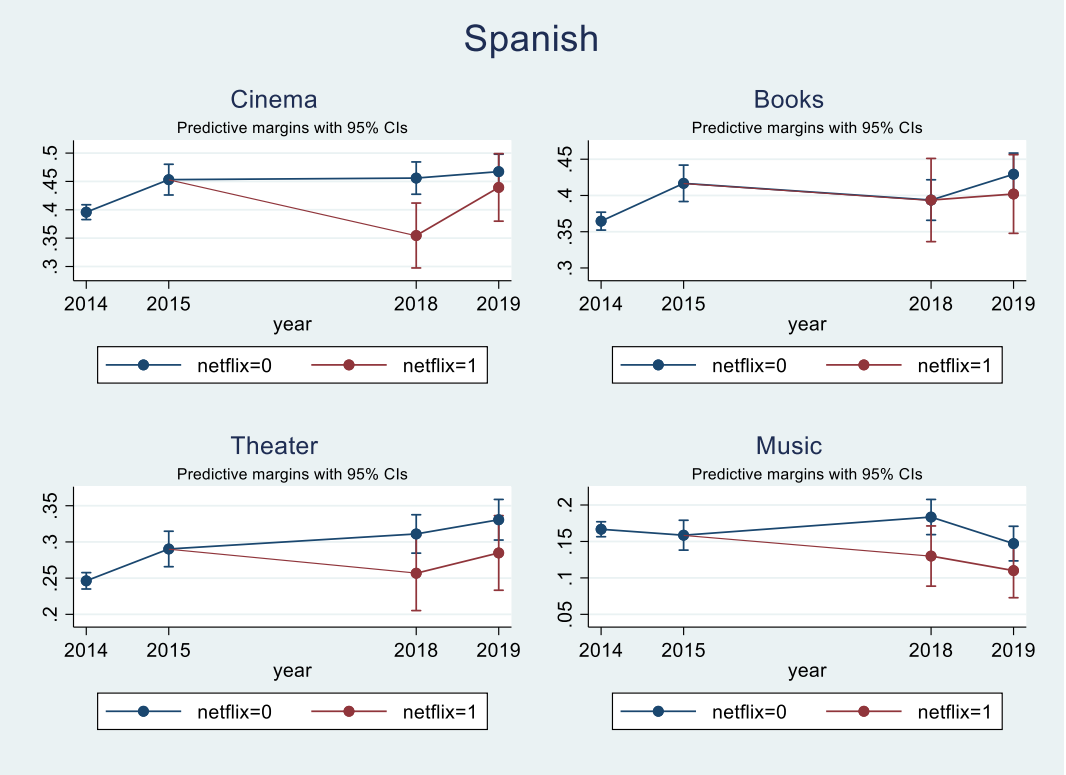


Figure 4.A3: Original Version. Margins with Full Specification for Cinema, Books, Theatre and Music

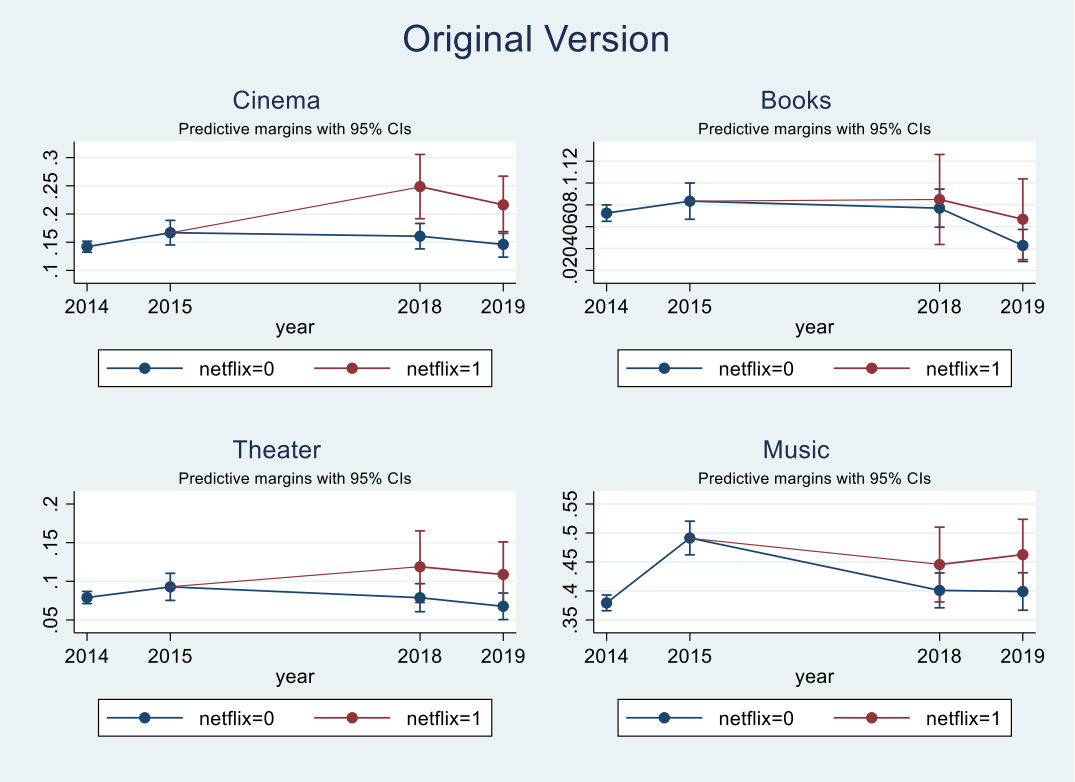


Table 4.A1.1: Basic Regression and Exogenous Characteristics

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
Basic regression				
Netflix	-0.0520***	-0.0779***	0.176***	-0.0425**
	(0.016)	(0.025)	(0.023)	(0.018)
ϕ_{2014}	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
ϕ_{2015}	-0.0592***	0.0783***	0.0355***	-0.0503***
	(0.0146)	(0.0171)	(0.0127)	(0.0126)
ϕ_{2018}	-0.112***	0.115***	0.0255**	-0.0311**
	(0.0132)	(0.0169)	(0.0126)	(0.0128)
ϕ_{2019}	-0.138***	0.143***	0.00141	-0.0136
	(0.0128)	(0.0172)	(0.0123)	(0.0136)
Constant	0.283***	0.376***	0.134***	0.203***
	(0.007)	(0.007)	(0.005)	(0.006)
Individual exogenous characteristics as controls				
Netflix	-0.0228	-0.0905***	0.113***	0.00426
	(0.016)	(0.024)	(0.023)	(0.018)
L1. Catalan	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
L1. Spanish	-0.380***	0.405***	-0.0349***	0.0125
	(0.010)	(0.010)	(0.008)	(0.009)
L1. Catalan-Spanish	-0.270***	0.163***	0.0112	0.0931***
	(0.022)	(0.028)	(0.023)	(0.025)
L1. Others	-0.341***	0.273***	0.0207	0.00595
	(0.017)	(0.027)	(0.022)	(0.020)
Age	0.0139***	-0.0114***	-0.00944***	0.00728***
	(0.001)	(0.002)	(0.001)	(0.001)
Age squared	-0.000107***	0.0000907***	0.0000516***	-0.0000400***
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.0382***	0.0203**	-0.00701	0.0267***
	(0.008)	(0.010)	(0.008)	(0.009)
ϕ_{2014}	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
ϕ_{2015}	-0.0323**	0.0479***	0.0415***	-0.0531***
	(0.0127)	(0.0155)	(0.0123)	(0.0125)
ϕ_{2018}	-0.0472***	0.0509***	0.0356***	-0.0382***
	(0.0118)	(0.0157)	(0.0124)	(0.0129)
ϕ_{2019}	-0.0782***	0.0822***	0.0139	-0.0224
	(0.0117)	(0.0162)	(0.0123)	(0.0137)
Constant	0.137***	0.462***	0.486***	-0.0989***
	(0.035)	(0.044)	(0.037)	(0.032)
N	8045			

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

ϕ are the year dummies

Table 4.A1.2: Full Specification

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
Netflix	-0.0204	-0.0626***	0.0786***	0.00754
	(0.016)	(0.024)	(0.021)	(0.018)
L1. Catalan	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
L1. Spanish	-0.313***	0.330***	-0.0462***	0.0333***
	(0.011)	(0.013)	(0.010)	(0.011)
L1. Catalan-Spanish	-0.242***	0.145***	-0.00581	0.100***
	(0.022)	(0.028)	(0.023)	(0.025)
L1. Others	-0.274***	0.243***	-0.0647**	0.0663***
	(0.020)	(0.033)	(0.025)	(0.024)
Age	0.0126***	-0.00826***	-0.0114***	0.00730***
	(0.001)	(0.002)	(0.001)	(0.001)
Age squared	-0.0000767***	0.0000339*	0.0000792***	-0.0000382**
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.0353***	0.0135	-0.00349	0.0275***
	(0.008)	(0.010)	(0.008)	(0.009)
Born. Catalonia	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
Born. Spain	-0.122***	0.110***	0.0362***	-0.0232*
	(0.011)	(0.016)	(0.011)	(0.013)
Born. European Union	-0.0339*	-0.0246	0.121***	-0.0896***
	(0.018)	(0.034)	(0.027)	(0.022)
Born. Non-EU	-0.0453***	-0.0235	0.118***	-0.0525**
	(0.015)	(0.029)	(0.026)	(0.020)
Municipality. Less than 50,001 inhabitants	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
Municipality. Between 50,001 and 150,001	-0.0345***	0.0681***	-0.00133	-0.0326***
	(0.010)	(0.013)	(0.009)	(0.011)
Municipality. Between 150,001 and 1M	-0.0520***	0.0339*	0.0249*	-0.00894
	(0.014)	(0.019)	(0.014)	(0.016)
Municipality. More than 1M	-0.0714***	-0.0321**	0.137***	-0.0322**
	(0.012)	(0.015)	(0.013)	(0.013)
Works	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
Does not work	-0.0201**	0.0246**	-0.00991	0.00432
	(0.010)	(0.012)	(0.009)	(0.011)
Primary education or less	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
Secondary education	0.0459***	-0.0457**	0.00549	0.00280
	(0.014)	(0.019)	(0.010)	(0.017)
Professional course	0.0641***	-0.117***	0.0429***	0.0223
	(0.015)	(0.019)	(0.011)	(0.017)
Tertiary education	0.0381**	-0.222***	0.183***	0.0120
	(0.016)	(0.020)	(0.013)	(0.018)
ϕ_{2014}	Ref. Cat.	Ref. Cat.	Ref. Cat.	Ref. Cat.
ϕ_{2015}	-0.0264**	0.0575***	0.0248**	-0.0524***
	(0.0127)	(0.0153)	(0.0121)	(0.0126)
ϕ_{2018}	-0.0384***	0.0531***	0.0203	-0.0326**
	(0.0124)	(0.0157)	(0.0125)	(0.0134)
ϕ_{2019}	-0.0702***	0.0794***	0.00217	-0.0101
	(0.0129)	(0.0167)	(0.0125)	(0.0145)
Constant	0.0987**	0.574***	0.417***	-0.112***
	(0.0386)	(0.0480)	(0.0377)	(0.0379)
N	8045			

Robust standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

ϕ are the year dummies

Table 4.A2: Main Results with Non-Linear Models

	(1) Catalan	(2) Spanish	(3) Original Version	(4) Indifferent
VoD (Logit)	-0.0429*	-0.0598***	0.0441***	0.0067
	(0.0240)	(0.0225)	(0.0145)	(0.0221)
VoD (Probit)	-0.0411*	-0.0584***	0.0480***	0.0063
	(0.0228)	(0.0225)	(0.0152)	(0.0213)
N	8045			

Standard errors in parentheses * p<0.1, ** p<0.05, *** p<0.01

Full specification. Marginal effects reported.

5. Conclusion

Economics and Language are closely intertwined subjects. Although the study of their relationship may be considered a relatively small field within the economics literature, its intrinsic interdisciplinary nature allows it to intersect with multiple fields. Therefore, the Economics of Language encompasses research that investigates language acquisition as a form of human capital investment with implications for various outcomes, which is situated within the literature on the Economics of Education and Economics of Migration. When this research examines the labor market impact of linguistic proficiency, it also contributes to the field of Labor Economics. Consequently, the first article in this dissertation also makes contributions to these economic subfields. On the other hand, as demonstrated in the second article, analyzing markets where language is a significant characteristic utilizes the framework of Industrial Organization. Since these markets often involve cultural goods, such studies also contribute to the field of Cultural Economics. Furthermore, the third article, despite focusing on the same cultural good (films), is somewhat connected to the literature on Behavioral Economics, as it explores changes in preferences resulting from market innovation.

This dissertation embodies the holistic view and broad scope of the Economics of Language. However, tackling such a comprehensive approach posed a major challenge throughout the research process. Thankfully, with the invaluable assistance of my advisors, who complemented each other perfectly in their respective areas of expertise, these challenges were successfully overcome. The interdisciplinary nature of this thesis necessitated presenting the articles at conferences spanning various fields, including Labour Economics[46], Economics of Education[47], Industrial Economics[48] and Cultural Economics[49].

In the first article, we conducted a novel analysis of the impact of subsidized language courses for adults offered by the CLN. The results reveal a modest yet robust effect of exposure to language centers on language proficiency, especially regarding the probability of being able to speak and write in Catalan. The impact of the local supply of language courses is more pronounced for younger and less educated individuals and, most strikingly, is relevant only for those who

[46] Workshop on Labour Economics 2022, IAAEU
[47] 7th LEER conference, KU Leuven
[48] 10th WIPE, URV; XXXVI Jornadas de Economía Industrial, ULPGC
[49] EWACE 2022, Università Di Torino

were born in Catalonia: they are indeed those who are more sensitive to an increase of spatial accessibility (i.e. having more language centres near to the place of residence). Actually, this result is consistent with the fact that the availability of language learning centres only affects the probability of speaking and writing in Catalan. In fact, the latter domain of language proficiency is precisely what people born in Catalonia (many of them being native Catalan speakers) wanted to achieve, because they were not exposed to language at school and had limited written skills. However, although competences in Catalan are positively related to labour market performance (especially employment probability and occupation), the effects induced by the geographical accessibility are too modest to lead to any improvements in any of the labour market outcomes considered in this work.

In the second article, I wanted to find whether the concentration of the movie-theatre industry, both at the distribution and the exhibition levels, led to an undersupply of language diversity in a bilingual context. This is, to the best of my knowledge, the first attempt to empirically analyse the language supply of a cultural market. In order to do so, I built a unique data set, using advanced web-scraping techniques and combining different sources, of all the cinema screenings in the bilingual region of Catalonia to explore a potential market failure. Such failure would consist in an undersupply of films in the Catalan version due to the concentration in two stages of the market. Applying a Linear Probability Model controlling by relevant factors and adding several fixed effects, I found that the concentration of the market indeed leads to an undersupply of 4.04 percentage points of films in Catalan, which means that if those market failures did not exist there would be 96% more screenings in Catalan. By using the genre of the film, I disentangled two types of consumers, children and adults. The market response differs if the film is targeted at children or at the general adult audience. Due to the higher intensity of the linguistic preference of the children compared to the adults, I found that the competitive pressure of the exhibitors only changes the supply of films in Catalan when it deals with the child demand (animation films), and the effect is negligible for the rest.

In the third and last article, I investigated the effect of the introduction of VoD platforms in Spain, a major change in the movies' market, on the preferences over language versions. The results clearly confirmed the positive effect of VoD exposure to the preference over original versions. The estimates indicate that

being exposed to a VoD platform increased at least by 7.86 percentage point de probability to prefer films in original versions. The general results also showed a negative but non-significant effect on the preference over Catalan versions. The effect on the preference over Spanish versions turned out to be negative too; being a VoD user, ceteris paribus, decreases the probability to prefer Spanish versions by 6.26 p.p. The empirical analysis shows that the vertical effect (the acceptance of original versions with subtitles) is intense, while the horizontal effect (Spanish versus Catalan in dubbing) is negligible. In this work I also looked at the heterogeneity of the impact of VoD exposure on the language preferences. The results show that it only reduced the preference over Catalan versions for the Catalan-speakers, most likely due to the fact that such preference was already low for non-Catalan-speakers. Such effect was actually very high, with a reduction of 10 p.p. for Catalan-speakers and 7.27 p.p. for bilingual speakers. The other heterogeneous effects demonstrate that more educated and younger individuals have a higher propensity to be affected by the exposure to VoD, indicating that they are more adaptable and/or they have the ability to understand other languages or read fast (to understand original versions with subtitles).

In addition to the specific implications discussed in the conclusion subsections of each article, there are overarching observations that can be made when considering the big picture.

First, that the language-related issues usually require some sort of public intervention and a subsequent policy evaluation. This is evident in the first paper, which represents the first evaluation of the extensive language programs of the CLN, despite them being implemented long ago. Economists have often been detached from language policies, resulting in inefficient policy planning. Evaluation is crucial for enhancing policy design. In our study, we found that geographical proximity to language courses is important. Nonetheless, over the past 20 years, there has been a trend of centralizing the distribution of language courses in large schools in each district, which is contrary to our findings. While there are valid cost-related reasons for this approach, such as benefiting from economies of scale, our results demonstrate that it can also reduce the effectiveness of the courses. Therefore, courses should be located where the targeted individuals reside in order to maximize their impact.

While the first paper illustrates the importance of the evaluation of language policies, the second article shows how the lack of economic research on this

type of matters can lead to a welfare loss, due to the absence of government, either through regulation or provision. Given that languages possess non-communicative and subjective value, which influences individuals' willingness to pay, it is crucial to closely examine the supply of language diversity in non-perfectly competitive markets.

Another joint implication of the articles, especially the second and the third, is that the structure of a cultural market can indirectly affect the influence language preferences. As demonstrated in the second article, market concentration can result in a reduction in the availability of films in the Catalan language. This partially explains the varying levels of language provision observed across different cultural markets, such as theater, cinema, radio, and books. Fragmented markets, in general, tend to offer a higher percentage of products in the weaker language, Catalan. However, when individuals express their language preferences for specific cultural goods, they may reveal differences among products. The findings of the third paper shed light on this aspect by demonstrating how consumers adapt their preferences to align with the prevailing language landscape. In other words, in terms of habit formation, individuals become accustomed to the language typically associated with each product they consume. These two effects together bring us to the conclusion that the structure of the market have a direct and an indirect effect on language provision. Initially, concentrated markets can result in suboptimal provision of the weaker language. Subsequently, consumers adapt their preferences based on the availability of languages in the market, thereby influencing language demand. This helps to explain why there is a lower revealed preference for Catalan versions in cinema compared to books or theater.

Hence, these two effects can converge in a vicious circle wherein reduced provision of the weaker language resulting from market concentration leads to decreased demand for that language, subsequently leading to a further decline in the supply of the weaker language version, and so forth. Conversely, the opposite dynamic occurs for the language favored by market concentration.

In terms of public policy, understanding the dynamic effects of language regulation in specific markets is crucial. Let's take the example of the cinema industry. As discussed in chapter 3.2, the Law of the Cinema was enacted in 2010, initially requiring 50% of productions to be either dubbed or subtitled in Catalan. However, this requirement was later reduced to 25% by the Constitutional Court. If the policy had been fully implemented, we would have

expected a gradual increase in the preference for Catalan versions over the medium term. Therefore, to assess the overall welfare impact of such a policy, it is essential to consider not only the immediate effects but also the long-term dynamic effects it produces. Moreover, such kind of evaluation is problematic, since the literature is divided between those who advocate that initial preferences should be taken into account and those who are in favor of considering the preferences resulting from the policy.

These considerations should be taken with caution, since the results of the third article are bounded by the limitation of the data. Nevertheless, I believe that this initial exploration of the endogeneity of linguistic preferences sets the stage for future research in this area, which is undoubtedly an intriguing yet complex topic.

Overall, the main objective of this dissertation was to examine previously unexplored language-related issues. As a result, it can be said that this thesis raises more questions than it provides answers, which is a common occurrence when delving into uncharted research territory. There is still much research to be conducted on market structure and language provision, particularly from an empirical perspective. In terms of the formation of language habits, I have demonstrated that economists can offer a distinct and innovative approach to its study, which was previously the domain of sociolinguists. With the growing availability of data on multilingualism and individual choices, econometric analysis plays a crucial role in future research endeavors.

Finally, Language Policy Evaluation should be incorporated as routinary part of the language planning of governments and institutions. The application of standard econometric tools has demonstrated its effectiveness in analyzing language policies. By integrating insights from various disciplines, policymakers can enhance the design of future policies or reevaluate and adjust existing ones.

References

Allain, M-L, Chambolle, C., & Rey, P. (2016). Vertical integration as a source of hold-up. The Review of Economic Studies, *83*(1) 1–25.

Anderson, C., & Andersson, M. P. (2013). *Long tail*. Bonnier fakta.

Aparicio Fenoll, A., & Kuehn, Z. (2016). Does foreign language proficiency foster migration of young individuals within the European Union. The economics of language policy, 331-355.

Arendt, J. N., Bolvig, I., Foged, M., Hasager, L., & Peri, G. (2020). Language Training and Refugees' Integration. National Bureau of Economic Research working paper No. w26834.

Åslund, O., & Engdahl, M. (2018). The value of earning for learning: Performance bonuses in immigrant language training. Economics of Education Review, 62, 192-204.

Bleakley, H., & A. Chin (2004), Language skills and earnings: Evidence from childhoold immigrants. *Review of Economics and Statistics* 86, 481-496.

Bleakley, H., & Chin, A. (2010). Age at arrival, English proficiency, and social assimilation among US immigrants. American Economic Journal: Applied Economics, 2(1), 165-92.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. Journal of Economic Literature, *36*(1), 75–111.

Butlletí d'informació sobre l'audiovisual a Catalunya (No. 20). (2022).

Caminal, R. (2010). Markets and linguistic diversity. Journal of Economic Behavior and Organization, *76*(3), 774-790

Caminal, R., Cappellari, L., & Di Paolo, A. (2021). Language-in-education, language skills and the intergenerational transmission of language in a bilingual society. Labour Economics, 70: 101975.

Caminal, R., & Di Paolo, A. (2019) Your language or mine? The noncommunicative benefits of language skills. Economic Inquiry, *57*(1), 726–750.

Cappellari, L., & Di Paolo, A. (2018). Bilingual schooling and earnings: Evidence from a language-in-education reform. Economics of Education Review, 64, 90-101.

Chen, Y., & Riordan, M., 2007. Price and variety in the spokes model. Economic Journal 117, 897-921.

Chiswick, B. R., & Miller, P. W. (2007). The economics of language: International analyses. Routledge.

Dahl, G., & DellaVigna, S. (2009). Does movie violence increase violent crime?. The Quarterly Journal of Economics, *124*(2), 677–734.

Davis, P. *(*2006). Spatial competition in retail markets: Movie theaters. *The RAND Journal of Economics*, *37*(4) 964–982.

DellaVigna, S., & Kaplan, E. (2007). The Fox News effect: Media bias and voting. *The* Quarterly Journal of Economics, *122*(3), 1187–1234.

Dietrich, F., & List, C. (2013). Where do preferences come from? International Journal of Game Theory, 42(3), 613–637.

Di Paolo, A. & Mallén, B. (2023). Does geographical exposure to language learning centres matter in a bilingual city?. IREA Working Paper 2022/03

Doh-Shin J., Jullien, B., & Klimenko, M. (July 2021). Language, Internet and platform competition. Journal of International Economics, *131*(103439).

Domínguez, M., & Montolio, D. (2021). Bolstering community ties as a means of reducing crime. Journal of Economic Behaviour and Organization, 191, 916-945.

Egger, P. H., & A. Lassmann (2015), The causal impact of common native language on international trade: Evidence from a spatial regression discontinuity design. *The Economic Journal* 125(584), 699-745.

Eisensee, T., & Strömberg, D. (2007). News droughts, news floods, and US disaster relief. The Quarterly Journal of Economics, *122*(2), 693–728.

Fauli-Oller, R., & Sandonis, J (2016). Welfare effects of downstream mergers and upstream market concentration. The Singapore Economic Review, *61*(05), 1550056.

Foged, M., Hasager, L., Peri, G., Arendt, J. N., & Bolvig, I. (2022). Language Training and Refugees' Integration. The Review of Economics and Statistics, 1-41.

Frankel, J., & A. Rose (2002), An estimate of the e§ect of currencies on trade and income. Quarterly Journal of Economics 117, 437-466.

Gans, J. S. (2007). Concentration-based merger tests and vertical market structure. The Journal of Law and Economics, *50*(4), 661–681.

Garcia–Lopez, M. À., Nicolini, R., & Roig, J. L. (2020). Segregation and urban spatial structure in Barcelona. Papers in Regional Science, 99(3), 749-772.

Gazzola, M. (2014) The Evaluation of Language Regimes. Amsterdam: John Benjamins

Gentzkow, M. (2006). Television and voter turnout. The Quarterly Journal of Economics, *121*(3), 931–972.

Giesecke, M., & Schuss, E. (2019). Heterogeneity in marginal returns to language training of immigrants. IAB-Discussion Paper No. 19/2019.

Ginsburgh, V., & Weber, S. (2011). How many languages do we need? The economics of linguistic diversity. Princeton University Press.

Gonzalez, L., & Ortega, F. (2011). How do very open economies adjust to large immigration flows? Evidence from Spanish regions. Labour Economics, 18(1), 57-70.

Grin, F. (1992). Towards a threshold theory of minority language survival. Kyklos, 45, 66–97.

Grin, F., & Vaillancourt, F. (1998). Language revitalisation policy: an analytical survey, theoretical framework, policy experience and application to Te Reo Maori (No. 98/06). New Zealand Treasury Working Paper.

Heller, B. H., & Slungaard Mumma, K. Immigrant Integration in the United States: The Role of Adult English Language Training. American Economic Journal: Economic Policy (forthcoming).

Institut d'Estadística de Catalunya (IDESCAT). Televisió i continguts audiovisuals. Audiència per llengua. https://www.idescat.cat/pub/?id=eac&n=3.1.2.02

Kivi, L. H., Sõmer, M., & Kallaste, E. (2020). Language training for unemployed non-natives: who benefits the most? Baltic Journal of Economics, 20(1), 34-58.

Ku, H., & Zussman, A. (2010). Lingua franca: The role of English in international trade. Journal of Economic Behavior & Organization, *75*(2), 250–260.

Lang, J. (2021). Employment effects of language training for unemployed immigrants. Journal of Population Economics, 35: 719–754.

Leung, T. C., Qi, S., & Yuan, J. (2020). Movie industry demand and theater availability. Review of Industrial Organization, 56, 489–513.

Lochmann, A., Rapoport, H., & Speciale, B. (2019). The effect of language training on immigrants' economic integration: Empirical evidence from France. European Economic Review, 113, 265-296.

Mallén, B. (2023). The Effect of Competition on Language Diversity in the Movie-Theatre Industry. IREA Working Paper 2023/05, revised May 2023

Melitz, J. (2008), Language and Foreign Trade. *European Economic Review* 52(4), 667-699.

Mèlitz, J. (2012). A framework for analyzing language and welfare. SIREDP-2012-89.

Micola, A. R., Fenoll, A. A., Banal-Estañol, A., & Bris, A. (2019). TV or not TV? The impact of subtitling on English skills. Journal of Economic Behavior & Organization, *158*, 487–499.

Pont-Grau, A., Lei, Y. H., Lim, J. Z., & Xia, X. (2023). The effect of language training on immigrants' integration: Does the duration of training matter? Journal of Economic Behavior & Organization, 212, 160-198.

Quan, T. W., & Williams, K. R. (2018). Product variety, across-market demand heterogeneity, and the value of online retail. The *RAND* Journal of Economics, 49(4), 877–913.

Reksulak, M., Shughart, W. F., & Tollison, R. D. (2004). Economics and English: Language growth in economic perspective. Southern Economic Journal, *71*(2), 232–259.

Tucker, G. R. (2001). A global perspective on bilingualism and bilingual education. In J. E. Alatis, & A.-H. Tan (Eds.), Roundtable on language and linguistics.

Orbach, B. Y., & Einav, L. (2007). Uniform prices for differentiated goods: The case of the movie-theater industry. International Review of Law and Economics, *27*(2), 129-153.

Wickstrom, B. A. (2005). Can bilingualism be dynamically stable? A simple model of language choice. Rationality and Society, *17*(1), 81-115.