Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

Document de Treball 2025/13 1/36 pág.
Working Paper 2025/13 1/36 pag.

Grup de Recerca Anàlisi Quantitativa Regional
Regional Quantitative Analysis Research Group

Document de Treball 2025/07 1/36 pág.
Working Paper 2025/07 1/36 pag.

Language of Instruction, Bilingualism, and Neighbourhood Quality: Do Local Language Skills Matter?

Antonio Di Paolo

UBIREA | Institut de Recerca en Economia Aplicada Regional i Pública

UBIREA | Institut de Recerca en Economia Aplicada Regional i Pública
UNIVERSITAT DE BARCELONA

WEBSITE: www.ub.edu/irea/ • CONTACT: irea@ub.edu

The Research Institute of Applied Economics (IREA) in Barcelona was founded in 2005, as a research institute in applied economics. Three consolidated research groups make up the institute: AQR, RISK and GiM, and a large number of members are involved in the Institute. IREA focuses on four priority lines of investigation: (i) the quantitative study of regional and urban economic activity and analysis of regional and local economic policies, (ii) study of public economic activity in markets, particularly in the fields of empirical evaluation of privatization, the regulation and competition in the markets of public services using state of industrial economy, (iii) risk analysis in finance and insurance, and (iv) the development of micro and macro econometrics applied for the analysis of economic activity, particularly for quantitative evaluation of public policies.

Abstract

This paper investigates whether acquiring proficiency in a local language improves neighbourhood quality in a bilingual region, focusing on Catalonia, Spain. The analysis uses rich microdata linked to census-tract measures of neighbourhood quality, including average local income, unemployment benefits per capita, and a composite socioeconomic status index. OLS results show that oral proficiency in Catalan among native Spanish speakers is associated with better residential outcomes. To address potential endogeneity of language skills, I exploit the implementation of a language-in-education policy that introduced Catalan as a medium of instruction, promoting Catalan-Spanish bilingualism among native Spanish speakers. Specifically, I construct an instrument consisting in the interaction between years of language exposure during compulsory education and an indicator for native Spanish speakers, considering that the reform did not affect oral Catalan proficiency among native Catalan speakers and assuming cohort trends unrelated to the reform are homogeneous across language groups. IV/TSLS estimates reveal no causal effect of increased oral Catalan skills, induced by school language exposure among native Spanish speakers, on any measure of neighbourhood quality. Falsification exercises aimed at validating the main identification assumption, along with robustness checks addressing potential confounders and alternative mechanisms, support the identification strategy and reinforce the main findings. Overall, the results suggest that although the reform significantly raised oral Catalan proficiency among native Spanish speakers, this variation in language skills does not translate into changes in residential sorting or neighbourhood quality.

Keywords: local language skills, bilingualism, language-in-education policy, neighbourhood quality

JEL Classification: I28, Z13 R23

Author:

Antonio Di Paolo, Department of Econometrics, Statistics and Applied Economics, Universitat de Barcelona. Address: Avinguda Diagonal 690, 08034 Barcelona (Spain). Email: antonio.dipaolo@ub.edu.

Acknowledgements: The research presented in this paper was carried out with the support of a research grant funded by the Institut d'Estudis de l'Autogovern (PRE166/23/000004). The author also benefited from funding provided by the Ministry of Science and Innovation (PID2021-122575NB-I00). There are no conflicts of interest to disclose, and all remaining errors are the sole responsibility of the author. **:** the research presented in this paper was carried out with the support of a research grant funded by the Institut d'Estudis de l'Autogovern (PRE166/23/000004). The author also benefited from funding provided by the Ministry of Science and Innovation (PID2021-122575NB-I00). There are no conflicts of interest to disclose, and all remaining errors are the sole responsibility of the author.

1) Introduction

Living in a good neighbourhood might improve several dimensions of individuals' socioeconomic outcomes and families' wellbeing. The literature on this topic is extensive, and more recent works based on experimental and quasi-experimental approaches have confirmed the existence of positive effects of neighbourhood quality. They have also highlighted that the effects of the quality of the residential environment are particularly relevant for young children who are exposed to better neighbourhood conditions (Chyn and Katz, 2021). Therefore, a natural research question consists in investigating the extent to which individuals' human capital and skills shape the opportunity to live in better neighbourhoods and, more broadly, contribute to the spatial sorting of individuals (Diamond and Gaubert, 2022). Indeed, a higher endowment of human capital can enhance earnings potential, which in turn increases the likelihood of residing in a high-quality neighbourhood, as neighbourhood quality is capitalized into housing prices. Moreover, individuals with higher human capital are more likely to access better social networks and information, which can reduce search costs, improve bargaining power and ultimately enhance their position in the housing market.

This paper investigates the effect of language skills — as a component of human capital — on various measures of neighbourhood quality. While most existing research has focused on the opposite direction of causality in the case of migrants, namely the impact of neighbourhood characteristics — particularly linguistic enclaves — on host-country language proficiency (see, among others, Beckhusen et al., 2013; Danzer and Yaman, 2016; Laliberté, 2019; Danzer et al., 2022), the literature examining the causal effect of migrants' language skills on neighbourhood quality remains limited. To the best of my knowledge, only two studies have explicitly addressed this question. First, Bleakley and Chin (2010) exploited the differential effect of migrating at a young age — specifically, before age 9 — on English proficiency among childhood migrants to the U.S., depending on whether individuals came from English-speaking or non-English-speaking

countries, using an instrumental variable approach. Among the various outcomes they examined, one consists in the composition of local enclaves. They found that, for females from non-English-speaking countries, improved English proficiency due to earlier age at arrival significantly reduced the proportion of co-nationals or individuals of the same ancestry residing in the same Public Use Microdata Area (PUMA). Second, Aoki and Santiago (2024) examined the case of migrants in the UK and employed an identification strategy similar to that of Bleakley and Chin (2010). Their instrumental variable for English proficiency is based on the interaction between the excess age at arrival beyond age 8 and the linguistic distance between the origin-country language and English. They not only focused on different measures of residential enclave, but also examined the causal effect of English proficiency on several indicators of neighbourhood quality — namely, the quintiles of three domains from the English Indices of Deprivation: income, employment, and health deprivation.[1] Overall, their results indicate that the deficiency in English skills have a negative impact on neighbourhood quality.

The focus of this paper differs from previous studies, as it examines the relationship between local language skills and neighbourhood quality among native residents in a bilingual region, rather than the impact of host-country language proficiency among migrants. Specifically, I study the Spanish bilingual region of Catalonia to assess whether improved oral proficiency in Catalan — the vernacular language of the region — affects various socioeconomic indicators of residential quality. The case of Catalonia is particularly interesting for several reasons. First, it is an asymmetrically bilingual region with two main native linguistic communities: native Catalan speakers, who are fully bilingual in Catalan and Spanish, and native Spanish speakers, only a fraction of whom are

[1] Specifically, Income deprivation reflects the proportion of the population experiencing low income within a neighbourhood, while employment deprivation captures the share of the working-age population who are involuntarily excluded from the labour market. Health deprivation captures age and sex specific premature mortality, as well as the share of the population experiencing reduced quality of life due to poor physical or mental health. The final health deprivation indicator is constructed as a single ordinal scale derived through factor analysis.

proficient in Catalan. Therefore, any role played by Catalan proficiency on residential sorting cannot be attributed to its communicative function, because communication is not at stake. Second, a large proportion of native Spanish speakers are descendants of internal migrants who arrived in Catalonia during the mass migration waves of the 1950s and 1960s. These migrants tended to be residentially segregated in low-income areas, primarily located on the periphery of Barcelona and in different municipalities of its Metropolitan Area (Garcia-López et al., 2021). To some extent, this residential segregation may have been perpetuated across generations, maintaining a certain degree of persistence. Third, these individuals experienced not only spatial but also linguistic segregation, with limited or no knowledge of Catalan. Both dimensions of segregation are likely positively correlated and may have reinforced one another. Fourth, and related to the previous point, during the early years of democracy following Franco's regime, the Catalan government (*Generalitat de Catalunya*) implemented a broad language policy aimed at restoring the social use and prestige of Catalan—severely repressed during the dictatorship—and at ensuring bilingualism regardless of regional or linguistic origin: the Language Normalization Act (LNA) of 1983. This language policy received a large political support from all political parties at that time and was considered to be one of the main drivers of upward social mobility of descendants of internal migrants and native Spanish speakers in general, and generally promoted equal opportunities in the Catalan society.[2] As explained below, I take advantage of this reform as a natural experiment, to investigate the existence of a causal impact of local language skills on neighbourhood quality.

Prior research has highlighted the positive effects of Catalan language skills on labour market outcomes (e.g., Rendón, 2007; Di Paolo and Raymond, 2011) and other social outcomes such as the formation of linguistically mixed couples

[2] The details of the LNA reform are thoroughly described by Cappellari and Di Paolo (2018), who also analyzed its impact on the returns to education. Specifically, the authors showed that the introduction of Catalan-Spanish bilingualism at school, induced by the language-in-education component of the reform, increased the wage returns to schooling. The effect was larger for individuals with a non-Catalan background and was mostly driven by increased skills in Catalan.

(Caminal and Di Paolo, 2019) and the intergenerational transmission of language (Caminal et al., 2021). However, the relationship between local language proficiency and neighbourhood quality in a bilingual context remains unexplored. Therefore, this paper is the first to address this gap in the literature. Moreover, another important contribution of this paper to the existing research consists in the use of geographically disaggregated and continuous measures of neighbourhood quality, defined at the census tract level.[3] Specifically, the outcome variables considered in this study are: gross and net average per-capita income; the average gross amount of unemployment benefits per person, used as a proxy for the intensity of welfare dependency related to unemployment at the local level; and a composite index of socioeconomic status (see Section 2 for details). These census tract–level outcome variables are merged with microdata from the 2013 and 2018 waves of the Survey of Language Use of the Catalan Population, using census tract identifiers corresponding to respondents' place of residence in the year of the survey.

To address the endogeneity of language skills, I exploit the introduction of Catalan as medium of instruction in compulsory education (alongside Spanish), implemented within the framework of the 1983 Language Normalization Act. Following the approach of Caminal and Di Paolo (2019) and Caminal et al. (2021), I leverage the differential effect of exposure to Catalan at school on oral proficiency between native Spanish and native Catalan speakers. This difference arises because native Catalan speakers were already orally proficient in their heritage language through intergenerational transmission within the family. Accordingly, the instrumental variable for oral proficiency in Catalan is defined as the interaction between years of potential exposure to Catalan during compulsory schooling and an indicator for being a native Spanish speaker. The corresponding 2SLS estimator identifies the causal effect of improved oral skills in Catalan among native Spanish speakers induced by language exposure in

[3] Specifically, Aoki and Santiago (2024) used data at the level of ONS Lower-layer Super Output Areas (LSOAs), which are geographical units roughly equivalent to census tracts. However, due to confidentiality restrictions, their measures of neighborhood quality are based on quintiles of the original continuous variables, making them essentially discrete.

school, under the assumption that cohort effects unrelated to the language-in-education policy are similar across the two language groups. Overall, an additional novel feature of this work consists in broadening the existing evidence on the effects of language policies aimed at preserving regional minority languages, by focusing on their previously unexplored impact on the quality of the residential environment. In this sense, the paper represents a first step toward a promising line of research on the relationship between language policies and residential sorting and segregation in multilingual settings.

The results indicate that oral proficiency in Catalan is strongly associated with various measures of neighbourhood quality, and in the expected direction. Specifically, it is positively correlated with local-level gross and net per-capita income and with the index of socioeconomic status, and negatively associated with the average amount of unemployment benefits per capita. However, the estimates obtained using the Instrumental Variables approach indicate that the increase in oral proficiency in Catalan induced by the language-in-education policy among native Spanish speakers does not causally affect any of the measures of neighbourhood quality. These results remain robust across a range of falsification exercises, controls for potential confounding factors, and additional sensitivity checks. Overall, the evidence reported in this paper indicate that native Spanish speakers who benefited from the introduction of Catalan in schools—thereby gaining the opportunity to become bilingual, like native Catalan speakers—did not experience improvements in residential quality. Consequently, the policy did not contribute to the existing language-related spatial sorting and residential segregation in the bilingual region of Catalonia.

The remainder of the paper is structured as follows. Section 2 presents the data and descriptive statistics. Section 3 outlines the empirical strategy. Section 4 reports the main results along with evidence from several robustness checks. Finally, Section 5 concludes.

2) Data and Descriptive Statistics

The empirical analysis presented in this paper combines multiple data sources, integrating individual-level microdata with information at the census tract level in the Spanish region of Catalonia. The census tracts represent the geographical units used to define the neighbourhood in this work.[4]

The individual-level data are drawn from the 2013 and 2018 waves of the Survey of Language Use of the Population (*Enquesta d'Usos Lingüístics de la Població*[5], EULP), which is carried out every five years by the Catalan Statistical Institute (IDESCAT). The survey is representative of the Catalan population aged 16 or more for the corresponding years and includes an extensive set of variables related to language background, skills and use, sociodemographic characteristics of the respondents and their parents, labour force status and current or previous occupation (1 digit classification). The definition of key linguistic variables is particularly important for the objectives of this study and warrants a detailed explanation. First, I use data on self-reported oral proficiency in Catalan, measured on a 0–10 scale.[6] Second, I exploit information on native language, defined as the language first spoken during childhood. This variable allows respondents to select a specific language or indicate a mixed native language (Catalan and Spanish). I define native Spanish speakers those individuals reporting only Spanish as native language.[7] Third, I employ several additional language-related variables for robustness checks, including the respondent's current self-identification language, the partner's language (if applicable), and the language used with friends. Finally, for the purpose of this project, I obtained

[4] In Spain, census tracts are small geographical units defined as subdivisions of municipal boundaries for electoral, statistical, and administrative purposes. They typically encompass a relatively uniform population ranging from 1,000 to 2,500 inhabitants, except in the case of smaller municipalities.

[5] Additional information regarding the EULP Survey is contained here: https://www.idescat.cat/pub/?id=eulp&lang=en.

[6] The dataset also contains information on additional language skills—comprehension, reading, and writing—for Catalan, Spanish, English, and French, although these are not used in this study.

[7] In a robustness check, I show that the main evidence is unaffected by the exclusion of the few individuals with mixed mother tongue.

confidential information about the census tract of residence of the respondents.[8] After geocoding the data, I obtained the area of each census tract, as well as the latitude and longitude of their centroids.

To characterize neighbourhood quality, I consider several variables defined at the census tract level, which represent the outcome variables examined in this paper. First, I use data from the Household Income Distribution Atlas, developed by the Spanish Statistical Institute (INE) and obtained yearly from tax declaration records.[9] Specifically, I include information about average net and gross income per person (in €), as well the average gross amount of unemployment benefits per person, as a proxy for the intensity of unemployed-related welfare dependency at the local level. Second, I also consider the synthetic Index of Socioeconomic Status (IST) developed by the IDESCAT.[10] This index is defined at the census tract level and was constructed using Principal Component Analysis, combining various administrative records. These include data on the employed population, individuals in low-skilled jobs and with low education levels, young people without post-compulsory education, immigrants from medium- to low-income countries, and average per capita income. The ISS index is standardized with a reference value of 100 for Catalonia as a whole, meaning that the value for each census tract reflects its deviation from the regional average.

The micro-level data were merged with census tract-level variables for both waves of the EULP survey (2013 and 2018), restricted to the census tracts covered by the sampling procedure. However, it is important to note that these measures of neighbourhood quality have been available only since 2015. To address this limitation, I tracked changes in the definition of census tracts between 2013 and

[8] Information about the census tract of residence is not available in the public-use version of the EULP database. Researchers interested in replication analysis can request access to the full, confidential EULP dataset for research purposes through IDESCAT (more information is available here: https://www.idescat.cat/serveis/dades/?lang=en).

[9] Technical details about data collection and definitions are available from the following link: https://www.ine.es/en/metodologia/metodologia_adrh_en.pdf. Notice that the information is not releases for census tracts with an insufficient number of inhabitants, affecting only around 60 out of more than 5000 units.

[10] Technical details are available here: https://www.idescat.cat/pub/?id=ist&lang=en.

2015 and imputed the 2015 values of the outcome variable to characterize neighbourhood quality in 2013, assuming that any potential differences between these two years are negligible. Moreover, monetary outcomes are expressed in real terms, relative to the Consumer Price Index of 2016.

The EULP dataset contains 7255 and 8780 observations for 2013 and 2018, respectively. To construct the main dataset used in the empirical analysis, I retain individuals born in Catalonia and those born in other regions of Spain who migrated to Catalonia before the age of six. This approach ensures a relatively homogeneous sample of individuals who were entirely educated in Catalonia. Additionally, I restrict the sample to individuals born between 1950 and 1990 to exclude those at an advanced stage of life, as well as those who are too young and may not yet be fully emancipated. This also ensures that the sample is symmetrically constructed around 1970, the pivotal year for exposure to the LNA reform during compulsory education. I also drop individuals who are still in education. Overall, these conditions imply that the age range spans from 23 to 68. Finally, I focus on individuals whose native language is either Catalan, Spanish, or both. After dropping few observations with missing values[11] in key individual characteristics, the final sample consists of 6920 individuals (3187 for 2013 and 3733 for 2018).

I also construct two additional samples for falsification analysis. First, I consider individuals who meet the same conditions as those in the baseline sample described above but were born between 1940 and 1969. This falsification sample consists of individuals who were entirely schooled in Catalonia but are too old to have been exposed to Catalan during their compulsory education. Second, I retain individuals born in the same cohorts as the baseline sample (1950–1990) but who were born outside Catalonia and migrated at age 16 or

[11] Specifically, I drop observations with missing values for average gross/net income (29 in 2013), oral skills in Catalan (10 observations), individual's completed education (39 observations) and parental origins (64 observations). However, for parental education, I treat the 236 missing values as a separate category. Additionally, for a robustness check, I incorporate information on working status and occupation, which results in 57 fewer observations due to missing data. Descriptive statistics for these variables are reported based only on the available observations.

older, thereby not being exposed to bilingual education during their compulsory schooling.

Descriptive statistics for the main variables used in the empirical analysis are presented in Table 1, for the pooled sample and separately by wave. Regarding the proxies for neighbourhood quality, the sample means of census tract average net and gross income per person (in 2016 prices) are approximately €12,500 and €15,200, respectively, while the average amount of unemployment benefits is €276. All monetary variables show an improvement across the two waves (i.e., higher income and lower unemployment benefits), driven by the general recovery of the economy following the financial crisis. Similar figures are obtained when using data at the census tract levels for all census tracts. Overall, 46% of individual in the estimation sample are Spanish native speakers and the average level of oral skills in Catalan is 8.9 (over 10).

The sample is relatively balanced by gender, with an average age of approximately 46. Regarding parental origins, 22% of the sample have one parent born outside Catalonia, while 34% have both parents born outside Catalonia. Additionally, 46% of the sample have parents whose highest level of education is primary schooling, while 17% have parents with no formal education. Among the respondents, 45% have a secondary education degree, while 30% have attained a university degree. At the time of the survey, 74% of the sample was employed, 12% was unemployed, and the remaining individuals were either retired or inactive. The share of unemployed individuals is lower in 2018, reflecting the general improvement in labour market conditions. Conversely, the proportion of retired individuals is higher in the second wave, likely due to the cohort restrictions applied in defining the estimation sample.

Table 1: descriptive statistics

	Pooled Sample		Wave 2013		Wave 2018	
Variable	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
outcomes						
net average income per capita (€)	12527.5	3075.4	12112.6	2810.2	12881.8	3243.6
gross average income per capita (€)	15242.4	4579.7	14608.2	4086.6	15783.9	4898.1
average unemployment benefits per capita (€)	275.82	79.97	319.12	76.33	238.85	62.56
ist	99.90	14.94	99.82	14.83	99.96	15.03
control variables						
wave 2013	0.461	0.498				
oral skills in Catalan	8.906	1.964	8.852	1.986	8.953	1.944
native Spanish speaker	0.464	0.499	0.450	0.498	0.475	0.499
male	0.480	0.500	0.486	0.500	0.474	0.499
age	45.82	10.89	43.13	10.53	48.11	10.67
parental origins						
both parents born in Catalonia	0.439	0.496	0.453	0.498	0.427	0.495
one parents born outside Catalonia	0.223	0.416	0.228	0.419	0.219	0.413
both parents born outside Catalonia	0.338	0.473	0.320	0.466	0.354	0.478
parental education						
no education	0.170	0.375	0.178	0.382	0.163	0.369
primary	0.465	0.499	0.469	0.499	0.462	0.499
secondary	0.214	0.410	0.208	0.406	0.219	0.413
tertiary	0.114	0.318	0.111	0.314	0.117	0.322
missing parental education	0.037	0.189	0.034	0.182	0.039	0.195
completed education						
primary or no education	0.242	0.429	0.248	0.432	0.238	0.426
secondary education	0.429	0.495	0.426	0.495	0.431	0.495
tertiary	0.291	0.454	0.288	0.453	0.292	0.455
other studies	0.038	0.192	0.037	0.190	0.039	0.193
labour force status						
employed	0.737	0.440	0.727	0.445	0.746	0.435
unemployed	0.116	0.321	0.159	0.365	0.080	0.271
retired	0.082	0.274	0.035	0.184	0.121	0.327
other	0.065	0.246	0.079	0.270	0.053	0.223
occupation (current or previous)						
managers	0.044	0.206	0.041	0.199	0.047	0.211
professionals	0.199	0.399	0.194	0.396	0.202	0.402
technicians and associate professionals	0.175	0.380	0.129	0.336	0.214	0.410
clerical support workers	0.073	0.260	0.112	0.315	0.040	0.195
service and sales workers	0.161	0.368	0.179	0.383	0.146	0.353
skilled agricultural, forestry and fishery workers	0.026	0.158	0.025	0.155	0.026	0.160
craft and related trades workers	0.112	0.315	0.117	0.321	0.108	0.311
plant and machine operators, and assemblers	0.080	0.271	0.071	0.256	0.087	0.282
elementary occupations	0.047	0.211	0.048	0.213	0.046	0.209
never worked before	0.084	0.277	0.084	0.278	0.084	0.277
number of observations	6959		3205		3754	

Table 2 reports the means and raw differences in the outcome variables by native language. Gross and net average income per capita, along with average unemployment benefits per capita at the census tract level, are log-transformed. For native Spanish speakers, the table also distinguishes between those who are

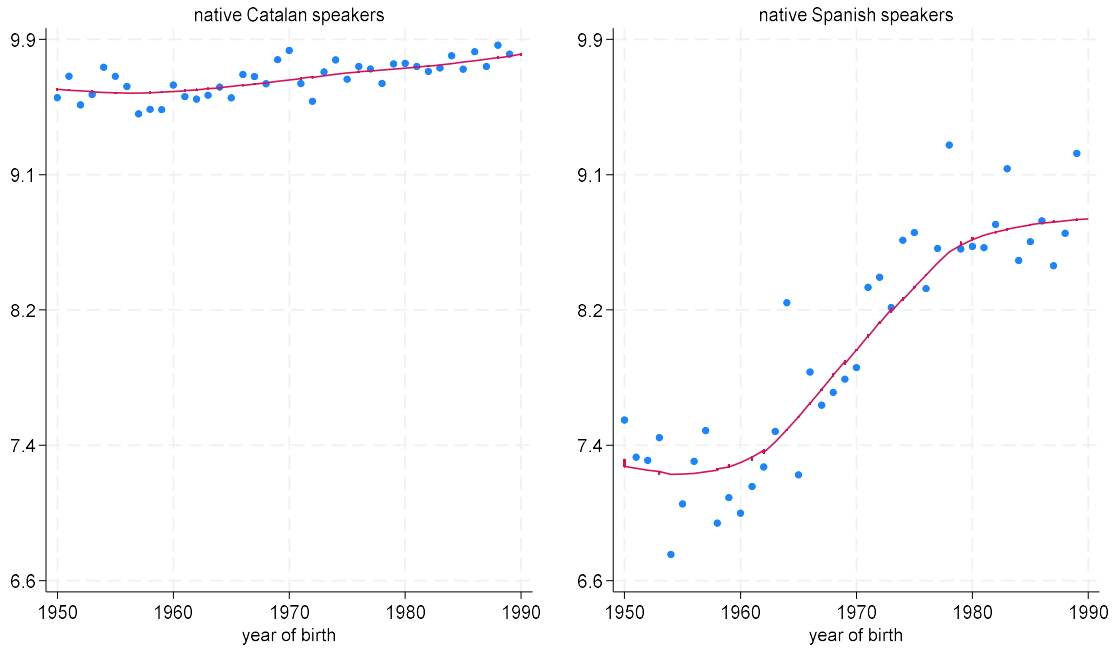
orally proficient in Catalan (with self-assessed skills of 8 or higher on a 10-point scale) and those who are not. Native Catalan speakers tend to reside in neighbourhoods with higher average per capita income, lower levels of unemployment benefits per person, and higher overall socioeconomic status compared to native Spanish speakers (column 5). However, these differences in local income and socioeconomic status disappear when focusing only on native Spanish speakers who are proficient in Catalan (column 6). Consistently, this subgroup enjoys better neighbourhood conditions than their counterparts who lack oral proficiency in Catalan (column 7). Nevertheless, this descriptive evidence should not be interpreted as causal. The following section outlines the identification strategy used to obtain estimates that can be more plausibly interpreted in causal terms.

Table 2: native language, language skills and neighbourhood quality

Variable:	native Catalan speakers (A)	native Spanish Speakers (B)	native Spanish speakers - proficient in Catalan (C)	native Spanish speakers - not proficient in Catalan (D)	A-B	C-A	D-C
ln(net income per capita)	9.422	9.395	9.417	9.343	0.027***	-0.005	-0.074***
ln(gross income per capita)	9.610	9.577	9.603	9.515	0.033***	-0.007	-0.088***
ln(unemployment benefits)	5.543	5.613	5.595	5.656	-0.070***	0.052***	0.061***
ist	100.65	99.00	100.40	95.66	1.652***	-0.245	-4.674***

The final descriptive evidence, shown in Figure 1, illustrates the evolution of average oral proficiency in Catalan across birth cohorts, separately for native Catalan speakers and native Spanish speakers. As can be observed, the former group reports consistently high and stable levels of oral proficiency across cohorts. In contrast, the latter group exhibits a clear upward trend, beginning with cohorts likely exposed to the language-in-education reform. As explained in the next section, this pattern provides the foundation for constructing the Instrumental Variable used in the analysis.

Figure 1: binned scatter plot for oral skills in Catalan by native language



Note: the red lines represent LOWESS fit curves.

3) Identification Strategy

The empirical analysis begins with the estimation of OLS regressions to explain four measures of neighbourhood quality: (i) average gross income per capita, (ii) average net income per capita, (iii) average unemployment benefits per capita (all logged), and (iv) a composite index of socioeconomic status. The model is specified as follows:

$$Y_{ci} = \alpha + \delta CATSK_i + \beta' X_i + \gamma' W_c + \theta_{\tau(i)} + \varepsilon_{ic} \quad if \ SPA_i = 1 \qquad (1)$$

Here, the outcome(s) of individual i residing in census tract c is regressed on the self-reported oral skills in Catalan ($CATSK_i$), a set of individual and family controls[12] (X_i), birth-cohort fixed effects ($\theta_{\tau(i)}$) and predetermined characteristics of the census tract of residence (W_i), including the latitude and longitude of the

[12] These control variables are introduced progressively using a stepwise approach. The initial specification includes only a wave dummy and a gender indicator. In the second step, the model is augmented with controls for parental origin and the highest level of education of the parents. Finally, the specification is further expanded by adding dummy variables for the individual's completed level of education.

tract's centroid and its total area.[13] The model is initially estimated using the subsample of individuals whose mother tongue is exclusively Spanish (SPA_i), since individuals from the other language group (only Catalan or Catalan and Spanish as native languages) are virtually fully proficient in Catalan. Since the outcome variables are defined at the census tract level, standard errors are clustered at that level to account for potential correlation of the error terms among individuals residing within the same census tract.

However, the OLS estimation of the coefficient of interest (δ) is likely to be biased and inconsistent due to several, potentially simultaneous, issues. First, there may be omitted variables that are correlated both with Catalan language skills and with the error term in equation (1). Second, language proficiency is self-reported and thus subject to measurement error. Third, the quality of the place of residence may directly influence Catalan proficiency — particularly if higher-quality neighbourhoods have more native Catalan speakers, facilitating language acquisition through social interaction. This introduces the possibility of reverse causality, further complicating identification. Hence, the OLS estimate of δ is unlikely to reflect a causal relationship.

To address these identification issues and obtain a causal estimate of the effect of oral Catalan skills among native Spanish speakers on neighbourhood quality, I adopt the Instrumental Variables (IV) approach used by Caminal and Di Paolo (2019) and Caminal et al. (2021). Specifically, I employ a variable that captures potential exposure to bilingual education during compulsory schooling, induced by the implementation of the LNA reform, which is defined as:

$$E_{\tau(i)} = \begin{cases} 10 & if\ \tau(i) \geq 1983 \\ 8 & if \quad 1977 \leq \tau(i) < 1983 \\ \tau(i) - 1969 & if \quad 1970 \leq \tau(i) < 1977 \\ 0 & if \quad \tau(i) < 1970 \end{cases}. \tag{2}$$

[13] The inclusion of these variables is intended to account for the spatial location of census tracts and to capture whether they are situated in rural, urban, or more densely populated areas. The size of the census tract serves as a proxy for population density, which is deliberately not included directly in the model to avoid concerns about potential reverse causality.

Accordingly, individuals born before 1970 were never exposed to Catalan during their compulsory education. Those born in 1977 or later completed all their compulsory schooling under the bilingual system. The duration of compulsory education was eight years for cohorts born up to 1983 and ten years for those born afterward, following the 1990 reform that extended compulsory schooling. Individuals born between 1970 and 1976 experienced partial exposure to bilingual education, with the extent of exposure depending on their exact year of birth. The exposure variable $E\tau_{(i)}$ is exogenous, as it depends solely on the individual's birth cohort. However, it may not serve as a valid instrument, since it could also capture cohort effects unrelated to the LNA reform, thereby violating the exclusion restriction. To construct a valid instrument, I exploit the heterogeneous impact of school-based language exposure by native language. This approach is motivated by evidence that the language-in-education reform had no significant effect on the oral Catalan skills of native Catalan speakers. To leverage this heterogeneous effect of exposure to the LNA reform and construct a more reliable instrument, the estimation sample is expanded to include native Catalan speakers. The corresponding model also includes a dummy variable for native language, yielding the following equation:

$$Y_{ci} = \alpha + \delta CATSK_i + \mu SPA_i + \beta' X_i + \gamma' W_c + \theta_{\tau(i)} + \varepsilon_{ic} \quad (3)$$

As shown below, this specification yields the same estimate of the δ coefficient as equation (1) and is equivalent to the estimate from an interaction between oral Catalan skills and the native language indicator. This equivalence arises because all variation in $CATSK_i$ in equation (3) is driven by native Spanish speakers. Consequently, the instrument used to obtain a consistent estimate of the effect of Catalan skills is the interaction between exposure to the reform, ($E\tau_{(i)}$), and the dummy variable for being a native Spanish speaker, (SPA_i). This leads to the following first-stage equation:

$$CATSK_i = \pi + \omega SPA_i + \rho(E_{\tau(i)} S_i) + \varphi' X_i + \psi' W_c + \sigma_{\tau(i)} + u_{ic} \quad (4)$$

The inclusion of native Catalan speakers in the estimation sample allows for flexible control of general cohort trends, captured by year of birth fixed effects. This specification ensures that the instrument ($E\tau_{(i)} \cdot SPA_i$) isolates the variation in language proficiency induced among native Spanish speakers by their differential exposure to Catalan at school due to the LNA reform, while accounting for non-language-related cohort trends. The key identifying assumption is that both language communities—native Catalan and native Spanish speakers—were subject to the same general cohort effects. Therefore, any cohort-specific changes observed among native Spanish speakers following the reform can be attributed to improvements in their Catalan language skills. Under this identification assumption, the IV estimate of the effect of oral Catalan skills among native Spanish speakers can be obtained by replacing the endogenous variable with its predicted values in the outcome equation, that is:

$$Y_{ci} = \alpha + \delta_{IV}\widehat{CATSK}_i + \mu SPA_i + \beta' X_i + \gamma' W_c + \theta_{\tau(i)} + \varepsilon_{ic} \qquad (5)$$

The IV estimate (δ_{IV}) identifies the causal effect of oral Catalan proficiency on neighbourhood quality among native Spanish speakers who improved their oral skills in Catalan as a consequence of their exposure to the LNA reform during compulsory education, representing a Local Average Treatment Effect (LATE). Because the level of variation of the instrument is by birth cohort, the standard errors are clustered by year of birth.[14]

Naturally, the validity of this IV approach is not without concerns. The most important issue is the possibility that the instrument captures spurious differential trends across cohorts that are directly related to the outcome, thereby violating the exclusion restriction. To address this concern, I conduct two

[14] As sensitivity check, I also repeat the estimation by adopting two-way clusters by year of birth and census tract. Moreover, because the outcome variables are clearly correlated among each other, I also check the coefficients' p-values obtained after correcting for multiple hypothesis testing, using the Romano-Wolf procedure (I run the bootstrap with 3000 replications).

falsification exercises using a reduced-form approach. The reduced-form equation is specified as follows:

$$Y_{ci} = \alpha + \lambda_{RF}(E_{\tau(i)}S_i) + \mu SPA_i + \beta' X_i + \gamma' W_c + \theta_{\tau(i)} + \nu_{ic} \quad (6)$$

In the first falsification exercise, I compare the reduced-form coefficient (λRF) with estimates from a series of placebo regressions. These regressions use a sample of older individuals born between 1940 and 1969 — either in Catalonia or elsewhere in Spain but who migrated to Catalonia at age six or earlier — and who were therefore never exposed to bilingual education during compulsory schooling. For these individuals, I construct a fake exposure variable by artificially shifting the year of LNA reform implementation backward in time, representing a placebo treatment. In the second falsification exercise, I use individuals from the same birth cohorts as the baseline sample (1950–1990) who migrated to Catalonia after reaching the legal age for compulsory education. I then impute to them the exposure variable "as if" they had completed their compulsory schooling in Catalonia.[15] In both cases, finding statistically significant and quantitatively meaningful reduced-form coefficients in the placebo regressions would indicate a violation of the identification assumption of common cohort trends between the two language groups, which would undermine the validity of the exclusion restriction.

A second issue that may threaten the internal validity of the identification strategy concerns the presence of potential confounders, related to the interplay between language of instruction, identity formation, and social networks. First, one may argue that exposure to Catalan in school could have influenced identity — specifically, that native Spanish speakers might develop a stronger identification as Catalan — thereby facilitating access to better networks and connections. To rule out the possibility that the results are driven by changing

[15] In this second falsification sample, there are few native Catalan speakers, who are either individuals originating from other Catalan-speaking areas (e.g., Valencia, the Balearic Islands, northeastern Aragón, and the Catalan-speaking area of Perpinyà in France) or, more generally, children of Catalan-speaking parents residing outside Catalonia.

identity patterns, I re-estimate the model using a restricted sample consisting only of individuals whose self-identified language at the time of the survey matches their native language, thereby excluding those who have switched their identity language relative to their linguistic background. Second, I restrict the sample to individuals whose partners share the same mother tongue. This approach aims to discard the role of couple formation, specifically, the possibility that native Spanish speakers exposed to the reform are more likely to live in affluent neighbourhoods because they are more likely to partner with Catalan speakers, a dynamic related to the findings of Caminal and Di Paolo (2019).[16] Finally, to provide additional suggestive evidence aimed at excluding the possibility that the baseline results are driven by personal networks, I use information on the language most frequently spoken with friends to test the robustness of the results by restricting the sample to individuals who predominantly use their mother tongue in social interactions with friends.[17]

A third aspect that deserves consideration relates to youth emancipation, which tends to occur relatively late in Spain compared to other European countries (Ayllón, 2015). This could be problematic, as individuals exposed to the LNA reform are younger, and some — particularly those in the most recent cohorts of the estimation sample — may still have been living with their parents at the time of the survey. Although the survey does not directly report whether respondents live with their parents, I construct a proxy based on information about household size, number of children, and couple status. Specifically, I consider an individual unlikely to be living with their parents if: (i) they live alone; or (ii) the household size corresponds to the expected number of household members based on their couple status and number of children (if any).

[16] Moreover, I further restrict the sample to individuals who not only have a partner with the same mother tongue but are also living with their partner.

[17] Specifically, I classify a native Spanish speaker as using only Spanish with friends if they report speaking exclusively Spanish, more Spanish than Catalan, or Spanish and other languages (excluding Catalan) in their social interactions. A symmetric definition is applied to native Catalan speakers. For the small number of individuals who report both Spanish and Catalan as their native languages, I consider them to use the same language(s) with friends if they report speaking Spanish and Catalan equally in social contexts. Relatedly, I conduct an additional robustness check by excluding individuals who report having a mixed mother tongue (both Spanish and Catalan).

Therefore, I repeat the estimation using only observations of individuals who, based on the available information, are unlikely to be living with their parents.[18]

Finally, I consider the role of labour market status as potential channels, in line with the findings of Cappellari and Di Paolo (2018), who showed that the implementation of the LNA reform increased the wage returns to education in Catalonia. Specifically, I re-estimate the model including additional control variables in the form of dummies for economic activity status and for current or previous occupation. Although these variables are likely to be "bad controls," observing a significant change in the coefficient of Catalan skills after their inclusion would suggest that labour market status plays an important role as an underlying channel in the relationship between local language skills and neighbourhood quality.

4) Results

Table 3 presents the OLS estimation results of equation (1) for the four outcome variables. The estimates reported in columns (1), (4), (7), and (10) are based on a basic set of control variables, including wave and gender dummies, year-of-birth fixed effects, as well as the latitude, longitude, and area of the census tract centroids. As shown, the oral Catalan skills of native Spanish speakers are positively associated with both measures of local income per capita and with the socioeconomic status index, and negatively associated with the average per capita amount of unemployment benefits in the census tract. Next, I control for parental education and origin. As expected, the inclusion of family background variables attenuates the conditional correlation between oral proficiency in Catalan and the outcome variables, although the sign and statistical significance of the corresponding coefficients remain unchanged. Additionally, the estimates for parental characteristics align with expectations. Specifically, individuals whose parents were both born in Catalonia tend to reside in more affluent

[18] A related issue, which cannot be directly addressed with the available data, is the possibility that the individual resides in an apartment inherited from their parents. Unfortunately, the EULP survey does not include information on home ownership or tenancy status.

neighbourhoods — particularly in comparison to those whose parents were both born outside Catalonia — with the exception of the socioeconomic status index, for which the estimates are not statistically significant. Likewise, parental education is positively associated with all measures of neighbourhood quality. Finally, for each outcome, the last columns of Table 3 present the results obtained after including individuals' educational attainment. The coefficients on oral Catalan skills are substantially reduced, indicating that part of the association operates through education. Indeed, education emerges as a significant predictor of neighbourhood quality. Nevertheless, oral language skills in Catalan among native Spanish speakers remain positively associated with better neighbourhood conditions.

Table 3: OLS estimations for native Spanish speakers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
outcome:	ln(net income per capita)			ln(gross income per capita)			ln(unemployment benefits)			ist		
constant	8.706***	8.390***	8.109***	8.656***	8.278***	7.947***	5.161***	5.400***	5.521***	121.637	105.614	88.436
	(0.902)	(0.784)	(0.726)	(1.074)	(0.928)	(0.858)	(1.869)	(1.990)	(2.044)	(76.850)	(69.512)	(64.860)
oral skills in Catalan	0.017***	0.012***	0.007***	0.020***	0.014***	0.008***	-0.013***	-0.010***	-0.008***	1.087***	0.813***	0.489***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.159)	(0.152)	(0.148)
wave 2013	-0.048***	-0.049***	-0.050***	-0.060***	-0.062***	-0.063***	0.310***	0.312***	0.312***	0.164	0.130	0.079
	(0.014)	(0.013)	(0.012)	(0.016)	(0.015)	(0.015)	(0.015)	(0.014)	(0.014)	(0.982)	(0.946)	(0.926)
male	-0.002	-0.006	-0.000	-0.002	-0.007	0.001	0.011	0.014*	0.010	0.071	-0.154	0.193
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.554)	(0.540)	(0.537)
both parents born in Catalonia						reference category						
one parents born outside Catalonia		-0.036*	-0.037*		-0.041*	-0.043*		0.055***	0.056***		-1.157	-1.204
		(0.020)	(0.019)		(0.024)	(0.023)		(0.021)	(0.021)		(1.213)	(1.187)
both parents born outside Catalonia		-0.044**	-0.052***		-0.054**	-0.063***		0.066***	0.070***		-1.367	-1.811
		(0.020)	(0.020)		(0.024)	(0.023)		(0.021)	(0.021)		(1.217)	(1.196)
parental education = no education						reference category						
parental education = primary		0.046***	0.032***		0.053***	0.037***		-0.034***	-0.029***		3.236***	2.339***
		(0.010)	(0.009)		(0.011)	(0.011)		(0.010)	(0.010)		(0.724)	(0.715)
parental education = secondary		0.076***	0.040***		0.092***	0.049***		-0.052***	-0.037**		5.107***	2.841***
		(0.013)	(0.013)		(0.016)	(0.015)		(0.016)	(0.015)		(0.943)	(0.936)
parental education = tertiary		0.256***	0.194***		0.305***	0.232***		-0.181***	-0.152***		14.442***	10.727***
		(0.027)	(0.024)		(0.032)	(0.029)		(0.027)	(0.025)		(1.343)	(1.277)
parental education = missing		0.042**	0.039**		0.050**	0.046**		-0.019	-0.018		2.038	1.833
		(0.018)	(0.018)		(0.021)	(0.021)		(0.021)	(0.021)		(1.396)	(1.361)
completed education = primary						reference category						
completed education = secondary			0.057***			0.066***			-0.008			3.957***
			(0.010)			(0.011)			(0.011)			(0.736)
completed education = tertiary			0.134***			0.159***			-0.058***			8.229***
			(0.014)			(0.016)			(0.016)			(0.939)
adjusted R-squared	0.061	0.139	0.171	0.062	0.144	0.176	0.317	0.346	0.350	0.044	0.094	0.119
number of observations	3201	3201	3201	3201	3201	3201	3201	3201	3201	3201	3201	3201

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Standard errors clustered at the census tract level in parenthesis. All regressions include as control variables year of birth dummies, census tract's area and latitude and longitude of the corresponding centroids.

Using the full specification, I re-estimate the model by including native Catalan speakers in the estimation sample, as outlined in equation (3). The results, presented in Panel B of Table 4, show that the coefficients on oral Catalan skills are virtually identical to those obtained when restricting the sample to native Spanish speakers only.

Table 4: OLS estimations for native Spanish and Catalan speakers

outcome:	(1) ln(net income per capita)	(2) ln(gross income per capita)	(3) ln(unemployment benefits)	(4) ist
Panel A: only native Spanish speakers				
oral skills in Catalan	0.007***	0.008***	-0.008***	0.489***
	(0.002)	(0.002)	(0.002)	(0.148)
adjusted R-squared	0.171	0.176	0.350	0.119
number of observations	3201	3201	3201	3201
Panel B: native Spanish speakers & native Catalan speakers				
oral skills in Catalan	0.007***	0.008***	-0.007***	0.477***
	(0.002)	(0.002)	(0.002)	(0.132)
native Spanish speaker	0.020*	0.023*	0.003	1.088*
	(0.011)	(0.013)	(0.013)	(0.652)
adjusted R-squared	0.191	0.197	0.277	0.128
number of observations	6920	6920	6920	6920

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Standard errors clustered at the census tract level in parenthesis. All regressions include as control variables year of birth dummies, census tract's area and latitude and longitude of the corresponding centroids.

This equivalence arises because native Spanish speakers account for all the variation in oral Catalan skills. Moreover, once individual and parental characteristics are controlled for, native language shows little association (in terms of statistical significance) with neighbourhood quality. Overall, the results suggest that oral Catalan skills among native Spanish speakers exhibit a statistically significant, albeit quantitatively modest, conditional correlation with neighbourhood quality. However, the potential endogeneity of language skills precludes a causal interpretation of the estimates presented thus far.

To address the endogeneity concerns outlined in Section 3 and obtain consistent estimates of the effect of local language skills on neighbourhood quality, I implement an identification strategy that leverages only the variation in oral Catalan skills among native Spanish speakers induced by the introduction of bilingual education under the LNA reform. Specifically, Table 5 presents the

results of the IV/TSLS estimation of equation (5), where the instrument is constructed as the interaction between the number of years of potential exposure to Catalan during compulsory education and an indicator for being a native Spanish speaker. Panel A of Table 5 reports the first-stage estimates obtained using the three previously described sets of control variables. In all specifications, the instrument exhibits a positive and highly significant effect on oral Catalan skills, with the corresponding F-statistics indicating strong instrument relevance.

Table 5: IV/TSLS estimations

	(1)	(2)	(3)
Panel A: First Stage			
outcome = oral skills in Catalan			
exposure × I(native Spanish speaker)	0.130***	0.122***	0.117***
	(0.011)	(0.011)	(0.010)
adjusted R-squared	0.218	0.242	0.281
F-stat (instrument)	95.527	79.072	77.080
Panel B: 2SLS estimation			
outcome = ln(net income per capita)			
oral skills in Catalan	-0.017	-0.004	-0.007
	(0.012)	(0.012)	(0.012)
outcome = ln(gross income per capita)			
oral skills in Catalan	-0.021	-0.006	-0.009
	(0.014)	(0.013)	(0.014)
outcome = ln(unemployment benefits)			
oral skills in Catalan	-0.005	-0.003	-0.002
	(0.012)	(0.013)	(0.013)
outcome = ist			
oral skills in Catalan	-0.801	-0.116	-0.310
	(0.727)	(0.734)	(0.751)
number of observations		6920	

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area and latitude and longitude of the corresponding centroids. Model (1) includes wave and gender dummies. Model (2) includes dummies for wave, gender, parental origins and parents' education. Model (3) includes dummies for wave, gender, parental origins, parents' education and respondent's education.

The point estimate decreases only marginally as additional control variables are included. Panel B presents the estimated coefficients of interest for the four outcome variables. The results suggest that the increase in oral Catalan skills among native Spanish speakers — driven by the implementation of the language-in-education policy — does not have a statistically significant causal effect on neighbourhood quality. The point estimates are quantitatively small (similar to those obtained using OLS) but have a negative sign and are associated with large standard errors, which cannot be attributed to the use of a weak instrument.[19] Overall, these results point to a null effect of oral skills in Catalan induced among native Spanish by exposure to bilingualism at school on neighbourhood quality.

To support the internal validity of the results and the conclusions drawn from them, I next present findings from falsification exercises and additional robustness checks. First, Tables 6a, 6b, 6c and 6d report the results of the initial falsification test, which uses cohorts of never-treated individuals and assigns them placebo reforms dated several years prior to the actual policy implementation.[20] The first column of each table reports the reduced-form coefficient from equation (6), which is virtually zero across all outcomes. This suggests that years of exposure to the LNA reform among native Spanish speakers have no direct effect on neighbourhood quality — consistent with the IV/TSLS estimates.

[19] Table A2 in the Appendix reports the p-values of the coefficients of interest, obtained using two-way clustered standard errors at the year-of-birth and census tract levels, as well as after correcting for multiple hypothesis testing using the Romano-Wolf procedure, confirming their lack of statistical significance.

[20] The years of implementation of the placebo reforms (1970–1963) are chosen to ensure a sufficient number of observations for each pseudo-treated and control cohort.

Table 6a: falsification 1 for logged average gross income per capita

	real reform	placebo reforms in:							
	in 1983	1970	1969	1968	1967	1966	1965	1964	1963
exposure × I(native Spanish speaker)	-0.001								
	(0.001)								
exposure* × I(native Spanish speaker)		-0.002							
		(0.002)							
exposure* × I(native Spanish speaker)			-0.001						
			(0.002)						
exposure* × I(native Spanish speaker)				-0.001					
				(0.002)					
exposure* × I(native Spanish speaker)					-0.001				
					(0.002)				
exposure* × I(native Spanish speaker)						-0.001			
						(0.002)			
exposure* × I(native Spanish speaker)							-0.001		
							(0.001)		
exposure* × I(native Spanish speaker)								-0.001	
								(0.001)	
exposure* × I(native Spanish speaker)									-0.002
									(0.002)
R-squared	0.195	0.212	0.212	0.212	0.212	0.212	0.212	0.212	0.212
number of observations	6920	4393	4393	4393	4393	4393	4393	4393	4393

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education.

Table 6b: falsification 1 for logged average net income per capita

	real reform	placebo reforms in:							
	in 1983	1970	1969	1968	1967	1966	1965	1964	1963
exposure × I(native Spanish speaker)	-0.001								
	(0.002)								
exposure* × I(native Spanish speaker)		-0.002							
		(0.002)							
exposure* × I(native Spanish speaker)			-0.001						
			(0.002)						
exposure* × I(native Spanish speaker)				-0.001					
				(0.002)					
exposure* × I(native Spanish speaker)					-0.001				
					(0.002)				
exposure* × I(native Spanish speaker)						-0.001			
						(0.002)			
exposure* × I(native Spanish speaker)							-0.001		
							(0.002)		
exposure* × I(native Spanish speaker)								-0.001	
								(0.002)	
exposure* × I(native Spanish speaker)									-0.002
									(0.002)
R-squared	0.200	0.220	0.220	0.220	0.220	0.220	0.220	0.220	0.220
number of observations	6920	4393	4393	4393	4393	4393	4393	4393	4393

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education.

Table 6c: falsification 1 for logged average gross unemployment benefits per capita

	real reform	placebo reforms in:							
	in 1983	1970	1969	1968	1967	1966	1965	1964	1963
exposure × I(native Spanish speaker)	-0.000								
	(0.002)								
exposure* × I(native Spanish speaker)		-0.002							
		(0.002)							
exposure* × I(native Spanish speaker)			-0.003						
			(0.002)						
exposure* × I(native Spanish speaker)				-0.002					
				(0.002)					
exposure* × I(native Spanish speaker)					-0.002				
					(0.002)				
exposure* × I(native Spanish speaker)						-0.002			
						(0.002)			
exposure* × I(native Spanish speaker)							-0.002		
							(0.002)		
exposure* × I(native Spanish speaker)								-0.002	
								(0.002)	
exposure* × I(native Spanish speaker)									-0.001
									(0.002)
R-squared	0.281	0.272	0.272	0.272	0.272	0.272	0.272	0.272	0.272
number of observations	6920	4393	4393	4393	4393	4393	4393	4393	4393

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education.

Table 6d: falsification 1 for the index of socioeconomic status

	real reform	placebo reforms in:							
	in 1983	1970	1969	1968	1967	1966	1965	1964	1963
exposure × I(native Spanish speaker)	-0.036								
	(0.089)								
exposure* × I(native Spanish speaker)		-0.072							
		(0.110)							
exposure* × I(native Spanish speaker)			-0.058						
			(0.109)						
exposure* × I(native Spanish speaker)				-0.050					
				(0.104)					
exposure* × I(native Spanish speaker)					-0.036				
					(0.099)				
exposure* × I(native Spanish speaker)						-0.036			
						(0.091)			
exposure* × I(native Spanish speaker)							-0.030		
							(0.086)		
exposure* × I(native Spanish speaker)								-0.022	
								(0.085)	
exposure* × I(native Spanish speaker)									-0.017
									(0.088)
R-squared	0.132	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147
number of observations	6920	4393	4393	4393	4393	4393	4393	4393	4393

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education.

However, finding non-zero coefficients in the placebo reduced-form regressions would indicate the presence of spurious cohort trends that differ between the two language groups. Reassuringly, this is not the case: the coefficients associated with the placebo years of exposure are consistently small and statistically insignificant across all outcomes. Similarly, Table 7 presents the results of the second falsification exercise, which focuses on individuals born outside Catalonia who migrated to the region after completing compulsory education. For these individuals, placebo years of exposure to Catalan at school are imputed under the assumption that they were entirely educated in Catalonia. Once again, the results from this alternative falsification exercise confirm that the main findings are not affected by spurious heterogeneous trends across birth cohorts.

Table 7: falsification 2

outcome:	ln(net income per capita)		ln(gross income per capita)		ln(unemployment benefits)		ist	
	baseline	placebo	baseline	placebo	baseline	placebo	baseline	placebo
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
exposure × I(native Spanish speaker)	-0.001	-0.003	-0.001	-0.002	-0.000	-0.009	-0.036	-0.576
	(0.001)	(0.008)	(0.002)	(0.010)	(0.002)	(0.008)	(0.089)	(0.587)
R-squared	0.195	0.114	0.200	0.117	0.281	0.296	0.132	0.130
number of observations	6920	1459	6920	1459	6920	1459	6920	1459

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education.

With the aim of ruling out the effect of potential confounders related to identity, partnership formation and social networks, and parental co-residence, in Table 8 I present the reduced-form coefficients and the second-stage estimates obtained after: (a) retaining individuals whose self-identification language matches their native language (column 2); (b) retaining those who have a partner with the same native language (column 3) and who also live with their partner (column 4); (c)

selecting individuals who tend to use their native language with friends (column 5); and (d) excluding those likely to live with their parents.[21]

Table 8: potential confounders

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: reduced-form estimation						
outcome = ln(net income per capita)						
exposure × I(native Spanish speaker)	-0.001	-0.000	-0.002	-0.003	0.001	-0.000
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
outcome = ln(gross income per capita)						
exposure × I(native Spanish speaker)	-0.001	-0.000	-0.002	-0.003	0.001	-0.000
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
outcome = ln(unemployment benefits)						
exposure × I(native Spanish speaker)	-0.000	0.001	0.000	-0.000	-0.001	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
outcome = ist						
exposure × I(native Spanish speaker)	-0.036	0.043	-0.107	-0.135	0.094	-0.007
	(0.089)	(0.112)	(0.155)	(0.166)	(0.123)	(0.117)
Panel B: 2SLS estimation						
outcome = ln(net income per capita)						
oral skills in Catalan	-0.007	-0.002	-0.014	-0.016	0.005	-0.001
	(0.012)	(0.009)	(0.016)	(0.016)	(0.007)	(0.015)
outcome = ln(gross income per capita)						
oral skills in Catalan	-0.009	-0.003	-0.015	-0.017	0.006	-0.003
	(0.014)	(0.010)	(0.018)	(0.018)	(0.009)	(0.018)
outcome = ln(unemployment benefits)						
oral skills in Catalan	-0.002	0.005	0.002	-0.000	-0.004	-0.013
	(0.013)	(0.010)	(0.016)	(0.016)	(0.008)	(0.015)
outcome = ist						
oral skills in Catalan	-0.310	0.235	-0.745	-0.870	0.399	-0.057
	(0.751)	(0.601)	(1.059)	(1.057)	(0.519)	(0.960)
F-stat (instrument)	77.080	114.137	81.656	77.129	112.243	59.175
number of observations	6920	5586	4059	3236	4555	4974

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education. Column (1): baseline sample. Column (2): only individuals with mother tongue = identity languge. Column (3): only individuals with mother tongue = partner's language. Column (4): only individuals with mother tongue = partner's language, who are living with their partner. Column (5): only individuals with mother tongue = language with friends. Column (6): excluding individuals who are likely to live with their parents.

Finally, to assess the relevance of labour market status as a potential channel, I re-estimate the model including dummies for labour force status and (current or previous) occupational categories as additional control variables. The results, presented in Table 9, show that controlling for labour market status does not alter

[21] Table 3A in the Appendix also shows that the results remain unchanged when individuals reporting a mixed native language (Catalan and Spanish) are excluded from the sample.

the baseline findings. This is an expected result, given both the lack of a significant effect of oral Catalan skills — induced by the LNA reform among native Spanish speakers — on neighbourhood quality, and the fact that labour market status would only be a relevant mediating channel if such an effect were present. This interpretation is also consistent with prior evidence documenting a positive relationship between local language skills and labour market outcomes, as well as between favourable labour market status and residential quality.

Table 9: additional estimations controlling for labour market status

	(1)	(2)
Panel A: First Stage		
outcome = oral skills in Catalan		
exposure × I(native Spanish speaker)	0.117***	0.112***
	(0.010)	(0.010)
adjusted R-squared	0.281	0.296
F-stat (instrument)	77.080	78.985
Panel B: 2SLS estimation		
outcome = ln(net income per capita)		
oral skills in Catalan	-0.007	-0.012
	(0.012)	(0.012)
outcome = ln(gross income per capita)		
oral skills in Catalan	-0.009	-0.014
	(0.014)	(0.014)
outcome = ln(unemployment benefits)		
oral skills in Catalan	-0.002	0.001
	(0.013)	(0.014)
outcome = ist		
oral skills in Catalan	-0.310	-0.598
	(0.751)	(0.774)
number of observations	6920	6863

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education. Regression in column (2) additionally includes controls for working condition and (current or previous) occupation.

5) Conclusions

This paper investigates whether skills in a local language affect neighbourhood quality in a bilingual region. I focus on the Spanish region of Catalonia, where there are two main native language communities: native Catalan speakers, who are bilingual in Catalan — the region's vernacular language — and Spanish, and native Spanish speakers, among whom proficiency in Catalan is not universal.

The empirical analysis relies on two combined data sources. Specifically, I merge survey microdata—containing rich information on respondents' sociodemographic and linguistic characteristics, as well as their census tract of residence—with census tract-level data on average gross and net income per capita, the average amount of unemployment benefits per capita, and a composite index of socioeconomic status. These census tract-level indicators of neighbourhood quality serve as the outcome variables in the analysis. OLS estimates suggest that oral Catalan skills among native Spanish speakers are associated with better residential quality: specifically, they are positively related to local income and socioeconomic status, and negatively related to unemployment benefits per capita, which is taken as an indicator of dependency on the welfare state.

To address the endogeneity of oral language skills, I adopt an instrumental variable approach that exploits variation in Catalan proficiency among native Spanish speakers induced by exposure to Catalan-Spanish bilingualism during compulsory education, following the implementation of a language-in-education reform in the 1980s. Specifically, I include native Catalan speakers in the estimation sample and, considering that their oral proficiency in Catalan was unaffected by the reform, I construct an instrument based on the interaction between years of potential exposure to bilingual education during compulsory schooling and an indicator for being a native Spanish speaker. The key identifying assumption is that cohort effects unrelated to the reform are homogeneous across native Spanish and native Catalan speakers. Under this

assumption, the IV/TSLS estimation identifies the effect of Catalan proficiency among native Spanish speakers whose language skills improved due to exposure to Catalan during compulsory education.

The results from the instrumental variable approach indicate that the increase in oral Catalan skills induced by the reform among native Spanish speakers does not translate into improvements in residential quality for any of the outcomes considered in this study. The findings are supported by a series of falsification exercises designed to rule out spurious heterogeneous cohort trends that would undermine the validity of the exclusion restriction, as well as by robustness checks addressing potential confounders and alternative mechanisms related to identity, social networks, and labour market outcomes. Overall, the evidence presented in this paper suggests that, although the language-in-education policy successfully promoted bilingualism and has been associated with various positive socioeconomic effects, it did not influence residential sorting and, therefore, did not mitigate existing language-related disparities or residential segregation in the bilingual region of Catalonia.

References

Ayllón, S. (2015). Youth poverty, employment, and leaving the parental home in Europe. *Review of Income and Wealth*, *61*(4), 651-676.

Aoki, Y., & Santiago, L. (2024). Where to live? English proficiency and residential location of UK migrants. *Journal of Economic Behavior & Organization*, *221*, 73-93.

Beckhusen, J., Florax, R. J., Poot, J., & Waldorf, B. (2013). Living and working in ethnic enclaves: English language proficiency of immigrants in US metropolitan areas. *Papers in Regional Science*, *92*(2), 305-329.

Bleakley, H., & Chin, A. (2010). Age at arrival, English proficiency, and social assimilation among US immigrants. *American Economic Journal: Applied Economics*, *2*(1), 165-192.

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

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.

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

Chyn, E., & Katz, L. F. (2021). Neighborhoods matter: Assessing the evidence for place effects. *Journal of Economic Perspectives*, *35*(4), 197-222.

Danzer, A. M., & Yaman, F. (2016). Ethnic concentration and language fluency of immigrants: Evidence from the guest-worker placement in Germany. *Journal of Economic Behavior & Organization*, *131*, 151-165.

Danzer, A. M., Feuerbaum, C., Piopiunik, M., & Woessmann, L. (2022). Growing up in ethnic enclaves: language proficiency and educational attainment of immigrant children. *Journal of Population Economics*, *35*(3), 1297-1344.

Diamond, R., & Gaubert, C. (2022). Spatial sorting and inequality. *Annual Review of Economics*, *14*(1), 795-819.

Di Paolo, A., & Raymond, J. L. (2012). Language knowledge and earnings in Catalonia. *Journal of Applied Economics*, *15*(1), 89-118.

Garcia-López, M. À., Nicolini, R., & Roig Sabaté, J. L. (2021). Urban spatial structure in Barcelona (1902–2011): Immigration, spatial segregation and new centrality governance. *Applied Spatial Analysis and Policy*, *14*(3), 591-629.

Laliberté, J. W. (2019). Language skill acquisition in immigrant social networks: Evidence from Australia. *Labour Economics*, *57*, 35-45.

Rendon, S. (2007). The Catalan premium: language and employment in Catalonia. *Journal of Population Economics*, *20*(3), 669-686.

Appendix

Table A1: descriptive statistics for placebo samples

	Placebo Sample 1		Placebo Sample 2	
Variable	Mean	Std. dev.	Mean	Std. dev.
outcomes				
net average income per capita (€)	12680.5	3300.3	12306.0	3165.6
gross average income per capita (€)	15467.9	4934.7	14999.5	4697.5
average unemployment benefits per capita (€)	271.78	80.10	287.65	81.37
ist	100.01	15.36	97.91	16.16
control variables				
wave 2013	0.468	0.499	0.418	0.493
oral skills in Catalan	8.725	2.199	3.891	3.155
native Spanish speaker	0.386	0.487	0.971	0.169
male	0.480	0.500	0.439	0.496
age	58.88	8.76	46.87	11.22
parental origins				
both parents born in Catalonia	0.511	0.500	0.001	0.037
one parents born outside Catalonia	0.176	0.380	0.010	0.101
both parents born outside Catalonia	0.314	0.464	0.988	0.107
parental education				
no education	0.243	0.429	0.151	0.359
primary	0.511	0.500	0.395	0.489
secondary	0.148	0.355	0.248	0.432
tertiary	0.065	0.246	0.179	0.383
missing parental education	0.034	0.181	0.027	0.161
completed education				
primary or no education	0.391	0.488	0.263	0.440
secondary education	0.402	0.490	0.432	0.496
tertiary	0.206	0.405	0.305	0.461
number of observations	4393		1459	

Table A2: robustness to two-way clusters and multiple hypothesis testing

p-value:	baseline	two-way clusters	resample	Romano-Wolf
Panel A: reduced-form estimation				
	outcome = ln(net income per capita)			
exposure × I(native Spanish speaker)	0.5463	0.555	0.4252	0.7021
	outcome = ln(gross income per capita)			
exposure × I(native Spanish speaker)	0.5142	0.522	0.3905	0.6628
	outcome = ln(unemployment benefits)			
exposure × I(native Spanish speaker)	0.8659	0.860	0.8234	0.8271
	outcome = ist			
exposure × I(native Spanish speaker)	0.6858	0.695	0.5951	0.8271
Panel B: 2SLS estimation				
	outcome = ln(net income per capita)			
oral skills in Catalan	0.5424	0.558	0.4235	0.7011
	outcome = ln(gross income per capita)			
oral skills in Catalan	0.5107	0.525	0.3865	0.6711
	outcome = ln(unemployment benefits)			
oral skills in Catalan	0.8624	0.863	0.8211	0.8211
	outcome = ist			
oral skills in Catalan	0.6799	0.699	0.5978	0.8194

Note: All regressions include as control variables year of birth dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education.

Table A3: robustness to the exclusion of individuals with mixed native languages

	(1)	(2)
Panel A: reduced-form estimation		
outcome = ln(net income per capita)		
exposure × I(native Spanish speaker)	-0.001	-0.001
	(0.001)	(0.001)
outcome = ln(gross income per capita)		
exposure × I(native Spanish speaker)	-0.001	-0.001
	(0.002)	(0.002)
outcome = ln(unemployment benefits)		
exposure × I(native Spanish speaker)	-0.000	-0.000
	(0.002)	(0.002)
outcome = ist		
exposure × I(native Spanish speaker)	-0.036	-0.018
	(0.089)	(0.092)
Panel B: 2SLS estimation		
outcome = ln(net income per capita)		
oral skills in Catalan	-0.007	-0.006
	(0.012)	(0.012)
outcome = ln(gross income per capita)		
oral skills in Catalan	-0.009	-0.007
	(0.014)	(0.014)
outcome = ln(unemployment benefits)		
oral skills in Catalan	-0.002	-0.002
	(0.013)	(0.013)
outcome = ist		
oral skills in Catalan	-0.310	-0.158
	(0.751)	(0.778)
F-stat (instrument)	77.080	79.426
number of observations	6920	6668

Note: *** significant at 1%; ** significant at 5%; * significant at 10%. Robust standard errors clustered by year of birth in parenthesis. All regressions include as control variables age dummies, census tract's area, latitude and longitude of the corresponding centroids, as well as dummies for wave, gender, parental origins, parents' education and respondent's education. Column (1): baseline sample. Column (2): excluding individuals with mixed (Catalan & Spanish) mother tongue.

UBIREA

Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

WEBSITE: www.ub.edu/irea • **CONTACT:** irea@ub.edu

AQR

Grup de Recerca Anàlisi Quantitativa Regional
Regional Quantitative Analysis Research Group

WEBSITE: www.ub.edu/aqr/ • **CONTACT:** aqr@ub.edu

Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona