Language-in-education, language skills and the intergenerational transmission of language in a bilingual society[☆]

Ramon Caminal^a, Lorenzo Cappellari^b, A. Di Paolo^{c,*}

^a Institute of Economic Analysis, CSIC, and Barcelona GSE

^b Catholic University of Milan and LISER

^c AQR-IREA, University of Barcelona

ARTICLE INFO

JEL codes:

I28
J13
J24
J62
Z13

Keywords:

Language
Intergenerational transmission
Language proficiency
Language-in-education reform

ABSTRACT

We investigate the causal effect of parents' linguistic skills on the intergenerational language transmission in a bilingual society. We consider the case of Catalonia, where the two main speech communities, Spanish and Catalan, are of similar sizes, and both languages are official, and exploit the natural experiment generated by a language-in-education reform. We discuss theoretically the main alternative channels through which such policies may affect the language parents speak to their children. Empirically, we find that the reform improved oral skills in Catalan among native Spanish speakers. These greater skills increased the propensity to speak Catalan with the offspring, which highlights the intergenerational spillovers of the reform. We also show that the effect is not confounded by spurious trends, potential changes in language identity, and linguistically mixed partnership formation.

1. Introduction

Language skills impinge on many social and economic outcomes. In particular, it has been extensively documented that immigrants proficient in the host-country language enjoy enhanced employability and substantial wage premia. Bleakley and Chin (2004) and Chiswick and Miller (2007) are just two prominent examples of a fairly large and geographically diverse literature.¹ Thus, policies aimed at promoting additional language skills are likely to alleviate wage inequality and raise efficiency, owing to the positive externality generated by the acquisition of a second language (Selten and Pool, 1991; Church and King, 1993).

A few studies have also examined the transmission of language skills between first- and second-generation immigrants. More specifically, Bleakley and Chin (2008) and Casey and Dustmann (2008) for the cases of the US and the UK, respectively, found that the parents' English-speaking proficiency is positively correlated (in a causal sense) to children's English proficiency. These findings imply that the family language background influences the linguistic skills of the next generation; that is, the language deficiencies of the first-generation immigrants

[☆] We thank seminar participants at JRC-Ispira, University of Cagliari, EALE, Lisbon Workshop on the Economics, Statistics and Econometrics of Education, International Conference on Education Economics, SEHO, AIEL, ESPE, SAEe, WEAI, and the LEER Workshop on Education Economics for providing useful comments and suggestions, as well as the IDESCAT staff for their support with the data. We gratefully acknowledge funding from: Spanish Ministry of Finance, Industry, and Competitiveness ECO2017–88129-P (AEI/FEDER, UE), Severo Ochoa Programme for Centres of Excellence in R&D SEV-2015–0563 and Generalitat de Catalunya 2017-SGR-1136 (Caminal); Università Cattolica D32 ESEM strategic research grant (Cappellari); Spanish Ministry of Finance, Industry, and Competitiveness ECO2016–75805-R, RecerCaixa 2016ACUP00020 (Di Paolo). The usual disclaimers apply.

* Corresponding author.

E-mail address: antonio.dipaolo@ub.edu (A. Di Paolo).

¹ In a similar vein, starting with Frankel and Rose (2002), it has been shown that the knowledge of foreign languages significantly influences the pattern of international trade (see also Méltz, 2008, and Egger and Lassmann, 2015, among others).

are not fully offset by other institutions, such as the school.² Therefore, language-related earning inequality (as well as other distortions in other social outcomes) are likely to persist across generations.

Thus, the focus of the existing literature has been on the behavior of immigrants in societies with a predominant language. The main concern there is the speed of assimilation of minority groups. However, the economic value of language skills and the issue of language transmission across generations are not only relevant for migrants, but also for general populations living in multilingual environments. Many countries and regions are bilingual, in the sense that there exist two speech communities of significantly large size and a relevant fraction of native-born individuals with native competences in both languages.³ Here, the primary issues are different: the economic value of bilingualism, the endogamy of different speech communities, and the long-run contest between competing languages for dominance.

In this paper, we analyze for the first time the intergenerational transmission of language in bilingual societies. Our results indicate that such a channel reinforces the impact of policies that promote bilingualism and contributes to the large and persistent effect of language skills on labor market performance, as well as on other social outcomes. In addition, we argue that the implementation of language-in-education policies and their intergenerational spillovers may actually improve the vitality and, eventually, the chances of survival of minority languages.

We specifically focus on the effect of the parents' language skills on the language(s) spoken with the children, thus examining one of the potential channels that can drive the positive correlation between the level of proficiency of parents and children found in the existing literature (Bleakley and Chin, 2008; Casey and Dustmann, 2008 and Kuziemko, 2014). The language spoken by parents at home may influence the children's linguistic repertoire (provided the school and other institutions do not fully compensate the family's influence), as well as linguistic preferences and identity formation. These factors may influence off-springs' productivity and network formation, or may even shelter them from discrimination, with long-term effects on labor market outcomes.

Our empirical analysis focuses on the case of Catalonia, one of the bilingual regions of Spain. Catalonia is a particularly interesting case study. Firstly, communication is not at stake in this bilingual society due to the universal knowledge of Spanish (a very strong global language, with more than 400 million native speakers). Thus, the consequences of acquiring additional skills in Catalan, the local language, cannot be attributed to its communicative benefits. However, the degree of bilingualism is asymmetric, since only a fraction of native Spanish speakers is proficient in Catalan, whereas all native Catalan speakers are fully bilingual (high proficiency in both Catalan and Spanish). Previous research showed that knowledge of Catalan increases employability (Rendon, 2007), earnings (Di Paolo and Raymond, 2012; Cappellari and Di Paolo, 2018) and the formation of linguistically-mixed couples (Caminal and Di Paolo, 2019).

Secondly, Catalonia experienced a language-in-education reform that introduced bilingualism at school, which can be exploited to identify the causal effects of parents' language skills on language transmission. The language-in-education reform we study in this paper (Language Normalization Act, LNA) was implemented in 1983 by regional authorities and introduced Catalan as a medium of instruction (together with Spanish) in Catalan schools. This significantly enhanced Catalan proficiency among native Spanish speakers, without altering their proficiency in Spanish. The sociolinguistic characteristics of the Catalan

population in the early 1980s were crucially affected by the massive immigration flows from the rest of Spain that took place in the 1960s and early 1970s. Since Catalan was strongly repressed during the dictatorial regime (1939–1977), a significant fraction of the Spanish-speaking population at that time had limited or no knowledge of Catalan. The main purposes of the LNA reform were to foster bilingualism and biliteracy in the society (regardless of language background) and to “normalize” the use of the local language as to preserve its vitality.

Previous papers studied the effects of the LNA reform on national identity (Clots-Figueras and Masella, 2013), the wage returns to education (Cappellari and Di Paolo, 2018) and partnership formation (Caminal and Di Paolo, 2019). This paper adds to this literature by examining the intergenerational spillovers generated by the new language skills induced by the reform.

We begin our analysis by presenting a simple choice-theoretic framework of the intergenerational transmission of language in a bilingual society that portrays the optimal choice of the language(s) parents use with their children as balancing their own linguistic preferences with the welfare of their children (as measured by their labor market outcomes). The goal of the framework is twofold. First, to understand the main channels through which a language-in-education policy, like the LNA reform, may influence the language parents speak to their children. Second, it helps interpreting some of the stylized facts and empirical results. While emphasizing the effect of the reform on language skills, the model also points to two additional channels: (i) the potential effect of the reform on linguistic preferences, and (ii) the positive effect of the reform on the formation of mixed-couples, and its implications on the language used with the children. We show that, under certain conditions, the pattern of intergenerational transmission of language in an asymmetric bilingual society is indeed affected by language skills and, moreover, such a positive influence only concerns (somewhat surprisingly) the weak (local) language.

We next turn to the empirical analysis by proposing an identification strategy that takes advantage of the LNA reform as a natural experiment. We exploit the heterogeneous effect of exposure to bilingualism during compulsory schooling by native language to identify the causal effects of oral skills in Catalan among native Spanish speakers on the language spoken with their children. Since Catalan was orally transmitted within Catalan-speaking families (even during Franco's era), the reform boosted oral Catalan skills only among native Spanish speakers. We leverage on year of birth and native language to construct an instrumental variable (IV) for the Catalan skills of native Spanish speakers. The validity of the instrument relies on the assumption that non-language cohort effects in language transmission are common between Catalan and Spanish speakers, and we provide corroborating evidence based on several robustness checks.

We show that parents' language skills matter in the intergenerational transmission of language. The reform had a significant impact on the Catalan proficiency of native Spanish speakers, increasing it by on average around one point on a 10-points scale (15% increase) for those who were exposed to the reform for their entire compulsory education period (10 years), relative to the benchmark of no exposure. Using the proposed IV, we find that the increase in oral proficiency induced by the reform among native Spanish speakers raises their propensity to use Catalan with their children by two fifths of a point of the 1–5 scale with which language use is categorized in the data. Such an effect implies an 18% reduction in the average gap in Catalan use between native Spanish and native Catalan speakers. Additional estimates obtained from a non-linear model for ordered outcomes with an endogenous covariate indicate that increasing oral skills in Catalan by one unit raises the probability that a native Spanish speaker uses only Catalan with the first child by 13 percentage points. We also estimate the reduced-form effects of the reform, which provide evidence about the direct effect of exposure to bilingualism during compulsory education on language transmission. The reduced-form model predicts that the probability of using only Catalan for Spanish speakers with no exposure to the reform

² More recently, Kuziemko (2014) demonstrated the existence of reversed intergenerational transmission, with the proficiency of immigrant parents being negatively affected by their children's proficiency. That is, parents lean on, rather than learn from, their children.

³ Just in Europe, examples include some of the former soviet republics (like Ukraine, Estonia, and Kazakhstan), Belgium, Cyprus, Wales, Basque Country, and Catalonia.

is 22%, whereas for those fully exposed to the reform such a probability jumps to 40%.

Previous studies have documented the positive effect of additional skills in the local language on labor market and other social outcomes. Policies that promote the knowledge of minority languages not only tend to reduce wage inequality, increase efficiency and reduce endogamy among the affected generation. We show that those policies may also generate intergenerational spillovers, extending the above benefits to future generations and reinforcing the short-run responses. Clearly, such long-run effects may significantly contribute to the vitality and preservation of local languages.

The remainder of the paper proceeds as follows: In [Section 2](#) we provide a brief historical account of the sociolinguistic dynamics of Catalonia. [Section 3](#) presents the outline of the theoretical framework, which is formally presented in the Appendix. [Section 4](#) contains a description of the data used in the empirical analysis. The empirical framework is detailed in [Section 5](#). [Section 6](#) discusses the main results and [Section 7](#) examines the evidence from the robustness checks. [Section 8](#) concludes.

2. Historical background

Catalan and Spanish (Castilian) are both Romance languages. Catalan is the variety of Vulgar Latin that developed in the territory from the 9th century. Spanish arrived in Catalonia as early as the 15th century and consolidated its position among the elite in the 18th century. In the 19th and early 20th centuries, the expansion of elementary education increased the knowledge of Spanish among the general population.

Franco's regime (1939–1977) represented a serious setback for Catalan. It imposed Spanish as the only official language and the only language to be used in education. Catalan was excluded from most social activities but, nevertheless, it was still intensively used in the private sphere and, more importantly, transmitted across generations. The massive migration flows from the south of Spain to Catalonia in the 1960s and early 1970s dramatically changed the demographics of the region. In the mid-1970s, 40% of the Catalan population was born outside Catalonia. Urban segregation of immigrants and the ban on the use of Catalan in schools and media meant that, for the first time in history, a very large fraction of the Catalan population was monolingual in Spanish ([Woolard and Gahng, 1990](#); [Siguan, 1991](#)).

The political environment drastically changed with the advent of democracy. In particular, new language-in-education policies were introduced after the approval of the Language Normalization Act (LNA) in 1983. The primary goal of the LNA was for all students to be fully bilingual in both Catalan and Spanish by the end of their compulsory education. It also established an education system in which students were not separated on the basis of their native language. With the implementation of this reform, education slowly evolved from a system in which Catalan was excluded to one in which Catalan was the main language of instruction in compulsory education.⁴ The reform led to a significant improvement of the Catalan skills of native Spanish speakers, whereas the Spanish skills of both native Spanish and Catalan speakers remained essentially unchanged (at excellent levels). The reform also improved written skills among native Catalan speakers, while their oral skills were already high for cohorts unaffected by the policy change, because Catalan was orally transmitted within Catalan-speaking families (even during the most repressive phases of the dictatorship). In this paper we focus on oral skills, which represents the most important dimension for intergenerational language transmission.

Nowadays, the position of the two languages in Catalan society is still asymmetric, but the situation is clearly more complex. Spanish is the dominant language in Spain, the official language throughout all its territory, and backed by the central government. Moreover, it is one

of the most important global languages with almost 400 million native speakers and probably more than 200 million second-language speakers. Clearly, the private and social incentives to learn Spanish are huge. Catalan enjoys strong support from the regional government. Besides Catalonia, it is also spoken in other territories of the ancient Kingdom of Aragón, mainly Valencia and the Balearic Islands. Overall, less than six million people speak Catalan on a daily basis. Consumption of media and cultural goods in Catalonia is largely in Spanish, with the exception of radio and books where the share of Catalan fluctuates around 50% and 30%, respectively. Local and regional politics are conducted mostly in Catalan, but in other branches of government (such as judicature) Catalan is almost non-existent.

Three facts deserve special consideration. First, all Catalan citizens are fully proficient in Spanish, but not all are proficient in Catalan. This asymmetry is not likely to die out any time soon. Even though basic skills in Catalan are widespread among the younger generations, the distribution of Catalan skills still exhibits high variance. Thus, in terms of proficiency, Spanish is still the dominant language, but less so than a few decades ago. Second, the labor market seems to consistently reward bilingual individuals, as shown by [Rendon \(2007\)](#), [Di Paolo and Raymond \(2012\)](#), and [Cappellari and Di Paolo \(2018\)](#). Third, bilingualism also provides non-monetary private and social rewards. In particular, increased fluency in Catalan among native Spanish speakers has increased their chances of finding a partner from the other language community ([Caminal and Di Paolo, 2019](#)).

3. Theoretical framework

Before moving to the empirical strategy, it is useful to discuss some possible mechanisms that affect the choice of intergenerational language transmission from a theoretical point of view. We provide a formal model in the online Appendix (A), and in this section we sketch its main features and conclusions. The model offers some guidance for understanding the channels through which a language-in-education reform may affect the patterns of language transmissions. The aim is to provide a theory of intergenerational language transmission that applies to asymmetric bilingual societies.

3.1. Benchmark framework

Consider a society with two languages. The status of these two languages is quite different. One is a strong global language, dominant in cultural goods and media products. Hence, the incentives to learn it are huge. In contrast, the incentives to learn the other language are only linked to its local use, and hence are much more limited.

When parents choose the intensity with which they speak each language with their children, they pursue two separate goals: i) maximizing the wellbeing (labor market and social outcomes) of their children and ii) transmitting their ethnic or cultural heritage (i.e., parents have a preference for their *identity* language). In a society with a clear dominant language (say the US), the optimal family linguistic background from the point of view of children's wellbeing is clearly the dominant language (English). Thus, only immigrant parents with very strong preferences for their identity language and/or very poor skills in the dominant language will transmit their identity language. In an asymmetric bilingual society, the optimal choice from the point of view of the children's wellbeing is less clear-cut. In the remainder of the section we consider Catalonia as the example of asymmetric bilingual society, where Spanish is the strong global language and Catalan the local one.

In the case of Catalonia, the existing evidence indicates that bilingualism pays off in both the labor market and in other social contexts. In addition, native Catalan speakers are bound to become bilingual, whereas this is less likely for native Spanish speakers. This suggests that parents' optimal choice from the point of view of the children's wellbeing is to use a combination of both languages. The actual way

⁴ See [Clots-Figueras and Masella \(2013\)](#), or [Cappellari and Di Paolo \(2018\)](#), for more details about the LNA reform.

in which the two languages will be used will depend on parents' preferences and skills. Native Catalan speakers will transmit Catalan more intensively than native Spanish speakers. In any case, provided they are relatively fluent and their preferences for their identity language are not too strong, native Spanish speakers will typically choose to transmit Catalan together with Spanish.

Given this framework, the main prediction of the model is that a language-in-education reform that reduces the cost of learning Catalan will induce native Spanish speakers to acquire better skills in Catalan, which in turn will reduce the cost of switching to their second language. Consequently, they will choose to speak Catalan with their children more intensively. Hence, the policy change will not only affect the current generation but will also alter the distribution of native languages of the next generation, contributing to the vitality and preservation of the weak language in the long-run.

According to our framework, the reason why native Spanish speakers end up transmitting (at least, partially) Catalan is that bilingualism is rewarded in the labor market and in society at large. Thus, our framework cannot adequately represent societies with an undisputed dominant language (for instance, English in the US). In those cases, a bilingual immigrant worker, who has learnt the dominant language in addition to her native language, could be worse off than a monolingual, native speaker of the dominant language as far as their labor market position is concerned (Chiswick and Miller, 2018). The potential premium associated to the knowledge of the immigrant's native language is very small, and probably dominated by the possibility of being discriminated against on the basis of a peculiar accent (Grogger, 2019). Thus, exclusively transmitting the dominant language to the next generation (language shift) is often an optimal strategy. In a more balanced bilingual society, discrimination on the basis of accents is much less likely and in contrast the economic value of bilingualism will tend to be much higher.

3.2. Alternative channels

While skills are perhaps the most intuitive channel for the impact of the language-in-education reform on language transmission, our model allows for the possibility that the policy triggers alternative mechanisms.

a) Preference switching

In the baseline framework linguistic preferences are exclusively determined by the individual's native language, but it may well be that they reflect other contextual influences, such as those coming from the school environment. If the language-in-education reform affects the linguistic preferences of native Spanish speakers or, more drastically, replaces Spanish with Catalan as the preferred language of some of them, then they will use Catalan more intensively with their children. Thus, the reform may affect both skills and preferences. These two separate channels may cause similar effects on the pattern of language transmission, but they tell quite different stories. The change in linguistic preferences can be interpreted as an act of cultural assimilation, whereas the change in skills is compatible with a stable composition of cultural communities as defined by their identity.

b) Mixed couples

In the benchmark framework parents have the same native language, and thus share the same goals. As a result, they agree on the language or combination of languages that they use with their children). Since existing evidence shows that the reform has increased the formation of linguistically mixed couples (Caminal and Di Paolo, 2019), it is interesting to consider the impact of native language heterogeneity among parents on intergenerational transmission.

The members of a mixed couple must deal with the potential conflict of interest arising from their different linguistic preferences. As a result, they might end up using different languages. Whereas the effect on the children's linguistic background still depends on the average

language use of the parents, each member of the couple may now be differently affected by their own language use and that of the partner's. In other words, a native Spanish speaker incurs a cost when using Catalan with the children but is also negatively affected when the partner uses Catalan, although the latter effect is likely to be smaller. The existence of such spillover across parents is very natural since some family activities involve both parents. In fact, the size of the spillover effect turns out to be crucial in the analysis.

Indeed, the main question concerns the consequence of having a native Catalan speaker as a partner on the intergenerational language choices of a native Spanish speaker. In one extreme, if there is no spillover (that is, if each member of the couple is not directly affected by the language the partner uses with their children), then the native Spanish speaker uses Catalan less intensively. The reason is that the all-out use of Catalan by her partner is sufficient to guarantee that the children will reach the optimal linguistic background, and hence she finds it optimal to use her preferred language (substitution effect). In the other extreme, if the spillover is so large that each member places the same weight on its own language use than on the partner's, then the native Spanish speaker will use Catalan more intensively, due to the averaging effect. That is, both members share exactly the same goals, and hence they choose the same language policy, which is an average of the optimal choices that each individual parent would make in a homogeneous couple. Thus, the effect of being in a mixed couple for native Spanish speakers will depend on which effect dominates, averaging or substitution. If the first effect dominates and the policy fosters the formation of mixed couples, then the reform would enhance the intergenerational transmission of Catalan.

4. Data description

The data are drawn from the Survey of Language Use of the Catalan Population, which is produced by the Catalan Statistical Institute (IDESCAT), is representative of the Catalan population, and is available for survey years 2008 and 2013.

The survey contains unique information on sociolinguistic characteristics. It includes specific questions about native language, defined as the first language spoken at home during childhood, self-identification language at the time of the survey, parental native language, as well as current and former partner's native language. Moreover, there are separate questions about proficiency in both Catalan and Spanish and, most importantly, about the language normally spoken with the first three children. Finally, the survey also provides detailed sociodemographic characteristics (gender, year of birth, place of birth, place of residence, education, etc.) and parental background (father's and mother's place of birth and the level of education attained). All variables are self-reported by the respondent.

We focus on individuals who were entirely educated in Catalonia and exclude from the estimation sample those born in other Spanish regions who migrated to Catalonia at age 6 or older. Foreigners are also excluded from the estimation sample. The first birth-cohort that experienced some exposure to bilingualism in compulsory education contains those born in 1970, who attended the last grade of compulsory school in 1983. We build our sample symmetrically around this pivotal cohort and include in the analysis individuals born from 1950 up to 1990, implying that the youngest respondent is aged 18 and the oldest 63 at the date of interview.⁵

Virtually all selected respondents report that either Spanish or Catalan is their native language or the language they speak with children, and we deleted the few (120) observations for which this is not the case. Roughly 60% of respondents who meet the previous criteria have at least one child and can contribute to the analysis.⁶ After applying

⁵ We use data for older cohorts to conduct placebo exercises, see Section 7.

⁶ All the children of the respondents in our sample are exposed to bilingualism at school.

Table 1
Descriptive Statistics by Native Language.

Variable	pooled sample		native Catalan speakers		native Spanish speakers	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
language spoken with first child: only Catalan	0.610	0.488	0.912	0.283	0.246	0.431
language spoken with first child: more Catalan than Spanish	0.047	0.213	0.028	0.166	0.070	0.256
language spoken with first child: equal Catalan and Spanish	0.088	0.283	0.036	0.187	0.149	0.356
language spoken with first child: more Spanish than Catalan	0.070	0.255	0.010	0.099	0.143	0.350
language spoken with first child: only Spanish	0.185	0.388	0.013	0.113	0.392	0.488
age of the first child	16.75	10.41	17.45	10.50	15.89	10.24
oral skills in Catalan	8.714	2.154	9.578	0.911	7.672	2.692
oral skills in Spanish	9.558	0.943	9.415	1.094	9.729	0.683
native Spanish speaker	0.454	0.498	0.000		1.000	
Catalan (or Catalan and Spanish) as self-identification language	0.696	0.460	0.982	0.131	0.349	0.477
partner Catalan (or Catalan and Spanish) native speaker	0.580	0.494	0.769	0.422	0.351	0.477
both parents Spanish speakers	0.408	0.491	0.033	0.178	0.859	0.348
both parents born outside Catalonia	0.356	0.479	0.043	0.203	0.732	0.443
years of compulsory language exposure	1.660	2.852	1.446	2.702	1.918	3.003
wave 2013	0.554	0.497	0.546	0.498	0.563	0.496
Age	45.56	8.495	46.59	8.528	44.32	8.290
Male	0.456	0.498	0.470	0.499	0.439	0.496
number of observations	3667		2004		1663	

all the data selections described above and eliminating some observations with missing values in the relevant variables, we obtain a sample of 3667 observations (out of about 14,000 cases available in the two waves of the survey).⁷

Table 1 reports descriptive statistics for the main variables used in the empirical analysis (descriptive statistics of all the explanatory variables used in the empirical analysis are displayed in Table 1B of the online Appendix B). Native language (of the respondents, their parents and their partners) is originally classified into three categories: 1) only Catalan, 2) Catalan and Spanish, and 3) only Spanish. There are very few cases of mixed native language (4.2%). We consider individuals to be native Spanish speakers if they report their native language to be only Spanish (results are robust to the use of alternative definitions). According to this definition, 45% of individuals are native Spanish speakers. Although the survey reports separate information on several dimensions of proficiency (understanding, speaking, writing, and reading), here we focus on oral proficiency in Catalan, since oral fluency is the most relevant domain of language proficiency for language transmission from parents to children. Self-reported speaking skills in Catalan are coded on a 0–10 scale, with higher levels referring to more advanced language skills.

Native Catalan speakers are virtually fully proficient in their language (77% report a value of 10), whereas the average oral skills in Catalan are 1.9 points lower for their native Spanish speaker counterparts (the difference is statistically significant). Both language communities are virtually fully proficient in oral Spanish (9.4 for Catalan speakers and 9.7 for Spanish speakers), reflecting the asymmetric bilingualism of the Catalan society. Fig. 1, shows that the level of oral proficiency in Catalan among the Catalan native speakers remained stable over birth cohorts. The same variable displays a substantial increase for native Spanish speakers among cohorts of individuals born since the late 1960s, the ones experiencing the change in the language of instruction.⁸

The survey provides separate information on the language spoken with each of the first three children. However, we only consider the first

child, both to preserve sample size (respondents with at least two children are 64% of the selected sample) and because there is little between-children variation in language spoken with children by birth order. The language spoken with children is recorded into five categories: only Spanish (1), more Spanish than Catalan (2), equal Spanish and Catalan (3), more Catalan than Spanish (4), and only Catalan (5). The large majority of native Catalan speakers transmit their native language to the firstborn child. Conversely, a substantial amount of variation is observed among native Spanish speakers. Only 39% of them speak only Spanish to their firstborn child, whereas 25% speak only Catalan, and the remainder mixes both languages.

5. Empirical framework

As mentioned in the previous sections, younger cohorts of native Spanish speakers are more fluent in Catalan due to bilingual exposure at school. We now set out an empirical framework to identify the causal effect of the increase in oral language skills among Spanish-speaking parents on the choice of the language used with the next generation. We start with a simple model, estimated for the subsample of native Spanish speaker ($S_i = 1$ if individual i reports Spanish as the native language), in which we treat language skills as exogenous. The outcome is the language spoken with the firstborn child (an ordinal variable L_i taking values from 1 to 5 on an increasing scale of intensity of use of Catalan), and the main regressor is the self-assessed measure of oral proficiency in Catalan (C_i), which takes values from 0 to 10 for increasing levels of proficiency.

The equation of interest is as follows:

$$L_i = \alpha + \beta' X_i + \delta C_i + \theta_{\tau(i)} + \epsilon_i \text{ if } S_i = 1 \quad (1)$$

where $\tau(i)$ is person i 's year of birth and θ is a year of birth fixed effect to control for trends across the cohorts, while the vector X_i contains a parsimonious set of exogenous covariates (age, gender and a wave dummy).⁹ In this model, the parameter δ represents the effect of the Catalan linguistic skills of native Spanish speakers on the intensity with which they use Catalan with their children. This is the key parameter

⁷ It seems worth mentioning that almost 50% of reside in the Metropolitan Area of Barcelona, where the relative use of Catalan (versus Spanish) is less intense than in other areas of Catalonia. Reassuringly, the main results reported below are virtually the same when estimating the model for separate subsample for individuals residing in the Metropolitan Area of Barcelona or elsewhere in Catalonia (available upon request).

⁸ Individuals born between 1966 and 1969 were partially exposed to the change in the language of instruction if they remained at school after compulsory education.

⁹ In robustness checks we extend the set of controls in a stepwise fashion, including respondents' parental (place of birth and education) and personal (place of birth, place of residence and education) characteristics. Moreover, we also considered interacting the vector of controls with the native language dummy to allow for heterogeneous effects. As detailed in Section 6, extending the set of control variables or interacting them with native language does not affect our results.

Fig. 1. Oral skills in Catalan by year of birth and native language.
Note: bins are defined according to the year of birth; local linear smooth plots included.

measuring the impact of skills on the intergenerational transmission of language, which is expected to be greater than zero ($\delta > 0$) as long as better skills reduce the costs of using Catalan for native Spanish speakers.¹⁰

Our theoretical framework clearly highlights the direction of causality from parental skills to intergenerational transmission. However, it is hard to claim empirically that measured language skills (C_i) are exogenous with respect to the unobservables that may affect the choice of the language spoken (ϵ_i). For instance, native Spanish speakers with a taste for Catalan culture will invest in their own linguistic skills and will also be more likely to use Catalan with their children. As long as preferences are unobservable, the “naïve” estimation of (1) will deliver biased estimates of the parameter of interest.¹¹

We tackle the endogeneity issue with an empirical strategy based on the exogenous variation in oral skills in Catalan plausibly induced by the LNA reform of 1983. We employ an exogenous measure of potential exposure to Spanish-Catalan bilingualism at school during compulsory education, which only depends on birth cohort and not on post-compulsory schooling choices. One option is to separate cohorts on the basis of the year of birth, individuals born after 1970 representing the treated cohorts. However, not all treated cohorts received the same amount of bilingual education during their compulsory schooling, and a binary pre-post indicator would miss this key aspect of the analysis. One reason for differential exposure is the sharp implementation of the reform across all educational levels in 1983, implying that for cohorts already attending school in that academic year the actual amount of exposure depends on the remaining number of years before completion of compulsory education. Secondly, a subsequent reform that took place in 1990 extended

compulsory education from 8 to 10 years, translating into longer exposure to bilingual schooling for cohorts born in 1983 and onwards. Therefore, to capture differential treatment intensity within the group of treated cohorts, the exposure measure is defined as a piecewise function of the birth year:

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

This is the variable that Clots-Figueras and Masella (2013) used to measure the degree of exposure to the educational reform during compulsory education, and that they considered as an exogenous determinant of political identity in Catalonia.

Although the language exposure measure $E_{\tau(i)}$ is clearly exogenous, as it depends exclusively on year of birth ($\tau(i)$) and on the duration of compulsory schooling, it might not be a good instrument for C_i as it could be related to the error term of Eq. (1), violating the exclusion restriction assumption. This is because $E_{\tau(i)}$ may capture before-after changes across the cohorts that might affect the language spoken with the children by native Spanish speakers through other channels than proficiency in Catalan. For example, the increased societal prestige of Catalan might have increased the propensity to use this language with the offspring regardless of language skills.

In order to construct a valid instrument, we exploit the heterogeneous effect of language exposure at school by native language, motivated by the fact that the language-in-education reform did not exert any significant effect on the oral Catalan skills of native Catalan speakers, as shown in Section 4, while their Spanish skills remained very high and stable over cohorts. To allow for such heterogeneous effects, we begin by including Catalan native speakers in the analysis, which implies that Eq. (1) becomes:

$$L_i = \alpha + \beta' X_i + \delta C_i + \pi S_i + \theta_{\tau(i)} + \epsilon_i, \quad (3)$$

¹⁰ We estimate equation (1) both by OLS and Ordered Probit.

¹¹ Another bias in the estimation of equation (1) may arise from reverse causality, as long as parents learn from their children who are exposed to non-native language at school, see Kuziemko (2014). Furthermore, self-reported language skills may be subject to measurement error, a possibility that we consider in the robustness checks.

which is estimated on the full sample and where we have added a control for having only Spanish as native language (S_i).¹² Since there is virtually no variation in Catalan skills among native Catalan speakers, the associated parameter (δ) estimated from Eq. (3) is essentially equivalent to its counterpart in Eq. (1), and also equivalent to one that would be estimated by entering the $S_i C_i$ interaction in Eq. (3), replacing C_i , as in each case the source of variation used in estimation is the same.

As in Caminal and Di Paolo (2019), we construct our instrument as the interaction between the native language indicator S_i and compulsory language exposure $E_{\tau(i)}$. The corresponding first stage equation for instrumenting skills in Catalan is:

$$C_i = \mu + \varphi' X_i + \omega S_i + \rho(E_{\tau(i)} S_i) + \sigma_{\tau(i)} + u_i \quad (4)$$

where $\sigma_{\tau(i)}$ is a year of birth fixed effect. Indeed, the inclusion of native Catalan speakers enables controlling for general trends across the cohorts in a flexible way (i.e. dummies for year of birth). This implies that the instrument $E_{\tau(i)} S_i$ captures the change in language proficiency induced among native Spanish speakers who were increasingly exposed to Catalan at school due to the LNA reform, after partialling out non-language trends across the cohorts through the inclusion of (common) year of birth fixed effects. The underlying assumption of our identification strategy is that both language communities were subject to the same general cohort effects in intergenerational language use. Therefore, specific cohort effects experienced by native Spanish speakers after the policy change should be attributed to their improved skills in Catalan.

This approach is in the spirit of the identification strategy proposed by Bleakley and Chin (2004, 2008, and 2010). They estimate the assimilation effects of English proficiency among US immigrants, exploiting the existence of a “critical period” of language acquisition (i.e., immigrants who arrive in the host country at a very young age assimilate the language more easily). Their identifying variable is the interaction between age at arrival and a dummy that takes the value one if the immigrant comes from a non-English-speaking country. Under the assumption that the non-language assimilation effects of early migration are the same for immigrants arriving from English-speaking countries as for those from non-English-speaking countries, the differential effect of age at arrival for those who migrated from a non-English-speaking country should be purged of non-language-related effects and thus would represent a valid exclusion restriction. In our case, the IV estimator for the effect of skills on intergenerational language use can be obtained from the following equation

$$L_i = \alpha + \beta' X_i + \pi S_i + \delta_{IV} \hat{C}_i + \theta_{\tau(i)} + \varepsilon_i, \quad (5)$$

which includes the prediction from the first stage equation ($\hat{C}_i$) rather than the observed variable. The consistency of δ_{IV} is based on the assumption that non-language cohort effects that could be captured by the exposure variable are common between Spanish and Catalan speakers, and are thus partialled out by the (common) year of birth fixed effects included in the main equation ($\theta_{\tau(i)}$). Under this assumption, the IV estimator identifies the causal effect of oral proficiency in Catalan on language transmission among native Spanish speakers affected by the LNA reform (a Local Average Treatment Effect, LATE).¹³

Analysing the direct effect of the reform on the outcome in a reduced form equation is of independent interest, since it enables quantifying the long-run effects of the reform on the intergenerational language choices

of Spanish speakers. The reduced form equation

$$L_i = \alpha + \beta' X_i + \pi S_i + \gamma_{RF} E_{\tau(i)} S_i + \theta_{\tau(i)} + \varepsilon_i \quad (6)$$

provides evidence regarding the effect of the policy on the preservation of a local language and its vitality. Moreover, it is also useful to check if the results are driven by pre-existing trends that could differently affect the choice of the language used with the children depending on the parents' native language. To this aim, we perform several falsification exercises based on placebo reduced form regressions like Eq. (6), in which we include placebo exposure variables obtained by artificially shifting back in time the roll-out of the reform (see Section 7).

We also consider several possible violations of the exclusion restriction or other issues that might undermine the internal and external validity of our results. First, the variable capturing native language is self-reported and the reform (or a contextual increase in the societal prestige of Catalan) might have induced some individuals to miss-report their native language. Therefore, we check that the results are robust to alternative proxies for native language, such as parents' language and parents' origins. Second, the availability of these two alternative proxies for language background enables specifying over-identified models and to test the validity of the instruments, an exercise that seems useful as long as each of the proposed instruments affects subpopulations that are to some extent largely overlapping. Specifically, we present two alternative overidentified models in which we use as instruments the interactions with both native language (as in our baseline specification) and any of the two proxies (parents' language and origins) and test for overidentification.¹⁴ Third, as discussed in Section 3, the LNA reform might have changed linguistic preferences of native Spanish speakers independently of their language skills, alleviating the trade-off between socioeconomic incentives and linguistic preferences. In order to provide suggestive evidence that changes in preferences are not the main driver of the results, we exploit the information about self-identification language of the respondents and exclude from the estimation those who have a self-identification language that is different from their mother tongue (i.e. the “language switchers”). Finally, the theoretical model also indicates that changing patterns of partnership formation might play a role, since the increase in language proficiency of native Spanish speakers induced by the reform also affected their propensity to find a Catalan-speaking partner. While strictly speaking this is not a violation of the exclusion restriction, it anyway represents a different mechanism that is not considered in our empirical strategy. In order to rule out this alternative mechanism as the main driver of the results, we repeat our estimations after removing mixed couples from the sample.

6. Results

In Table 2 we report the OLS estimates of the impact of linguistic skills on language transmission choices for native Spanish speakers, corresponding to parameter δ in Eq. (1).¹⁵ The positive parameter estimate indicates that linguistic skills in Catalan are associated with more intense use of Catalan with children. In particular, in column (1) a unit change in the linguistic skill variable (C_i , measured on a 0–10 scale) is associated with a 0.281 increase in the intensity of Catalan use with children (L_i , measured on a 1–5 scale). The estimated effect is highly

¹² This implies that the indicator for being native Spanish speaker is treated as exogenous in equation (3). We recognize that this could be controversial, due to the self-reported nature of the variable. However, in the robustness checks section we provide evidence that our results are virtually invariant using alternative proxies for native language.

¹³ It is also possible to estimate the same model with endogenous language skills adapting the Ordered Probit model to the presence of an endogenous regressor. We report this additional evidence to provide the quantitative interpretation of the causal estimates.

¹⁴ We also test for overidentification using dummies for exposure (interacted by native language) as exclusion restriction. Overidentification test are also useful to dispel doubts about exposure to the LNA correlating with language use besides its effect on language skills, for example via changes in political, cultural or linguistic identity, or because it could affect other socioeconomic aspects of life. Our empirical framework does not assume the absence of such potential effects, but only restricts them to be the same for the two language groups. In this context, testing for instruments validity helps assessing the heterogeneity of the alternative channels by language group.

¹⁵ Complete sets of coefficient estimates are reported in the Appendix (B).

Table 2

OLS regressions (outcome = language spoken with the first child), only Spanish speakers.

	(1)	(2)	(3)	(4)
oral skills in Catalan	0.281*** (0.011)	0.278*** (0.011)	0.244*** (0.013)	0.242*** (0.013)
parents' controls	NO	YES	NO	YES
individual controls	NO	NO	YES	YES
adjusted R-squared	0.254	0.260	0.293	0.295
number of observations	1663	1663	1663	1663

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for individual's and child's age and individual's year of birth dummies. Regression in column (2) contains controls for paternal and maternal place of birth (with missing indicators) and highest parental education (with missing indicators). Regression in column (3) includes controls for individual's place of birth, place of residence and completed education. Complete results are reported in Tables B2 of Appendix B.

significant. It indicates that one standard deviation increase in oral proficiency in Catalan is associated with an increase of 0.46 standard deviation points in the intensity of the use of Catalan with the first child. Alternatively, considering that the average of the language use variable is 4.8 for native Catalans and 2.6 for native Spanish speakers, the result indicates that a unit increase in linguistic skills reduces the mean difference in Catalan use in the two groups by 13%. Adding as controls respondents' characteristics and those of their parents in columns (2) to (4) does not affect the substance of the evidence.

As discussed in Section 5, we tackle possible biases with an IV strategy that exploits the differential change in linguistic skills between native Spanish and Catalan speakers induced by the 1983 reform. We begin by including also the information on native Catalan speakers in the analysis and estimate Eq. (2) by OLS using the full sample. Results are presented in Table 3 and show that the point estimate of the impact of linguistic skills is virtually the same as the one estimated in the sample with only native Spanish speakers. This is because there is essentially no variation in linguistic skills among native Catalans. Moreover, the coefficient of interest is barely affected by the inclusion of parental and individual controls, even when interacted with the native language indicator (column (5)).

Using the full sample, we can now move to the IV estimates, corresponding to Eq. (5). The upper panel in Table 4 reports the first stage estimates of the impact of bilingual compulsory education on skills in Catalan. The positive estimate of ρ indicates that for native Spanish speakers 1 year of exposure to bilingual education increases reported oral proficiency by 0.114 points on the 0–10 scale. The predicted proficiency score among Spanish native speakers educated before the reform is 7.45, implying that someone starting school after the reform (and therefore experiencing 10 years of exposure to bilingual school) experiences a proportional increase in proficiency of approximately 15% relative to the benchmark of no exposure ($= 10 \times 0.114/7.45$). The effect is relevant and also statistically significant; in particular, the F-statistic for the instrument shows that it is a rather strong predictor of proficiency. Reading across columns, we note that the first stage estimate is rather stable after introducing additional controls, which speaks in favor of the exogeneity of the proposed IV.

The lower panel of Table 4 reports the IV estimates for the causal effect of linguistic skills on intergenerational language use. Under the assumption that cohort effects in language transmission, other than the ones operating through skills, are common across speech communities, the IV estimate captures the causal effect of the increase in oral skills in Catalan of native Spanish speakers (induced by the LNA reform) on the language spoken to the first child. The estimated effect is larger than in the OLS counterpart: for example, in column (1) a unit increase in proficiency increases the intensity of intergenerational Catalan use for native Spanish speakers by 0.4 points on the 1–5 scale, versus the OLS coefficient equal to 0.26. In the likely case of absence of defiers to the reform, the IV estimate can be interpreted as a LATE, the causal effect of skills for the sub-group of compliers, native Spanish speakers whose proficiency increased because of the reform. The fact that the IV estimate is larger than the OLS counterpart suggests that the intergenerational transmission decision may be more reactive to increases in linguistic skills for compliers compared with the general population.

The finding of a higher impact of language proficiency on language transmission estimated by IV might be also explained by reverse causality issues, or by the presence of measurement error in the language skills variable, which tends to bias OLS downward (under classical measurement error). With the aim of understanding the role played by measurement error, in the Appendix Table B6 we show additional estimates obtained using information about the language of the interview as an independent proxy for oral skills in Catalan. Since the survey is administered by the Catalan Government, the interviewers are requested to

Table 3

OLS regressions (outcome = language spoken with the first child), pooled sample of Spanish and Catalan speakers.

	(1)	(2)	(3)	(4)	(5)
native Spanish speaker	-1.647*** (0.047)	-1.413*** (0.069)	-1.468*** (0.049)	-1.334*** (0.068)	-0.686 (0.964)
oral skills in Catalan	0.264*** (0.009)	0.258*** (0.009)	0.237*** (0.009)	0.233*** (0.009)	0.224*** (0.011)
parents' controls	NO	YES	NO	YES	YES
individual controls	NO	NO	YES	YES	YES
interacted controls	NO	NO	NO	NO	YES
adjusted R-squared	0.567	0.572	0.584	0.587	0.595
number of observations	3667	3667	3667	3667	3667

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for individual's and child's age and individual's year of birth dummies. Regression in column (2) contains controls for paternal and maternal place of birth (with missing indicators) and highest parental education (with missing indicators). Regression in column (3) includes controls for individual's place of birth, place of residence and completed education. Regression in column (5) includes parent's and individual controls interacted with the native Spanish speaker indicator. Complete results are reported in Tables B3 of Appendix B.

Table 4
2SLS regressions, pooled sample of Spanish and Catalan speakers.

	(1)	(2)	(3)	(4)	(5)
FIRST STAGE — outcome: oral skills in Catalan					
native Spanish speaker	-2.102*** (0.104)	-1.735*** (0.117)	-1.613*** (0.103)	-1.473*** (0.116)	-2.555 (1.681)
exposure × native Spanish speaker	0.114*** (0.025)	0.118*** (0.026)	0.109*** (0.024)	0.110*** (0.025)	0.071*** (0.023)
adjusted R-squared	0.222	0.235	0.281	0.284	0.324
F-test of excluded instruments [p-value]	21.78 [0.000]	21.41 [0.000]	20.40 [0.000]	19.70 [0.000]	9.12 [0.005]
2SLS — outcome = language spoken with the first child					
native Spanish speaker	-1.379*** (0.243)	-1.244*** (0.186)	-1.270*** (0.186)	-1.179*** (0.175)	0.617 (1.027)
oral skills in Catalan	0.404*** (0.121)	0.369*** (0.116)	0.376*** (0.119)	0.355*** (0.121)	0.371** (0.182)
parents' controls	NO	YES	NO	YES	YES
individual controls	NO	NO	YES	YES	YES
interacted controls	NO	NO	NO	NO	YES
adjusted R-squared	0.512	0.555	0.559	0.568	0.569
number of observations	3667	3667	3667	3667	3667

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for individual's and child's age and individual's year of birth dummies. Regression in column (2) contains controls for paternal and maternal place of birth (with missing indicators) and highest parental education (with missing indicators). Regression in column (3) includes controls for individual's place of birth, place of residence and completed education. Regression in column (5) includes parent's and individual controls interacted with the native Spanish speaker indicator. The F-test on excluded instruments refers to the Angrist-Pischke multivariate F-test on the interactions between years of exposure to Catalan at compulsory schooling and the indicator for Spanish as native language. Complete results of the first-stage regressions and the outcome equation are reported in Tables B4 and B5 of Appendix B, respectively.

carry out the interview in Catalan, unless the respondent is unable to comfortably answer the questions in this language. If this is the case, they switch to Spanish or other languages. Therefore, the information for whether the entire interview was in Catalan reported by the interviewer is likely to be a powerful predictor of “true” oral proficiency and thus is used as an instrument for self-reported language skills. The results are suggestive that measurement error is actually biasing the OLS estimates toward zero, since the corresponding 2SLS coefficient is much higher, even than our baseline IV results, possibly because the interview language indicator is not a valid instrument to deal with endogeneity and potentially related to the error term of the structural equation.¹⁶

In Table 5 we provide additional evidence from reduced form equations (Eq. (6)) and from Ordered Probit models, also allowing for the presence of an endogenous covariate (i.e. oral skills in Catalan). All the estimates from linear models (column (1)) are roughly the same than the corresponding coefficients for the latent index equation of ordered models (column (2)). This is quite reassuring regarding the use of these non-linear models for the quantitative interpretation of the results, which are expressed through average marginal effects on the predicted probabilities of each possible value of the outcome (columns (3)–(7)).

In the second panel of Table 5 we report the reduced form estimate of the effect of the reform on the intergenerational language choice of native Spanish speakers, controlling for common trends across the cohorts. We focus on the average marginal effects, which indicate that one additional year of exposure increases the probability of using only Catalan

with the child by 1.6 percentage points. The estimated model predicts that the likelihood of using only Catalan for Spanish speakers exposed to Catalan throughout compulsory education (10 years of exposure) is equal to 0.4 versus 0.22 for native Spanish speakers never exposed to bilingualism at school. This result indicates that cohorts of native speakers of the strong language (Spanish) exposed to the policy change have a higher propensity to use the weak language (Catalan) with their children. Therefore, language-in-education reforms that promote bilingualism at school, like the Catalan LNA, generate longstanding effects that foster local language preservation and vitality.

The third panel of the same table contains the coefficient from the structural equation for the outcome, as well as the corresponding marginal effects on the predicted probabilities obtained from the endogenous Ordered Probit model. These estimates are informative about the causal effect of oral skills in Catalan, induced by reform among exposed cohorts of native Spanish speakers, on the language used with the children. According to our theory of language transmission in asymmetric bilingual societies, increased language skills, due to the reduction in learning costs, should be the main channel through which the language-in-education policy should affect the choice of the language used with the offspring. Indeed, the marginal effects show that the predicted probability of using only Catalan increases by 0.13 percentage points for each unitary increase in oral language skills among native Spanish speakers, which represents a sizable effect relative to the baseline probability of using Catalan with the child.

7. Robustness

7.1. Validity of the instrument

We now subject our findings to a number of robustness checks. We begin in Table 6 by addressing the sensitivity of our results to different functional form specifications of the first stage regression. Specifically,

¹⁶ Indeed, as suggested by a referee, it is also possible that our language skills variable defined in a discrete 0–10 interval represents an imperfect measure of language skills, thus leading to misclassification errors rather than measurement errors, which would also introduce bias in the coefficient but in an unpredictable direction. See Dustmann and van Soest (2004) for a more detailed discussion about misclassification errors in self-reported language skills.

Table 5
Ordered Probit and Endogenous Ordered Probit regressions.

<i>outcome = language spoken with the first child</i>	ols	oprobit	$\Delta\text{Pr}(L = 1)$	$\Delta\text{Pr}(L = 2)$	$\Delta\text{Pr}(L = 3)$	$\Delta\text{Pr}(L = 4)$	$\Delta\text{Pr}(L = 5)$
exogenous language skills							
native Spanish speaker	-1.647*** (0.047)	-1.741*** (0.049)	0.326*** (0.011)	0.089*** (0.006)	0.073*** (0.005)	0.018*** (0.002)	-0.506*** (0.014)
oral proficiency in Catalan	0.264*** (0.009)	0.278*** (0.014)	-0.082*** (0.003)	-0.006*** (0.001)	0.006*** (0.001)	0.008*** (0.001)	0.074*** (0.003)
reduced form regressions							
native Spanish speaker	-2.228*** (0.052)	-2.173*** (0.053)	0.385*** (0.010)	0.124*** (0.009)	0.120*** (0.007)	0.042*** (0.004)	-0.671*** (0.010)
exposure $\times$ native Spanish speaker	0.046*** (0.014)	0.052*** (0.014)	-0.019*** (0.005)	-0.001*** (0.000)	0.002*** (0.001)	0.002*** (0.001)	0.016*** (0.004)
endogenous language skills							
native Spanish speaker	-1.379*** (0.243)	-1.114*** (0.412)	0.114*** (0.037)	0.067** (0.032)	0.087** (0.040)	0.042** (0.019)	-0.310** (0.126)
oral proficiency in Catalan	0.404*** (0.121)	0.483*** (0.087)	-0.111*** (0.009)	-0.013*** (0.002)	-0.005 (0.006)	0.002 (0.004)	0.128*** (0.020)
unconditional probability for native Spanish speakers			39.21%	14.25%	14.91%	7.04%	24.59%
number of observations	3667	3667					

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for child's age and individual's year of birth dummies. $L = 1$ (only Spanish), 2 (more Spanish than Catalan), 3 (equal Spanish and Catalan), 4 (more Catalan than Spanish), 5 (only Catalan). Oprobit = Ordered Probit model; endoprobit = ordered probit model with endogenous language skills.

rather than a linear specification (as in the baseline case, reported in column (1)), we run our IV using either a quadratic function of exposure (column (2)), dummies for partial or full exposure versus no exposure (column (3)) or dummies for each year of exposure (column (3)). The point estimate of the parameter of interest obtained under alternative functional forms is virtually the same as in the more parsimonious specification based on linear effects in the first stage.¹⁷

Second, we address the concern that the first stage regression is not capturing a genuine effect of bilingual schooling on skills, but rather pre-existing trends in the outcome (possibly unrelated to the policy change) that affect differently Spanish and Catalan speakers. This circumstance would invalidate the causal interpretation of the IV estimates. To gain insights into this potential issue, we proceed by estimating reduced form models like Eq. (6), either for the real reform (as in Table 5) and for several fake reforms for never-treated cohorts of individuals. To generate these placebo reforms, we constructed a sample of cohorts that never actually received bilingual education, and pretend that some of them were exposed to bilingualism at school. More specifically, we exclude from our original estimating sample cohorts of individuals treated by the reform (those born from 1970 to 1990) and include older cohorts of non-treated (1949–1945, stopping in 1945 owing to limitations in the availability of information of year of birth). Next, for the resulting sample of non-treated individuals born in 1945–1969, we assume that the reform was implemented in a placebo year, and re-estimate the reduced form effect for this placebo reform. We conduct the exercise for placebo years 1963 to 1970. The results presented in Table 7 show that the reduced form effect of the interaction between compulsory exposure and the dummy for being a native Spanish speaker is positive and statistically significant, as expected, with 1 year of exposure to bilingual education raising the probability of transmitting Catalan to children by 4.6 percentage points (p.p.) in the group of native Spanish speakers. In-

stead, the coefficients of placebo exposure interacted with the dummy for being a native Spanish speaker are insignificant and even negative in many cases, which goes against the idea that our exposure variable is capturing spurious heterogeneous trends across cohorts. Thus, the evidence suggests that our first stage regression is plausibly capturing the effect of the policy change, supporting our causal interpretation of the findings from the IV model.

Having obtained reassuring evidence with respect to the main component of our exclusion restriction (the compulsory exposure variable), we scrutinize its second component, the native language variable.¹⁸ This variable is self-reported and we are concerned about the possibility that the misreporting error could be endogenously driven by the reform. The reason is that some treated native Spanish speakers could define themselves as a native Catalan speaker just because they were exposed to Catalan at school. Therefore, we repeat our estimations using alternative proxies for native language, namely the language of the parents (a dummy that takes the value 1 if both parents have Spanish as their native language) and the origins of the parents (a dummy that takes the value 1 if both parents were born outside Catalonia). The results are reported in columns (2) and (3) of Table 8 and show that the point estimate when we use these two alternative proxies for native language (especially regional origins) is somewhat higher than the baseline, but the overall results remain unchanged. We also take advantage of the availability of these two proxies for native language in order to provide further evidence in favor of the validity of our identification strategy, which is ultimately based on the assumption that there is no direct effect of language-specific trends across the cohorts on the language spoken with the first child (the exclusion restriction). Specifically, we estimate two alternative over-identified models in which we use as exclusion restrictions the interactions between language exposure at school and both the original native language indicator and any of the two proxies for language background. Notice that the subpopulations of compliers that are captured by the available instruments are relatively similar (especially regarding native and parental language). Therefore, rejecting the null

¹⁷ As mentioned before (see footnote n. 14), we also took advantage of these overidentified model that include more flexible functional forms for exposure to implement the Hansen J-test for overidentification, which in all cases suggest that the (linear combination of) the instruments are not related to the residuals of the language use equation. We also checked that the results are stable to different specifications of the functional form of the effect of individual and child's age, as shown in Tables B7 and B8 in the Appendix.

¹⁸ As mentioned in Section 4, there are few individuals in our sample that report having both Spanish and Catalan as native language. Dropping these observations from the estimation sample does not alter the results (available upon request).

Table 6
Sensitivity to alternative definitions of the exclusion restriction.

	(1)	(2)	(3)	(4)
FIRST STAGE — outcome: oral skills in Catalan				
native Spanish speaker	-2.102*** (0.104)	-2.130*** (0.108)	-2.142*** (0.111)	-2.143*** (0.111)
exposure × native Spanish speaker	0.114*** (0.025)	0.232*** (0.085)		
exposure ² × native Spanish speaker		-0.016 (0.011)		
no exposure (year birth < 1970)			ref. cat.	
partial exposure (1970 ≤ year birth < 1977)			0.672*** (0.177)	
full exposure (year birth ≥ 1977)			0.764*** (0.236)	
exposure = 0 × native Spanish speaker				ref. cat.
exposure = 1 × native Spanish speaker				0.591*** (0.111)
exposure = 2 × native Spanish speaker				-0.083 (0.114)
exposure = 3 × native Spanish speaker				1.027*** (0.115)
exposure = 4 × native Spanish speaker				0.662*** (0.110)
exposure = 5 × native Spanish speaker				0.709*** (0.118)
exposure = 6 × native Spanish speaker				1.009*** (0.118)
exposure = 7 × native Spanish speaker				0.963*** (0.116)
exposure = 8 × native Spanish speaker				0.676** (0.249)
exposure = 10 × native Spanish speaker				1.356*** (0.449)
adjusted R-squared	0.222	0.223	0.222	0.223
F-test of excluded instruments	21.78	13.33	9.74	208,513
[p-value]	[0.000]	[0.000]	[0.000]	[0.000]
2SLS — outcome = language spoken with the first child				
native Spanish speaker	-1.379*** (0.243)	-1.297*** (0.233)	-1.331*** (0.224)	-1.382*** (0.209)
oral proficiency in Catalan	0.404*** (0.121)	0.447*** (0.116)	0.429*** (0.111)	0.403*** (0.103)
Hansen J-test		2.228	0.236	8.882
[p-value]		[0.136]	[0.627]	[0.352]
adjusted R-squared	0.512	0.491	0.500	0.512
number of observations	3667	3667	3667	3667

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for individual's and child's age and individual's year of birth dummies. The F-test on excluded instruments refers to the Angrist-Pischke multivariate F-test on the interactions between exposure to Catalan at compulsory schooling (continuous or indicators) and the indicator for Spanish as native language.

hypothesis of the corresponding overidentification test would indicate that the linear combination of the instruments affects the outcome directly, thus invalidating our identification strategy. However, estimates for oral skills in Catalan obtained from overidentified models (columns (4) and (5) of Table 8) are statistically undistinguishable from the baseline, which again points to the overall robustness of our results. Most importantly, the Hansen test for overidentification does not provide sufficient evidence against the null hypothesis that the instruments are not directly related to the language spoken with the child. This confirms the results obtained from overidentification tests based on functional form assumptions regarding the exclusion restriction, indicating that (controlling for common cohort effects), the interaction between the exposure variable and native language is unlikely to be correlated with the outcome's error term and thus represents a valid instrument.

7.2. Alternative mechanisms

Both our theoretical model and our empirical results emphasize the importance of language skills in changing the incentives to transmit the native language to the children. However, as discussed in the theoretical section, there may be alternative mechanisms that are consistent with our empirical findings.

The first potential mechanism is that the reform may have also changed the language preferences of the compliers. Language preferences are likely to be closely related (although not perfectly correlated) to ethnic or national identity. Thus, the evidence reported in Clots-Figueras and Masella (2013), indicating that the 1983 reform promoted the Catalan identity, provides a motivation to explore such an alternative mechanism. In particular, our findings could be driven by a change

Table 7
Falsification analysis (Baseline and Placebo Reduced Form Equations).

	real reform in 1983	placebo reform in:							
		1970	1969	1968	1967	1966	1965	1964	1963
native Spanish speaker	-2.228*** (0.052)	-2.298*** (0.092)	-2.287*** (0.094)	-2.266*** (0.094)	-2.243*** (0.095)	-2.218*** (0.096)	-2.187*** (0.098)	-2.168*** (0.103)	-2.152*** (0.112)
exposure × native Spanish speaker	0.046*** (0.014)								
exposure* × native Spanish speaker		0.010 (0.014)							
exposure* × native Spanish speaker			0.007 (0.014)						
exposure* × native Spanish speaker				0.002 (0.014)					
exposure* × native Spanish speaker					-0.003 (0.014)				
exposure* × native Spanish speaker						-0.007 (0.014)			
exposure* × native Spanish speaker							-0.013 (0.015)		
exposure* × native Spanish speaker								-0.015 (0.015)	
exposure* × native Spanish speaker									-0.017 (0.017)
adjusted R-squared	0.472	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
number of observations	3667	3037	3037	3037	3037	3037	3037	3037	3037

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for individual's and child's age and individual's year of birth dummies. The regression in the first column is reduced form equation for the baseline sample. Regressions in columns 2–9 are based on a subsample of never-treated individuals (born between 1945 and 1969, in Catalonia or migrated before age 6); placebo compulsory exposure (exposure*) is imputed “as if” the reform was implemented from 13 to 20 years before the real reform of 1983 (i.e. in 1970 instead of 1983 for column 2, in 1969 for column 3, and so forth). Descriptive statistics for the subsample of never-treated individuals are reported in Table B9 of Appendix B.

Table 8
Alternative proxies for native language.

	(1)	(2)	(3)	(4)	(5)
proxy for native language:	native language	parents' language	parents' origins	parents' language	parents' origins
FIRST STAGE — outcome: oral skills in Catalan					
native Spanish speaker	-2.102*** (0.104)			-1.568*** (0.097)	-1.643*** (0.116)
exposure × native Spanish speaker	0.114*** (0.025)			0.052* (0.027)	0.032 (0.027)
both parents native Spanish speakers		-1.980*** (0.119)	-1.900*** (0.126)	-0.656*** (0.123)	
exposure × both parents native Spanish speakers		0.125*** (0.028)	0.166*** (0.028)	0.078** (0.036)	
both parents born outside Catalonia					-0.654*** (0.152)
exposure × both parents born outside Catalonia					0.125*** (0.030)
F-test of excluded instruments	21.78	19.87	36.04	11.85	15.50
[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
adjusted R-squared	0.222	0.192	0.166	0.227	0.229
2SLS — outcome = language spoken with the first child					
oral skills in Catalan	0.404*** (0.121)	0.441*** (0.122)	0.485*** (0.111)	0.406*** (0.111)	0.388*** (0.100)
native Spanish speaker	-1.379*** (0.243)			-1.164*** (0.184)	-1.321*** (0.188)
both parents native Spanish speakers		-1.182*** (0.231)		-0.265** (0.105)	
both parents born outside Catalonia			-0.949*** (0.185)		-0.129 (0.079)
Hansen J-test				0.237	0.014
[p-value]				[0.626]	[0.905]
adjusted R-squared	0.512	0.464	0.404	0.518	0.519
number of observations	3667	3622	3662	3622	3662

Note: standard errors clustered by year of birth within parenthesis.

*** significant at 1%,

** significant at 5%,

* significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for child's age and individual's year of birth dummies.

in language preferences. Our theoretical setup clearly indicated that a lower relative preference for using the native language, or a switch in the preferred language, had the same qualitative effects as an improvement in proficiency in the second language. That is, native Spanish speakers who increased their skills in Catalan owing to the exposure to the reform, may have also reduced their preferences towards the use of Spanish with their children. Fortunately, the survey does contain a question

about the respondents' language of self-identification, which is coded in the same way as the native language. We believe that those respondents with a language of self-identification that is different from their native language are the ones that might have experienced a change in preferences. Therefore, we provide some suggestive evidence about the role of changes in language preferences induced by the reform by re-estimating the model without “language switchers” (individuals whose

Table 9

Alternative mechanisms (identity changes and partnership formation).

	(1)	(2)	(3)
OLS — outcome = language spoken with the first child			
native Spanish speaker	-1.647*** (0.047)	-2.502*** (0.067)	-2.391*** (0.078)
oral proficiency in Catalan	0.264*** (0.009)	0.141*** (0.010)	0.175*** (0.011)
adjusted R-squared	0.567	0.743	0.748
FIRST STAGE — outcome: oral proficiency in Catalan			
native Spanish speaker	-2.102*** (0.104)	-3.096*** (0.133)	-2.697*** (0.126)
exposure × native Spanish speaker	0.114*** (0.025)	0.196*** (0.035)	0.141*** (0.032)
adjusted R-squared	0.222	0.364	0.308
2SLS — outcome = language spoken with the first child			
native Spanish speaker	-1.379*** (0.243)	-1.742*** (0.195)	-1.704*** (0.265)
oral proficiency in Catalan	0.404*** (0.121)	0.418*** (0.070)	0.454*** (0.108)
adjusted R-squared	0.512	0.625	0.616
number of observations	3667	3039	2549

Note: standard errors clustered by year of birth within parenthesis. *** significant at 1%, ** significant at 5%, * significant at 10%. All regressions include dummies for wave and gender, cubic polynomials for individual's age and individual's year of birth dummies. Results in column (1) are obtained using the entire sample. Results in column (2) are obtained by excluding individuals with self-identification language different from their native language. Results in column (3) are obtained by excluding individuals whose partner's language is different from their native language.

self-identification language is different from their mother tongue.) The results, reported in the second column of Table 9, are very stable with respect to the baseline in column 1, indicating that changes in language preferences are not the main driver of the results.

The second alternative mechanism highlighted by the theoretical model is endogenous partnership formation. Indeed, one of the extensions of our benchmark model suggests that transmission choices may depend on the language characteristics of the partner (mixed versus homogeneous couples). Therefore, it is possible that the effects we found are simply reflecting that those native Spanish speakers who become proficient in Catalan after its introduction as a medium of instruction use this language more intensively with their children, simply because they are more likely to be in a linguistically mixed couple. To appreciate the relevance of this alternative mechanism as a plausible interpretation of our empirical results, in the third column of Table 9 we repeat the analysis excluding mixed couples (*i.e.* individuals whose partner has a different native language). This exclusion does not alter the substance of our findings, providing suggestive evidence that changing assortative patterns induced by the reform are not the main driver of the estimated effects.

8. Concluding remarks

We have estimated the effects of parents' language skills on the intergenerational language transmission in a bilingual society. We consider the case of Catalonia, where we can exploit the natural experiment generated by a language-in-education reform. We provide a theoretical framework highlighting the main channels through which this type of

policy can impact intergenerational language choices. Empirically, we use exposure to the reform to isolate the causal effect of skills on intergenerational transmission, and we rule out other channels such as couple formation and preference switching.

Our results show that by increasing skills in Catalan among native Spanish speakers, the reform has changed their intergenerational language choices, inducing them to speak Catalan more intensively with their children. Existing literature shows relevant effects of reform-induced Catalan skills on labor market outcomes and pattern of social interactions. We add to this literature by showing that these effects are not confined within the generation that is exposed to the policy, but may well spill over to the offspring of the treated cohorts. In this way, we uncover an additional channel of effectiveness of language-in-education reforms through the language skills of the treated cohorts, which makes for a long-term impact on social cohesion in bilingual societies.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.labeco.2021.101975.

References

- Bleakly, H., Chin, A., 2004. Language Skills and Earnings: evidence from childhood immigrants. *Rev. Econ. Stat.* 86, 481–496.
- Bleakly, H., Chin, A., 2008. What holds back the second generation? The intergenerational transmission of language human capital among immigrants. *J. Hum. Resour.* 43 (2), 267–298.
- Bleakly, H., Chin, A., 2010. Age at arrival, English proficiency, and social assimilation among US immigrants. *Am. Econ. J.: Appl. Econ.* 2 (1), 165–192.
- Casey, T., Dustmann, C., 2008. Intergenerational Transmission of Language Capital and Economic Outcomes. *J. Hum. Resour.* 43 (3), 660–687.
- Caminal, R., Di Paolo, A., 2019. Your Language or Mine? *Econ. Inq.* 57 (1), 726–750.
- Cappellari, L., Di Paolo, A., 2018. Bilingual Schooling and Earnings: evidence from a Language-in-Education Reform. *Econ. Educ. Rev.* 64, 90–101.
- Chiswick, B., Miller, P., 2007. *The Economics of Language: International Analyses*. Routledge, London.
- Chiswick, B., Miller, P., 2018. Do native-born bilinguals in the US earn more? *Rev. Econ. Household* 16 (3), 563–583.
- Church, J., King, I., 1993. Bilingualism and network externalities. *Can. J. Econ.* 26, 337–345.
- Clots-Figueras, I., Masella, P., 2013. Education, Language and Identity. *Econ. J.* 123, 332–357.
- Di Paolo, A., Raymond, J.L., 2012. Language knowledge and earnings in Catalonia. *J. Appl. Econ.* 15 (1), 89–118.
- Dustmann, C., Van Soest, A., 2004. An analysis of speaking fluency of immigrants using ordered response models with classification errors. *J. Bus. Econom. Statist.* 22 (3), 312–321.
- Egger, P.H., Lassmann, A., 2015. The causal impact of common native language on international trade: evidence from a spatial regression discontinuity design. *Econ. J.* 125 (584), 699–745.
- Frankel, J., Rose, A., 2002. An estimate of the effect of currencies on trade and income. *Q. J. Econ.* 117, 437–466.
- Grogger, J., 2019. Speech and wages. *J. Hum. Resour.* 54 (4), 926–952.
- Kuziemko, I., 2014. Human Capital Spillovers in Families: do Parents Learn from or Lean on Their Children? *J. Labor Econ.* 32 (4), 755–786.
- Méltiz, J., 2008. Language and Foreign Trade. *Eur. Econ. Rev.* 52, 667–699.
- Rendon, S., 2007. The Catalan premium: language and employment in Catalonia. *J. Popul. Econ.* 20 (3), 669–686.
- Selten, R., Pool, J., 1991. The distribution of foreign language skills as a game equilibrium. In: Selten, R. (Ed.), *Game Equilibrium Models*. Springer-Verlag, Berlin vol. 4, pp. 64–84.
- Siguan, M., 1991. The Catalan Language in the Educational System of Catalonia. *Int. Rev. Educ.* 37 (1), 87–98.
- Woolard, K.A., Gahng, T.J., 1990. Changing Language Policies and Attitudes in Autonomous Catalonia. *Lang. Soc.* 19 (3), 311–330.