YOUR LANGUAGE OR MINE? THE NON-COMMUNICATIVE BENEFITS OF LANGUAGE SKILLS[*]

Ramon Caminal[†]and Antonio Di Paolo[‡]

October 2017

Abstract

Do languages matter beyond their communicative benefits? We explore the potential role of preferences over the language of use, theoretically and empirically. We focus on Catalonia, a bilingual society where everyone is fully proficient in Spanish, to isolate linguistic preferences from communicative benefits. Moreover, we exploit the language-in-education reform of 1983 to identify the causal effects of language skills. Results indicate that the policy change has improved the Catalan proficiency of native Spanish speakers, which in turn increased their propensity to find Catalan-speaking partners. Hence, the acquisition of apparently redundant language skills has reduced endogamy.

JEL Classifications: C26, C78, I28, J12, J15, Z13.

Keywords: bilingualism, non-communicative benefits, language preferences, endogamy, language skills, language-in-education policy.

[*] We thank Lorenzo Cappellari, Ana Rute Cardoso, Antonio Ciccone, Lidia Farré, Ada Ferrer-i-Carbonell, and Louis-Philippe Morin and for their helpful comments, Frederic Udina (IDESCAT's director) for facilitating the access to the dataset, Josep Maria Martínez and other IDESCAT staff for their assistance, and the Spanish Ministry of Education (grants ECO2013-41022-R and ECO2014-59959-P), Generalitat de Catalunya (SGR 2014-2017), and Research Caixa (2016ACUP00020, Obra Social laCaixa and ACUP) for their financial support. The usual disclaimers apply.

[†] **Corresponding author**. Institut d'Anàlisi Econòmica (CSIC) and Barcelona GSE: ramon.caminal(at)iae.csic.es, (34)935806612.

[‡] AQR-IREA, Universitat de Barcelona: antonio.dipaolo(at)ub.edu, (34)934037150.

1 Introduction

The days when most human beings could go through their life using exclusively their native language are long gone. The latest wave of globalization, and The Internet in particular, has dramatically increased individuals' exposure to multiple languages. It has been estimated that more than one-half of the world's population speak more than one language (Tucker, 2001). Thus, it is not surprising that multilingualism is attracting a great deal of attention, also among economists. Indeed, economic research has clearly established that language skills matter for economic outcomes. For instance, it has been shown that sharing a common language promotes international trade (e.g., Frankel and Rose, 2002; Melitz, 2008, Egger and Lassmann, 2015). Also, evidence from a variety of countries indicates that fluency in the host country's language has a large effect on immigrants' earnings (e.g., Bleakley and Chin, 2004; Chiswick and Miller, 2007). Not surprisingly, these results have been mostly attributed to the role of languages as communication devices. After all, the ability to communicate is crucial in trade, as well as in production.

Undoubtedly, the acquisition of additional language skills is bound to facilitate communication and reduce production and transaction costs. However, to focus exclusively on this dimension, and characterize languages as interchangeable communication codes, can easily lead to quite extreme views. Specifically, Church and King (1993) concluded that multilingual societies should only promote the majority language and hence restrict the use of minority languages to intra-community exchanges.[1] In a similar vein, Jones (2000) argued in favor of a convergence towards a single world language. The central argument is analogous to the benefits of technological compatibility. If languages are alternative, equally efficient standards, the social optimum requires standardization. From this perspective, the death of languages is seen as a natural, and even desirable, phenomenon in an increasingly globalized world. Similarly, policies that protect minority languages and promote linguistic diversity are suspected of pandering to narrow interests, and presumed harmful for the society as a whole.

By and large, economists have recognized that languages are much more than neutral communication devices. A prominent example is the recent book by Gins-

[1] They formalized the idea that learning a second language generates network externalities; as a result individuals underinvest in the acquisition of second languages, which opens the door to public intervention. The optimal policy includes a subsidy on learning the majority language.

burgh and Weber (2011). They note that preserving linguistic diversity involves non-negligible costs. However, individuals tend to develop some kind of emotional attachment to the language that better defines their identity; therefore, limiting the number of languages also generates losses. Hence, policy makers should pay attention to both the role of languages as means of communication as well as their subjective, emotional aspects.

The relevance of the non-communicative aspects of languages can be also inferred from two other strands of the economics literature. First, several studies (including Alesina et al., 2003 and Desmet et al., 2012) use language as a proxy for ethnicity or culture in order to examine the effects of ethnic or cultural diversity on civil conflict and redistribution. Second, certain language characteristics have been linked to values and economic behavior. In particular, Chen (2013) shows that languages that grammatically associate the future and the present foster forward-looking behavior. In a similar spirit, Gay et al. (2013), demonstrate that women speaking languages that more pervasively mark gender distinctions are less likely to participate in economic and political lives.[2]

In this paper we examine the non-communicative aspects of languages both theoretically and empirically. In contrast to the existing literature, we focus on the effects of acquiring a second language. In particular, we show that the acquisition of language skills that are redundant from a communicative viewpoint can significantly influence the pattern of social interactions, undermining endogamic behavior. We interpret such non-communicative effects as arising from a broad notion of linguistic preferences: most individuals develop an emotional attachment to their native language and, even if fully bilingual, prefer to use it over their second language. Clearly, linguistic preferences may also emerge from the ties between language and culture, and reflect ethnic or political identity. In any case, it is important to note that our theory focuses on the effect of language skills on social behavior, taking preferences as exogenous. Nevertheless, the interpretation of the empirical results may also depend on the nature of preferences, and hence we take this issue up again in Section 6.

More specifically, we first provide a theoretical framework that illustrates a new channel by which the distribution of language skills in a bilingual society affects the pattern of social interactions. We build on standard theory and assume that

[2] See also Galor et al. (2016) and their list of references.

sharing a common language enhances economic and social interactions.[3] On top of this, we assume that even fully bilingual individuals have a preference for using their native language or the language adopted as their own in later stages.[4] We model a bilingual society with an initial asymmetric distribution of language skills: all native speakers of the weak language are bilingual, with full command of both the strong and the weak language, but most native speakers of the strong language are either monolingual or only partially proficient in the weak language.[5] Thus, all agents share a common language, and hence the role of linguistic preferences can be isolated from the communicative benefits. Cooperation (trade partnerships, marriages, etc.) requires communication and hence the use of a particular language. Such a choice is trivial when all partners belong to the same speech community. However, in the case of mixed partnerships, individuals with strong linguistic preferences may reject optimal partners (in terms of non-linguistic dimensions) and instead match with less desirable, but linguistically homogeneous, partners. In other words, the formation of mixed partnerships requires a satisfactory resolution of a linguistic conflict. The crucial observation is that the intensity of the conflict varies with language skills. In particular, as native speakers of the strong language improve their skills in the weak language: (i) the frequency of mixed partnerships increases, (ii) the use of the weak language also increases.

It is important to note that, if we abstract from learning costs, such an improvement in language skills increases total surplus. That is, the promotion of language skills that do not expand the ability to communicate generates social benefits, that need to be measured against the learning costs. Thus, policies that promote minority languages can be justified not only in terms of fairness (Van Parijs, 2011) but also, under some conditions, on efficiency grounds. The intuition behind these benefits is that the equilibrium rate of mixed partnerships is inefficiently low, because individuals do not internalize the negative externalities inflicted on their potential partners when they unilaterally decide to match with an inferior but linguistically

[3] See, for instance, Selten and Pool (1991), Church and King (1993), and Weber et al. (2011).

[4] Some kind of linguistic preferences have already been introduced in a variety of economic frameworks. See, for example, Grin (1992), Wickström (2005), Caminal (2010), and Mèlitz (2012). Our main focus is on how language skills and preferences affect cooperation between speech communities.

[5] The relative strength of the two languages do not necessarily reflect the relative size of their local speech communities. A language may be strong because of its status and prestige, or because it is widely spoken outside the country or region (think of Russian in Latvia, or English in Quebec) and hence incentives to learn it may surpass its local communicative benefits. See the next section for precise definitions.

homogeneous partner. Thus, the increase in mixed partnerships generated by the additional language skills is bound to raise total surplus.

Next we empirically test these predictions using survey data originated in the particular but very fitting case of Catalonia (Spain). Two main reasons make Catalonia a unique test field. First, it is a bilingual society (Spanish and Catalan are the two main languages) where the ability to communicate is not at stake because everyone speaks the strong language (Spanish), just as in the theoretical model. Hence, any implications of additional language skills must be attributed to linguistic preferences. Second, new language-in-education policies were introduced three decades ago, after the approval in 1983 of the Language Normalization Act (LNA). With the implementation of this reform, education experienced a smooth transition from a system in which Catalan was excluded to one in which Catalan has become the main language of instruction in compulsory education. This reform led to a significant improvement of the Catalan skills of native Spanish speakers, whereas all other language skills remained basically unchanged.[6] Hence, the heterogeneous effect of language exposure during compulsory education allows us to generate quasi-experimental variation in the variables of interest.

The main goal of the empirical analysis is to study the influence of improved language skills among native speakers of the strong language (Spanish) on their propensity to form a linguistically-mixed couple and the use of the weak language (Catalan) with the partner.[7] In order to identify the causal effect, we exploit an Instrumental variable based on the differential effect by native language of exposure to Catalan as a language of instruction during compulsory schooling. Compulsory language exposure was already considered as an exogenous determinant of identity formation by Clots-Figueras and Masella (2013) in a reduced-form framework.[8] Here, we exploit the interaction between compulsory exposure and the indicator for

[6] We are referring to oral skills, which are the most relevant regarding the formation of a couple. As discussed in Section 4, written skills in Catalan improved for both Spanish and native Catalan speakers, although much less so for the latter group, and Spanish skills remained at very high levels for both speech communities.

[7] It has been shown (Bleakly and Chin, 2010, Furtado and Theodoropoylos, 2011; and Chiswick and Hoseworth, 2011) that the frequency of inter-ethnic marriages among US immigrants is positively affected by English-speaking ability. See also Meng and Meurs (2009) for the case of France. Since the proficiency of individuals in the strong language varies a lot from individual to individual, these studies cannot distinguish between linguistic preferences and communicative benefits.

[8] Thus, they study the effects of the same education reform, but focus on a different topic and use a different dataset. They find that attending compulsory schooling after the LNA reform reinforces individuals' self-identification as "Catalans". See also Aspachs et al. (2008). In Sections 5 and 6 we discuss whether identity considerations matter in interpreting our empirical results.

being native Spanish speaker as identifying variable in a Two-Stage Least Squares (2SLS) setting. This exclusion restriction captures the improvement in oral fluency in Catalan among native Spanish speakers that was induced by reform exposure during compulsory schooling. The main underlying assumption behind the validity of this identification strategy is that non-linguistic cohort effects are common for both linguistic communities (in the spirit of the identification strategy originally proposed by Bleakley Chin, 2004, 2008, 2010). Several robustness checks and falsification exercises are carried out in order to validate the use of such exclusion restriction.

Our results are in line with the theoretical predictions. In particular, the 1983 education reform, by improving the oral Catalan skills of native Spanish speakers, raised their propensity to find a Catalan speaking partner and to speak Catalan with the partner. These results are robust to a battery of sensitivity checks, and clearly indicate that linguistic preferences are relevant. In particular, the acquisition of language skills that appear redundant from a communicative viewpoint can significantly reduce segregation.

In the next section we lay out the theoretical framework and derive two testable hypothesis. In Section 3 we provide some historical background and describe the data. Section 4 discusses some descriptive evidence. The main results as well as the robustness and sensitivity tests are presented in Section 5. Finally, Section 6 summarizes the paper and discusses alternative interpretations.

2 The theory

Consider a country with two languages, A and B. A fraction α of the population is initially socialized in A (they are native A speakers), and a fraction $1-\alpha$ in B (native B speakers). Everyone is fully competent in their mother tongue. These two languages differ in their status and knowledge. In particular, all native B speakers are also fully proficient in language A, but only some native A speakers are proficient in language B. Because of the (domestically) universal knowledge we call language A the strong language, and B the weak language. Perhaps, these asymmetric language skills are induced by the fact that A is widely known in the rest of the world and hence very useful for communicating with foreigners.[9] In any

[9] Another reason could be that knowledge of A provides access to an abundant supply of media outlets and leisure goods produced in that language.

case, we take language skills as exogenous, and the identification of a language as strong or weak as country-specific. Thus, a particular language can be weak in one country or region and strong in another.[10] In spite of the universal knowledge of the strong language, the existence of different speech communities (defined according to native languages) still matters because individuals develop a preference towards their initial language, as specified below.

Individuals derive utility from forming partnerships with other compatriots (e.g., trade partnerships, couples).[11] In particular, each individual can match a single person. The level of utility obtained from a partnership depends on linguistic as well as non-linguistic factors. With respect to the latter, for each agent i there is a single best match, j, which is reciprocal (so that j's best match is also i). The best match generates, for each partner, a level of utility $g_{ij} > 0$ (pair-specific). For simplicity, we assume that all other potential matches provide the same level of utility, which is normalized to zero.

The activities of the partnership require communication, and hence the use of a particular language. Everyone has a preference for using their native language. Hence, if the two members of a best match belong to the same speech community,[12] then nothing prevents the formation of the best match, since each partner obtains g_{ij}, which is higher than any alternative. However, if they belong to different speech communities (a mixed match), then language preferences can prevent the formation of the best match. More specifically, let individual a be the native A speaker, and b the native B speaker of a mixed match. If they form the partnership and choose A as the language of communication, then a and b would obtain a payoff of g_{ab} and $g_{ab} - w_b$, respectively. That is, individual b incurs a cost w_b for using their second language. Individuals differ in the intensity of their linguistic preferences. In particular, w_b is the realization of a random variable w distributed over some interval $[0, \overline{w}]$ with density function $f(w)$, and distribution function $F(w)$. We

[10] The universal knowledge of the strong language guarantees communication, independently of the knowledge of the weak language. The model literally apply to cases like Catalonia, Wales or the Basque Country. However, in other cases like Belgium or Quebec some speakers of the weak language (Flemish and French, respectively) remain monolingual. The model can be easily extended to take into account a fraction of monolingual speakers of B. In that case, language skills will affect segregation not only through linguistic preferences but also through changing the ability to communicate.

[11] For simplicity, we ignore potential foreign partners.

[12] If everyone has the same probability of being i's best match, independently of their native language, then the probability of a linguistically homogeneous best match is α for a native A speaker and $1 - \alpha$ for a native B speaker.

assume that $f(w) > 0$ for all $w \in [0, \overline{w}]$ and there are no mass points. If instead they choose B, then their payoffs would be $g_{ab} - \eta_a - w_a$ and g_{ab}, respectively. That is, if individual a uses B instead of A, this incurs an extra cost of $w_a + \eta_a$, where w_a represents again the cost for using a's second language (pure preference), whereas $\eta_a \geq 0$ represents the disutility caused by a limited proficiency in the second language. Hence, individuals with a better command of B have lower values of η. For simplicity, we assume that both speech communities have identical distributions of pure preferences. That is, w_a and w_b are two independent realizations of the random variable w. Whereas w is a fixed individual characteristic, η vary as a becomes more proficient in B.[13] The value of the outside option for both partners is 0 since there is always a member of their own speech community among their second best partners.

Given the set of values $(g_{ab}, w_a, \eta_a, w_b)$, the two potential members of a mixed match must decide whether or not to form the partnership, and the language of use in case they do. Our main qualitative results rely on the existence of some kind of bargaining friction. For expositional convenience, we consider the following environment. First, partners negotiate under full information about the relevant parameters. Second, if both parties agree on forming the partnership, then they choose the language that maximizes the joint surplus. Thus, the only friction is the absence of monetary compensations (non-transferable utility). At the end of this section we discuss some alternative frameworks that provide very similar insights and qualitatively identical comparative statics and welfare results.

Hence, in our set up a will accept forming the partnership and use B only if $g_{ab} - \eta_a - w_a \geq 0$. Similarly, b will accept using A only if $g_{ab} - w_b \geq 0$. These two participation constraints imply that in equilibrium the coalition will be formed if and only if

$$\min\{\eta_a + w_a, w_b\} \leq g_{ab}$$

Thus, individuals do not internalize the negative externality imposed on their potential partners in case they unilaterally decide not to form the partnership. Therefore, if decisions were instead taken by a social planner aiming at maximizing total surplus (first best), then the best match would be formed if and only if

$$\min\{\eta_a + w_a, w_b\} \leq 2g_{ab}$$

[13] It would make sense to assume that a's limited competence in B, $\eta_a > 0$, can also reduce b's payoff. No qualitative result would be affected by such an adjustment.

Figure 1a depicts the equilibrium outcome (i.e., when individuals are allowed to unilaterally reject the best match), for the case $\overline{w} > 2g_{ab}$. The region marked with N (no best match) corresponds to the case where one of the parties prefers not to make the match. Regions marked with A and B correspond to the cases where the partnership is formed and that particular language, A or B, is selected.

Figure 1b represents the socially efficient outcome (the solution that maximizes total surplus). Comparing the two figures, it becomes apparent that there is a region of parameter values for which the best match is not formed in equilibrium but should form according to the first best.[14]

In order to avoid uninteresting technical issues, in the rest of the exposition we will focus on the case that $g_{ab} = g$ and η_a is distributed on $\left[\underline{\eta}, \overline{\eta}\right]$ with a density function that takes strictly positive values in this interval, and has no mass points. Moreover, η_a and w_a are assumed to be independent variables. It will be convenient to first compare two extreme scenarios. Suppose first that $\underline{\eta} \geq g$ (Scenario 0). That is, all as are essentially monolingual. In this case, B will never be used in a mixed match, and hence the best match will be formed if and only if $w_b \leq g$. Alternatively, suppose now that all as are fully competent in B: i.e., $\overline{\eta} = 0$ (Scenario 1). In this case, the two languages are in a symmetric position, which generates a symmetric outcome: each language is used with a fifty percent chance. Moreover, the fraction of best matches that materialize is higher than in Scenario 0. That is: (i) if $w_b \leq g$, as in Scenario 0, all best matches happen; moreover, (ii) if $w_b > g$, then those matches where $w_a \leq g$ also materialize.

The comparative statics are analogous if we consider gradual, but general changes in η_a. More specifically, for all $\alpha \in (0, 1)$, if we start from a situation where $\underline{\eta}< g$ (i.e., a positive fraction of as are willing to make the best match and use B) and there is a shift in the distribution of η_as such that the final distribution is first-order stochastically dominated by the initial distribution, then:

Result 1 (i) the fraction of successful mixed matches increases, and (ii) B is used more often in those matches.

See the Appendix for details.

[14] Instead of choosing between A and B, we could have allowed linear combinations of the two languages, assuming, for instance, that individual utility decreases linearly with the fraction of time in which the second language is used. The qualitative results would remain unchanged.

Result 1 contains the main hypothesis we want to test in the empirical analysis. That is, an exogenous improvement in the proficiency in the weak language on the part of native speakers of the strong language reduces segregation and fosters the use of the weak language.

We can now investigate the welfare consequences of such a change in language skills. First, we focus again on the two extreme scenarios. If all as are monolingual (Scenario 0), then the average payoffs to the as and bs, when their best match is linguistically mixed, are given by:

$$U_a^0 = F(g)\,g$$

$$U_b^0 = F(g)\,g - \int_0^g w_b dF(w_b)$$

Thus, the best match will materialize with probability $F(g)$, in which case each party obtains g. However, the bs bear all the costs of using their second language. That is, in Scenario 0, bilinguals are worse off than monolinguals.

Alternatively, if all as are also fully competent in B (Scenario 1), then the average payoffs are

$$U_a^1 = U_b^1 = F(g)\,g - \int_0^g \int_0^{w_a} w_b dF(w_b)\,dF(w_a) + [1 - F(g)]\left[F(g)\,g - \int_0^g w_b dF(w_b)\right]$$

Consider b's expected utility (it is symmetric for the as). With probability $F(g)$, $w_a < g$, the match is feasible and each member obtains g, which explains the first term of the above expression. However, in this region, b incurs the cost of using A whenever $w_b < w_a$, which is the second term. Also, $w_a > g$ with probability $1 - F(g)$. In this case, the match is feasible only if $w_b < g$, in which case b always incurs the full costs of using A, which is the third term.

Note that the bs are better off in Scenario 1: $U_b^1 > U_b^0$. Also, the total surplus is higher in Scenario 1: $U_a^1 + U_b^1 > U_a^0 + U_b^0$.[15] However, the as may be better off in Scenario 0 or in 1: $U_a^1 \lesseqgtr U_a^0$. The reason for this ambiguity is the following. Compared to Scenario 0, in Scenario 1, on the one hand, a benefits from the higher frequency of successful best matches, which increases from $F(g)$ to $F(g)[2 - F(g)]$. On the other hand, they lose their power to impose their preferred language, and

[15] It is important to emphasize that the welfare of individuals involved in a homogeneous match does not change across regimes. Thus, changes in total welfare are entirely driven by changes in the welfare of individuals involved in a mixed match, U_a and U_b; and since the number of a's and b's involved in a mixed match is the same, the sign of the change in total welfare is the same as the sign of the change in $U_a + U_b$.

have to bear half of the costs of using their second language.[16]. In other words, even abstracting from learning costs, native A speakers may or may not benefit from learning B. In contrast, native B speakers always benefit from this change, since on top of the higher frequency of successful best matches, they enjoy a better language treatment.[17] Finally, the total surplus is always higher in Scenario 1. That is, in case native A speakers lose, they lose less than the amount gained by native B speakers. The reason is twofold. Scenario 1 generates: (i) a higher rate of occurrence of best matches, and (ii) it allows a reduction in the total discomfort from using the second language, since B can now be used whenever $w_a < w_b$.[18]

In the Appendix we show that the same comparative statics hold for gradual but general changes in η_a. That is, for all $\alpha \in (0,1)$ if we start from a situation where $\underline{\eta} < g$, and there is a shift in the distribution of η_as, such that the final distribution is first-order stochastically dominated by the initial distribution, then:

Result 2 (i) Native B speakers are better off, (ii) native A speakers may be better-off or worse-off, and (iii) aggregate welfare increases.

Thus, if we abstract from learning costs, an exogenous improvement in the proficiency in the weak language among native speakers of the strong language raises total welfare. However, it may also have non-trivial distributional implications.

The model presented in this section is highly stylized. In the working paper version (Caminal and Di Paolo, 2015) we discuss various possible extensions and interpretations. None of these additional considerations affects the main message. In particular, one may argue that the assumption of non-transferable utility could be highly restrictive in some applications. If we allowed for monetary compensations then we would need to invoke informational asymmetries (on linguistic preferences, for example). As it is well known, bargaining under asymmetric information results in excessively frequent break-ups (Myerson and Satterwaite, 1983). In such a framework, changes in the language skills of native A speakers also reduce the inefficiency associated to asymmetric information, and Results 1 and 2 still hold.[19]

[16] For example, if $f(w) = \frac{1}{\overline{w}}$, then $U_a^0 - U_a^1$ takes a positive value if $\overline{w} - g$ is sufficiently small, and takes a negative value if $\overline{w} - 2g$ is also sufficiently small.

[17] Notice that the third term of U_b^1 is positive and the second term has a lower absolute value than the second term of U_b^0.

[18] Notice again that the third term of U_a^1 is positive. Also, $2\int_0^g \int_0^{w_a} w_b dF(w_b) dF(w_a) < \int_0^g w_b dF(w_b)$.

[19] Alternatively, we could model the matching process as the result of directed (costly) search decisions. Individuals might join a bunch of social activities in order to find their optimal partners.

3 Empirical analysis: preliminaries

3.1 Historical background

Catalan can be regarded as the native language of Catalonia. It is a Romance language, originating from Latin in the territory in the ninth century. Spanish (Castilian), another Romance language, arrived in Catalonia as early as the fifteenth century and consolidated its position among the elites during the eighteenth century. The general population remained primarily monolingual in Catalan, and only gained access to Spanish with the expansion of elementary education, which was relatively slow.[20]

During Franco's dictatorship (1939–1975), Catalan was restricted to the private sphere, and nevertheless transmitted (mostly orally) from parents to children in a large fraction of the native Catalan families. Towards the second half of this period, efforts to revive Catalan as vehicle of culture intensified, although those efforts systematically clashed with the legal frame and often resulted in fines, or exile and jail sentences. In contrast, Spanish was the only official language and the only language used in education. Moreover, the social use of Spanish in Catalonia was strongly reinforced by the massive migration from southern Spain (especially in the 1960s). By the end of the 1970s, Catalan was the native language of almost one-half of the population, who at the same time were fully competent in Spanish. In contrast, most of the native Spanish speakers (40% of the population of Catalonia had been born outside the region) were monolingual or only passively bilingual (Woolard and Gahng, 1990; Siguan, 1991). Regarding attitudes and social prestige, Catalan was in a somewhat awkward position. On the one hand, it was a language excluded from public life, but at the same time the language of a large fraction of the better educated: the middle and the upper-middle class.[21] The social composition of its native speakers is probably crucial to explain the vast political support for "normalizing" the use of Catalan in the post-Franco era.

Right after the constitution of the Catalan regional government (the Autonomous

If different activities are conducted in different languages, then language skills and preferences will also affect the formation of mixed partnership in a way similar to the stylized model we have presented in the main text.

[20] Massive school enrollment did not take place in Spain until the twentieth century. In 1872 the percentage of the primary-school age population enrolled in school was only 42%, far below the levels prevailing in contemporary France and England (Nohoglu Soysal and Strang, 1989).

[21] The economic elite and those social groups in direct contact with Franco's regime adopted Spanish as the unique language in their repertoire.

Community), the regional parliament passed in 1983 (unanimously) the “Language Normalization Act” (LNA), which set the legal framework that allowed the dramatic changes in language-in-education policy that occurred over the next two decades. The LNA aimed at making all pupils fully competent in both languages (Spanish and Catalan) by the end of compulsory education. It also defined an integrative education model, in which children were not separated on the basis of the language spoken at home. The application of the LNA was gradual. In the period 1984–1993, the two languages were both used as the language of instruction in proportions that varied geographically, depending on the linguistic characteristics of the students and teachers’ language skills. Throughout this period the average fraction of subjects taught in Catalan increased significantly over time.

As a result, at the beginning of the 1990s, Catalan had become the preferred language of instruction in most primary schools, although Spanish was still dominant in secondary education (Artigal, 1997). Since 1994, the authorities gave Catalan full priority as the language of instruction in all public educational institutions, but in practice Spanish has also been used, particularly in secondary education (Muñoz, 2005). In summary, education experienced a gradual transition from a system from which Catalan was excluded to one in which Catalan has become the main language of instruction, at least in compulsory education.[22]

Such an asymmetric treatment of the two languages has apparently produced a fairly symmetric distribution of language skills. At the end of compulsory education, students’ levels of proficiency in Catalan and Spanish are similar (Consell Superior d’Avaluació del Sistema Educatiu, 2013). Moreover, the level of proficiency in Spanish of students coming out of Catalan schools is similar to the rest of Spain (Instituto de Evaluación, 2011). From a dynamic perspective, the educational reform improved the oral Catalan skills of native Spanish speakers (and the written skills of both native Catalan and native Spanish speakers), with basically no effect on the Spanish skills of either speech community.[23]

The regional authorities also sought to promote the knowledge and use of Catalan using a variety of means, including a Catalan-only TV channel, several catalanization campaigns, and language proficiency requirements for public sector jobs.

[22] The education reform affected not only the language of instruction. New textbooks and instructional materials replaced the ones produced under the supervision of Franco’s educational authorities, and new generations of school teachers, better educated and more proficient in Catalan, joined the system. Also, specialized teachers were hired to fulfil the LNA’s objectives.

[23] See also Vila (2008) and references contained there.

The results of these policies have been mixed. The use of Catalan by the overall population has never exceeded 50%. Regarding specific environments, the use of Catalan is preeminent in the regional and local governments and, more generally, in the political life of the region. In contrast, its use in other branches of government (for example, the judiciary) is close to zero. Similarly, cultural activities and media outlets also exhibit very heterogenous linguistic patterns. For example, whereas about 50% of the radio audiences consume programs in Catalan, less than 5% of movies projected in Catalan theaters are either originally filmed or dubbed into Catalan.

3.2 Data and descriptive statistics

The data used in the empirical analysis are drawn from the *Survey of Language Use of the Catalan Population*, a representative survey that is carried out by the Catalan Statistical Institute (IDESCAT). We use two cross-sections (waves 2008 and 2013), which originally contain 6,767 and 7,255 observations, respectively. The database is unique, especially regarding sociolinguistic characteristics. On top of the standard socio-demographic variables (gender, year of birth, place of birth, place of residence, education, etc.), it reports various linguistic variables of special interest for our analysis: the respondent's native language (first language spoken at home during childhood), the language of self-identification, as well as the respondent's proficiency (understanding, speaking, writing and reading) in both Catalan and Spanish. All these variables are self-reported. The survey also includes several questions about the respondent's (current or former) spouse or partner.[24] We pay special attention to the partner's language[25] and to the relative use of Catalan (with respect to Spanish) with the partner. Moreover, the survey also includes detailed information about family background and parental language habits.

The restricted sample used in the baseline analysis includes individuals born in Catalonia and those born in the rest of Spain who migrated to Catalonia at age 6 or earlier. The goal is to focus exclusively on individuals who completed their entire schooling in Catalonia. In order to reduce possible recall bias and selective

[24] We do not know the legal status of their relationship (married or not), but we do know whether or not they live together. In fact, some of our results are strengthened when we restrict the analysis to *stable* couples (those who live together).

[25] Unfortunately, we do not know the partner's year of birth or his/her language skills (only native language). Hence, we need to restrict attention to the respondents' language skills and year of birth.

mortality, we exclude individuals born before 1950. Respondents born after 1990 (i.e. individuals younger than 18 in 2008) and those who were students at the time of the survey are also excluded from the analysis. Given the main research question, it is also natural to exclude individuals who never had a partner (less than 7% of the restricted sample). Finally, in order to reduce the degree of unobserved heterogeneity in the data, we also discard the very few remaining observations of individuals whose native language or whose partner's native language is neither Spanish nor Catalan. The resulting restricted sample has 5,357 observations, 2,553 from the 2008 wave and 2,804 from the 2013 wave.

Individuals' native languages (as well as self-identification language) are classified into three categories: (1) only Catalan, (2) both Catalan and Spanish, and (3) only Spanish. In the baseline analysis we define a native Spanish speaker if the respondent chose option (3), only Spanish, as their native language; and a native Catalan speaker, otherwise. According to this definition, native Spanish speakers amount to about 45% of the restricted sample. Of course, we checked that the main results are robust to alternative definitions.

The language proficiency variables are coded with a 0–10 scale. In our analysis we focus on oral skills (and in particular, the ability to speak), which are much more relevant in couple formation. Figure 2 displays the average oral proficiency in Catalan and Spanish (and a quadratic fitted line) by year of birth, for both native Spanish speakers and native Catalan speakers. As expected, oral Catalan proficiency is uniformly high for native Catalan speakers (who acquired oral competency during childhood within the family), whereas successive cohorts of Spanish speakers exhibit a clear positive trend. Moreover, oral Spanish fluency is very high and stable across cohorts for both speech communities (differences in average proficiency across speech communities for each cohort are not statistically significant).

Thus, native Catalan speakers are largely bilingual (with an full command of both languages), whereas earlier generations of native Spanish speakers had a limited command of Catalan, and younger generations are becoming increasingly bilingual. Although several factors could be responsible of the trend in oral proficiency observed for native Spanish speakers observed in the raw data, it seems plausible that the language-in-education reform of 1983 is one of the main reasons behind such a positive trend. Indeed, in the identification strategy that we adopt to recover causal estimates, we only exploit the variation in oral language skills that induced by the different degree of exposure of successive cohorts of native Spanish speakers to the language-in-education reform (which is arguably an exogenous component of the positive trend in language fluency). For the sake of comparison, Figure 3a displays written Catalan skills. Note that written proficiency improves for the younger cohorts of both speech communities, with a more pronounced increase for native Spanish speakers. Also, the level of written Spanish proficiency (Figure 3b) is uniformly high and virtually identical for both speech communities.[26]

The partner's language is also classified into the same three categories as the respondent's native language. In the baseline analysis, consistently with the definition of the respondent's native language, we define a respondent's partner as a Catalan speaker if either option (1) or (2), Catalan-only or Catalan and Spanish, is reported. Language use with the partner is instead coded with an ordinal scale (from 1 to 5): (1) only Catalan, (2) more Catalan than Spanish, (3) equal Catalan and Spanish, (4) more Spanish than Catalan, and (5) only Spanish.[27] In the empirical analysis we choose a strict definition of the use of Catalan: we say a respondent uses Catalan with the partner if option (1) has been reported: i.e., only Catalan. Once again, various robustness checks have been conducted.

Table 1 shows that Catalan society is noticeably fragmented along linguistic attributes. In particular, about two-thirds of native Spanish speakers have a partner who speaks only Spanish. Since we have assigned intermediate cases to the Catalan speaking community, the level of endogamy for native Catalan speakers is even higher (about three-quarters). An important observation is that endogamy is related to language skills. More specifically, native Spanish speakers with high oral proficiency in Catalan (with an index greater than or equal to 8) have a significantly lower level of endogamy (about 7 percentage points less). Similarly, the fraction of native Spanish speakers that use only Catalan with their partner also increases by a similar amount when we condition on high proficiency in Catalan.

[26] Note that this evidence clearly identifies Spanish as the *strong* language, as defined in the theoretical model: that is, the language shared by all speech communities. This evidence is also compatible with the results of the systematic tests mentioned above conducted by the national educational authorities.

[27] The distribution of this variable is quite concentrated on the extreme options, (1) and (5): only 16% of the sample report an intermediate option.

4 Descriptive evidence: OLS estimates

We consider two different left-hand-side variables: (i) an indicator that takes the value of 1 if individual i is matched with a Catalan-speaking partner, and zero otherwise, and (ii) an indicator that takes the value of 1 if individual i uses only Catalan with their partner, and zero otherwise. For each of the two outcomes, we specify a linear probability model (OLS):

$$Y_{it} = \alpha + \beta' X_i + \delta Cat_i + \theta_t + \varepsilon_{it} \tag{1}$$

where the outcome Y of individual i born in year t depends on a set of controls, X, oral proficiency in Catalan, Cat, year of birth fixed effects, θ, and a random disturbance, ε. The coefficient of interest is δ. We start with a parsimonious specification that includes as controls a dummy for wave, a gender indicator, and a cubic polynomial of age, which picks up age differences that are not fully captured by cohort dummies.[28]

We next include several controls for parental background (parents' place of birth, education, native language) and for individual attributes (place of birth, place of residence, and completed education). The full set of control variables is presented in Table 2, together with basic descriptive statistics.

We start by presenting the results obtained for the subsample of native Spanish speakers. Selected estimates for the two outcomes are presented in Table 3 (the complete results can be found in Tables A1a and A1b in the online Appendix). The estimates from the baseline specification (column a) indicate that a marginal increase in oral proficiency in Catalan is associated with an increase by about 4.5 percentage points in the probability of having a Catalan-speaking partner. Similarly, better skills in Catalan is associated, to a similar extent, with a higher likelihood of using only Catalan with the partner. These conditional correlations are similar, but slightly lower, when we control for parental characteristics, individual characteristics or both set of controls simultaneously (columns b, c, and d respectively).[29]

[28] Notice that the use of two different cross-sections enables the simultaneous inclusion of age and year of birth (since the sample contains individuals born in the same year but of different ages), which is especially useful for the identification strategy discussed in the next section.

[29] We are aware of the fact that the above-mentioned controls are unlikely to represent exogenous covariates. This is because some of the individual characteristics (like place of residence and education) are choice variables, potentially related to the error term of the outcome equation(s). Moreover, parental characteristics, as well as individual place of birth, could reflect unmeasured

Overall, the evidence using observational data seems to be consistent with the theoretical predictions of the model. Nevertheless, these conditional correlations might not represent the causal mechanism portrayed by the theoretical model. First, partner choice/language use and language skills are likely to be correlated with common unobserved factors, opening the door to the typical omitted variable bias. Second, language competence is self-reported, and hence measurement error bias could also be an issue due to the systematic tendency to over-report language skills. Third, we observe language skills only at the time of the interview, but this variable itself is likely to be affected by the linguistic characteristics of the partner. In other words, a native Spanish speaker is likely to improve their Catalan proficiency if matched with a Catalan speaker. This implies that reverse causality might also generate an additional source of inconsistency.

5 Causal evidence: Identification strategy and IV estimates

5.1 Empirical framework

We exploit the change in the language of instruction that took place in Catalan schools after the implementation of the "Language Normalization Act" (LNA) of 1983. Two important remarks are in order. First, oral skills in Catalan improved only for native Spanish speakers, since Catalan was in any case orally transmitted within Catalan-speaking families. Second, exposure to the language-in-education reform depends on the year of birth but also on the number of years of schooling. However, the second variable is endogenous. Therefore, in order to isolate the exogenous component we adopt the strategy followed by Clots-Figueras and Masella (2013), who restricted attention to exposure during compulsory education. They constructed a variable that measures the (potential) number of years of compulsory schooling under the linguistic regime introduced by the 1983 reform, which can be interpreted as an "Intention to Treat" variable.[30] More specifically, Clots-Figueras

parental characteristics that are potentially endogenous with respect to the two outcomes. Therefore, the evidence regarding these control variables must be interpreted with caution and are not discussed in details for brevity resons.

[30] That is, the number of years of schooling in Catalan, assuming: a) no grade repetition, b) perfect compliance with compulsory age of school attendance, and c) uniform use of Catalan as medium of instruction in the schools. The last assumption is the most restrictive, since in the early years of application of the reform, the use of Catalan for general teaching purposes was weaker in schools with a majority of native Spanish speakers. However, the focus of our analysis

and Masella (2013) assumed that individuals born in 1977 or after received all their compulsory schooling in Catalan, while those born between 1970 and 1976 were just partially exposed to the reform, with one year of exposure for the former cohort, up to seven years for the latter cohort. Individuals born before 1970 were never affected. The length of compulsory education in Spain was eight years under the legal framework implemented in 1974 ("Ley General de Educación") from ages 6 to 14. A new law passed in 1990 (LOGSE) extended the number of years of compulsory education to ten (from ages 6 to 16). This means that individuals born before 1983 were subject to eight years of compulsory schooling, and those born in 1983 or after to ten years. [31]

Thus, the variable capturing compulsory exposure to Catalan at school, ce_t, can be expressed in the following way:

$$ce_t = \begin{cases} 10, & \text{if } t \geq 1983 \\ 8, & \text{if } 1977 \leq t < 1983 \\ t - 1969, & \text{if } 1970 \leq t < 1977 \\ 0, & \text{if } t < 1970 \end{cases} \tag{2}$$

Notice that the variation in ce_t is only determined by the individual's year of birth, which is obviously not a choice variable. Indeed, ce_t seems to be an appealing way to extract an exogenous component from the positive trend in oral language skills observed over the successive cohorts of native Spanish speakers. However, this variable itself is unlikely to be a valid exclusion restriction to identify the causal effect of language proficiency on outcomes. In fact, ce_t could capture both the language proficiency effect of the LNA as well as other cohort effects that potentially affect directly the outcomes of interest (i.e., partnership formation and language use), through non-language-related channels.

In order to control for the direct (common) effects of birth cohort on the outcomes of interest, we include native Catalan speakers in the analysis. This is in the spirit of the identification strategy proposed by Bleakley and Chin (2004, 2008 and 2010). They estimate the (private and social) returns to English proficiency among US immigrants, exploiting the well-established fact of the existence of a "critical period" of language acquisition (i.e., immigrants who arrive in the host country at

is precisely the effect of the reform on native Spanish speakers (for whom the treatment was less intense). In this sense, we are probably capturing a lower-bound effect.

[31] The results are unaffected by the change in the length of compulsory education, since we obtained virtually the same results imputing eight years of exposure (instead of ten) also to individuals born after 1982.

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 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, we exploit the fact that oral language skills are also acquired within the family at an early age. Hence, the language-in-education reform did not exert any significant effect on the oral proficiency of native speakers. Moreover, the Spanish skills of native Catalan speakers have remained very high and stable over cohorts.

Therefore, using the pooled sample of native Spanish speakers and native Catalan speakers, we use the interaction between exposure to Catalan during compulsory schooling (ce_t) and the indicator that identifies native Spanish speakers as an exclusion restriction, controlling for (common) cohort effects in the outcomes of interest. The underlying assumption of this identification strategy is that both language communities were subject to the same general cohort effects, except that we allow the treatment (compulsory policy exposure) to affect (with increasing intensity) the oral proficiency in Catalan of the treated cohorts of native Spanish speakers. In other words, we assume that any specific cohort effect experienced by native Spanish speakers affected by the policy change should be (plausibly) attributed to better language skills.

This identification setup can be easily represented by a two-equation system, where the oral skills in Catalan (Cat) of individual i, born in cohort t and a native speaker of l ($l = Spanish, Catalan$) is the dependent variable of the first-stage equation, which contains as right-hand-side variables a set of controls (X), year of birth fixed-effects (φ_t), an indicator for native Spanish speaker ($l = Spanish$), and its interaction with ce_t (as identifying variable):

$$Cat_{itl} = \mu + \lambda' X_i + \rho I\left(l = Spanish\right) + \gamma I\left(l = Spanish\right) \times ce_t + \varphi_t + u_{itl} \qquad (3)$$

The second-stage equation explains the two outcomes of interest (having a Catalan-speaking partner and use of Catalan with the partner). Alternatively, we could define the first outcome as having a mixed-couple and the second outcome as speaking the non-native language with the partner. Such a symmetric treatment of the two speech communities seems desirable. Unfortunately, the data do not support a symmetric approach. The problem is that the survey reports more information about the respondent than about the partner. If the respondent is a native Spanish speaker then we know his/her Catalan proficiency and year of birth (so that we can impute years of exposure to the reform). However, if the respondent is a native Catalan speaker then we ignore his/her partner's Catalan proficiency and year of birth. Thus, we need to define the first outcome as having a Catalan-speaking partner and the second outcome as the use of Catalan with the partner. The second-stage equation includes proficiency in oral Catalan as an endogenously determined covariate:

$$Y_{itl} = \alpha + \beta' X_i + \pi I\left(l = Spanish\right) + \delta_{IV} Cat_{itl} + \theta_t + \varepsilon_{itl} \tag{4}$$

Under the validity of the identifying assumption, the 2SLS estimation of Equations (3) and (4) should provide the causal effect of oral fluency in Catalan on each of the outcomes (δ_{IV}) among native Spanish speakers who improved their language proficiency due to exposure to the language in their compulsory schooling. This is because 2SLS provides an estimate of the endogenous right-hand-side variable that exploits only the variability of language skills that is produced by the instrument among the subpopulation of compliers (i.e., a "local" estimate of the treatment effect).[32]

5.2 Estimation results

Selected 2SLS estimates of Equations (3) and (4), estimated with the pooled sample of Spanish and Catalan speakers[33], are displayed in Table 4. Overall, the results obtained from our identification strategy are in line with those obtained by OLS and, more importantly, consistent with the theoretical predictions. More specifically, the causal effect of better Catalan skills among Spanish speakers on the probability of having a Catalan-speaking partner is just slightly higher (but not statistically different) than the OLS estimate. Using the parsimonious set of controls, a unit

[32] In the empirical analysis, we cluster the standar errors on year of birth, which is the level of variation of our instrument.

[33] The results obtained by applying OLS to the subsample of native Spanish speakers are virtually identical to those obtained from the pooled sample of both Spanish and native Catalan speakers, as shown in Table A2 in the online Appendix. This means that most of the conditional correlations between oral proficiency in Catalan and the two outcomes are driven by the variation observed within the Spanish speaking community.

increase in fluency in oral Catalan increases the likelihood of a mixed match by 7.6 percentage points (versus an OLS estimate of 4.5 percentage points for the joint sample -See Table A2). In order to gauge the magnitude of the effect, we must note that, according to the first-stage regression, the Catalan proficiency of native Spanish speakers fully affected by the reform is approximately one point higher (on a 0-10 scale) than that of those not exposed to the reform (8.5 versus 7.5, respectively). Also, the 2SLS estimates indicate that such an increase in the level of proficiency raises the probability that a native Spanish speaker is matched with a native Catalan speaker from 0.33 to 0.40. This is a sizable effect. The two speech communities have a similar size, which implies that in the absence of any language-related bias such a probability would be approximately 0.50. Hence, according to our estimates, the reform has eliminated roughly 40% of the initial bias.

As we add parental controls, the point estimate drops slightly. However, in contrast to the OLS strategy, including individual controls generates a modest increase in the coefficient of interest, while controlling for both parental and individual characteristics provides virtually the same estimate as in the baseline specification.

Regarding the second outcome (the use of Catalan with the partner), our IV approach generates estimates that are much more similar to those obtained by OLS. In particular, for the baseline specification (column (a)), one unit increase in fluency in oral Catalan increases the probability of speaking only Catalan with the partner by 5.3 percentage points, slightly above the OLS estimates of 4.3 percentage points -See Table A2. The effect of including parental and individual covariates on the second outcome are analogous to the first outcome case, and hence the results of the baseline specification appear very robust. Overall, the differences between the OLS and 2SLS estimates could be due to the fact that the latter estimator exploits all the variation that is observed in the data, whereas the former is based only on the variation generated by the instrument among the treated cohort of the subsample of native Spanish speakers. Moreover, the presence of measurement error in self-reported language proficiency, which could cause a downward bias in the OLS estimate, could be an additional (and probably complementary) explanation for this divergence. It is important to note that the first-stage estimates corresponding to our identifying variable (the interaction between language exposure during compulsory schooling and the indicator for being a native Spanish speaker), presented in the upper panel of Table 4, have the expected sign and are strongly significant (the complete results of the first-stage regressions can be found in Table A3 in the online Appendix). Thus, native Spanish speakers affected by the language reform did improve their oral proficiency in Catalan. The corresponding coefficients obtained using different specifications are quite stable. Moreover, the F test for weak identification indicates that the instrument is sufficiently strong in all specifications. Overall, the results obtained from the IV strategy provide empirical support for the causal predictions of the theoretical model. Thus, better proficiency in the weak language of native speakers of the strong language (generated by a plausibly exogenous source of variation) fosters their propensity to form mixed partnerships and use the weak language more intensively.

We have performed a battery of robustness checks about the specification of our baseline regression. In the online appendix, we present in detail the results of our sensitivity analysis. In particular, we show that the estimates of interest are quite stable when we run separate estimations for males and females (Table A4), use alternative specifications of the age polynomial (Table A5), use an alternative specifications of the exposure variable (Table A6). Moreover, we also display the results obtained by dropping cases of mixed languages for individuals and their partners (i.e. Catalan and Spanish), exclude individuals who do not have a partner at the time of the survey and focus on respondents who live with their partner (columns (a)-(e) of Table 7A in the online Appendix).

One of the sensitivity checks is very relevant to discuss the role of individual or group identity in explaining our baseline results. If we exclude from the sample those respondents with a language of self-identification different from their native language ("language switchers") then the main results remain basically unchanged. It is important to note that the vast majority of language switchers are Spanish native speakers who chose Catalan as their language of self-identification. As reported in column (b) of Table 5, the effect of oral skills in Catalan on partnership formation is slightly smaller, and the effect on the use of Catalan slightly higher when we exclude switchers. Moreover, the instrument becomes much stronger and the coefficients are estimated more precisely. In Section 6 we further discuss how this exercise helps interpreting the nature of the main results. Here it seems worth highlighting that the stability of the estimates obtained after dropping language switchers represents a first evidence in favor of our identification strategy. In fact, it can be argued that ethnic or political identity (Catalan or Spanish) could be a possible unobserved determinant of partnership formation and, as reported by Aspachs et al. (2008) and Clots-Figueras and Masella (2013), also affected by the reform. In other words, according to such alternative theory, some native Spanish speakers might have adopted, as a result of exposure to the reform, a Catalan identity, and this would help them in finding a Catalan-speaking partner. However, if identity was the main driving force behind our results, and as long as the language of self-identification is positively correlated with ethnic or political identity (a very plausible hypothesis), then we would expect a significant change in the main coefficients when language switchers are excluded. Instead, the observed invariance of the coefficients suggests that the exclusion restriction is not picking up unobserved identity traits that affect the potential to find a Catalan-speaking partner.

On top of these sensitivity analysis, in the next section we provide more detailed evidence on some key robustness checks concerning the two components of our identifying variable (the interaction between compulsory language exposure and the native language indicator) and the underlying identifying assumptions of our identification strategy. First, we present the results from several placebo experiments, which aim at providing evidence that our treatment variable (compulsory exposure) is not capturing any spurious effects due to pre-existing trends across cohorts. Second, we repeat the estimations using two alternative proxies of native language, namely parental language and parental regional origins, in order to ensure that our results are robust to the potential endogeneity of self-reported native language. Third, thanks to the availability of these two proxies for native language, we are able to (partially) relax the underlying hypothesis of our identification strategy, requiring that the non-linguistic effects that operate across cohorts are common for both language groups.

5.3 Falsification and identification checks

Evidence from placebo experiments. One component of the identifying variable, exposure to Catalan during compulsory schooling, only depends on the year of birth. We need to consider the possibility that compulsory exposure could capture spurious relations due to potential cohort-specific trends in (language-related) couple formation and/or language use. We have run a set of placebo experiments, which aim at providing evidence that our identifying variable is not contaminated by any spurious effects.

We consider the reduced form equation to test for falsification. Equation (5) shows the reduced form representation of our baseline 2SLS approach,

$$Y_{itl} = \alpha + \beta' X_i + \pi I\left(l = Spanish\right) + \delta_{RF} I\left(l = Spanish\right) \times ce_t + \theta_t + \varepsilon_{itl} \qquad (5)$$

where δ_{RF} is the coefficient that "directly" relates exposure to Catalan during compulsory schooling among native Spanish speakers with the outcomes of interest.

Then, we consider the placebo sample of never-treated individuals born between 1944 and 1969 who were schooled in Catalonia before the reform was implemented (i.e., they were never exposed to Catalan during compulsory schooling). Therefore, also in line with the falsification strategy adopted by Clots-Figueras and Masella (2013), we impute years of (pseudo) exposure to Catalan at school (ce_t^*), which are imputed "as if" the reform had been applied from 13 to 20 years before the true reform; that is, first in 1970 instead of 1983, then in 1969 and so forth (until 1963).[34] We estimate the reduced form model (5), but using the placebo sample of individuals born in Catalonia (or migrated from the rest of Spain, before age 6) who were never affected by the compulsory component of the reform:

$$Y_{itl} = \alpha + \beta' X_i + \pi I\left(l = Spanish\right) + \eta I\left(l = Spanish\right) \times ce_t^* + \theta_t + \varepsilon_{itl} \qquad (6)$$

Obtaining a positive and significant coefficient for placebo exposure would cast doubt on the reliability of our (real) exposure variable, because it could be reflecting pre-existing cohort trends that apply to the outcomes of interest. However, the battery of falsification experiments we performed suggest that this is not the case. In fact, while the reduced form estimates based on real reform exposure reflect a positive causal effect of our identifying variable on both outcomes (see the first column of Table 6a and 6b, respectively), all the coefficients associated with the different placebo exposure variables are small in size and not statistically different from zero. Overall, this evidence suggests that the compulsory exposure variable constructed à la Clots-Figueres and Masella (2013) is unlikely to be capturing spurious relations, unrelated to the policy reform, as also highlighted in their original paper.

Native languages. We also address the validity of the second component of the identifying variable: the definition of native Spanish speakers. It could be argued

[34] These boundaries have been chosen in order to keep a minimum number of observations in the "pseudo" control group individuals not exposed to the fake reforms. Moreover, we are unable to consider individuals born before 1944, since we do not dispose of exact information about year of birth for older cohorts.

that the self-reported native language might not be exogenous; respondents could be influenced by endogenous factors. In particular, some Spanish speakers might be tempted to misreport their true native language in favor of Catalan (or Spanish and Catalan), perhaps because of the influence of the language-in-education reform on their self-identification. In order to address these concerns, we have replaced the native language variable used in the baseline estimations by two alternative proxies. In particular, an individual is classified as a native Spanish speaker: (i) if both parents have Spanish-only as native language (parental language) or, alternatively, ii) if both parents were born outside Catalonia (parental origins). We then re-estimated our 2SLS model using these two alternative definitions of language groups. The results obtained for each of the two proxies of native language are presented in column (a) of Tables 7a and 7b, respectively. These estimates are generally similar than those obtained using the original native language variable. We only observe a mild reduction in the coefficient of Catalan skills on the partner's language equation when individuals are classified into language groups by parental language, and somewhat higher coefficients for both outcomes when the groups are formed by parental origins.[35]

This evidence indicates that the main results are robust to the use of alternative proxies of native language. Moreover, the fact that the estimates obtained using parental origins as proxy for native language are higher than in the baseline estimation is consistent with the idea that the sub-population of compliers that is captured by this new instrument are individuals affected by the reform with both parents born outside Catalonia, who are likely to be more sensitive to exposure to Catalan at school. In other words, native Spanish speakers with at least one parent born in Catalonia were probably exposed to Catalan through alternative channels, and hence were less sensitive to the reform than their counterparts with both parents born outside Catalonia.[36]

The availability of two alternative proxies to define language groups opens the possibility of analyzing the sensitivity of the results to the main identifying assump-

[35] Notice that using parental language as a proxy for native language creates some ambiguity in the (few) cases in which the individual declares that both parents had both Catalan and Spanish as their native languages. However, the results are virtually the same when these observations are excluded (detailed results available upon request).

[36] Nevertheless, defining language groups on the basis of parental origins is not ideal for the purpose of testing the predictions of the theoretical model (which is structured around the concept of native language), since there is a relevant fraction of individuals with Catalan origins (i.e. at least one parent born in Catalonia) who are native Spanish speakers (around 20%).

tions in our model. First, we were able to specify two alternative overidentified 2SLS models, in which we use exposure to Catalan interacted with both the native language indicator and each of the two alternative proxies as exclusion restrictions. The results obtained from the overidentified models are presented in column (b) of Tables 7a and 7b for Spanish speaking parents and parents of non-Catalan origin, respectively. In both cases, the point estimates of interest are very similar to those obtained from the baseline specification. More importantly, the Hansen J test for overidentification does not reject the null hypothesis that the exclusion restrictions can be reasonably excluded from the outcome equation(s). This result points out that the instrument seems to be uncorrelated with unobservable determinants of partnership formation and language use (i.e. identity feelings, aspirations, social networks, etc.). Second, we are also able to perform an additional (and related) exercise. We relax the hypothesis that the only channel through which exposure to Catalan during compulsory schooling of native Spanish speakers affects the outcomes is through language proficiency, by including the interaction between language exposure and each of these two proxies as a control in the outcome equations (column (c) of Tables 7a and 7b). In this case, we obtain higher point estimates for Catalan skills when we consider the first proxy, which also lose precision (and strength of the instrument) due to the correlation between the exclusion restriction and these control variables. When we instead control for parental origins interacted with exposure to Catalan, the coefficient of Catalan proficiency for the partner's language equation is virtually identical to the baseline (but again imprecisely estimated), while it becomes smaller for the language use equation. In any case, the coefficients for the interaction between exposure to compulsory schooling and the two alternative proxies for language groups is not statistically significant and very small in size (which is consistent with the evidence from the overidentification test).

Common non-language effects. We have also tried to relax the assumption that the direct cohort effects in the two outcomes are common to native Spanish speakers and native Catalan speakers, which is a non-trivial underlying hypothesis of our identification strategy. We allow for language-specific cohort effects by including interactions between year of birth and indicators of the above language group proxies. This should capture potentially heterogeneous cohort effects on each of the two outcomes. Therefore, the 2SLS equations become

$$Cat_{itl} = \mu + \lambda' X_i + \rho I\left(l = Spanish\right) + \gamma I\left(l = Spanish\right) \times ce_t + \varphi_{l^* t} + u_{itl} \quad (7)$$

$$Y_{itl} = \alpha + \beta' X_i + \pi I\left(l = Spanish\right) + \delta_{IV} Cat_{itl} + \theta_{l^* t} + \varepsilon_{it} \tag{8}$$

where l^* is one of the two proxies of native language, and the terms $\varphi_{l^* t}$ and $\theta_{l^* t}$ represent birth-cohort fixed effects that are allowed to differ by either parental language or parental origins. The corresponding estimates are presented in column (d) of Tables 7a and 7b, respectively, and show the same pattern that emerged from the models that contain the interactions between exposure and language proxy as controls. That is, the coefficients for Catalan skills are somewhat higher (and imprecisely estimated) when parental language is considered as a proxy, while controlling for parental origin-specific year of birth effects yields the same point estimate for Catalan proficiency on partnership formation and a small and insignificant coefficient for the language use equation.

Subsample of native Spanish speakers. As a final exercise, we repeat the 2SLS estimation for the subsample of native Spanish speakers using the same specification as our baseline model, but using the interaction between parental origins and exposure to Catalan as an exclusion restriction.[37]

We estimate the model(s) for the whole sample of native Spanish speakers and also excluding individuals whose partner has both Catalan and Spanish as a native language. These results are displayed in columns (a) and (b) of Table 8**.** They are qualitatively similar to those obtained from the whole sample, which exploits all the variation among Spanish speakers to identify the causal effects, while here the estimates reflect the variation among Spanish speakers with non-Catalan origins who improved their oral fluency in Catalan due to language exposure during compulsory education. Nevertheless, the estimations are less precise and the identification is somewhat weak, but still the results are in line with the evidence presented using the simple OLS.

6 Discussion and concluding remarks

We have presented empirical evidence and theoretical arguments that endorse the idea that languages are much more than neutral communication devices, due to the plausible existence of some form of emotional attachment. However, one may claim

[37] The heterogeneous effect of exposure to Catalan by parental language cannot be used as an exclusion restriction, since virtually all Spanish speakers have both parents who have only Spanish as native language.

that our results could also be compatible with alternative, plausible interpretations. Let us consider the following three alternatives:

Alternative 1: Results are driven by a combination of social mobility and assortative matching.

A large fraction of native Spanish speakers either migrated from the South of Spain the 1960's or are their descendants. Thus, native Catalan speakers have enjoyed in average a better socio-economic status. Some of these immigrants or their children have climbed the social ladder, which may have raised their propensity to match with members of these upper social groups, which in turn are more likely to speak Catalan. Finally, native Spanish speakers may more inclined to learn and use Catalan as they improve their socio-economic status, perhaps using the language as a signaling device.

Some of the control variables we use in the estimation actually reflect the socio-economic status of individuals or their families: education of the respondent, parental education, and even the place of birth or residence of the respondent and their families. Hence, if such alternative interpretation had a bite, the introduction of these control variables should affect the point estimates of the effect of language skills on both outcomes. Since this is not the case, and the main estimates are observed to be very stable to the inclusion of various sets of controls, we find little support for such an alternative interpretation.

Alternative 2: Results are driven by changes in ethnic or political identity.

It is well known that language is a key symbol of ethnic, national, or class identity. For the case of Catalonia, the American antropologist Kathryne Woolard (Woolard, 1989; Woolard and Ghang, 1990) pointed out that back in the 1980's ethnicity was critical to understanding language attitudes and choices. More specifically, she found that Catalan was perceived by non-Catalan speakers as the language of native Catalans and completely alien to everybody else. Moreover, the adoption of Catalan was interpreted as sheer assimilation. In contrast, Spanish was perceived by almost everyone as "the language of everybody", free of ethnic marks. Thus, one may wonder if our results may simply reflect the dynamics of ethnic politics in Catalonia. In particular, the educational reform may have affected the frequency of mixed couples (according to our definition) not so much by changing language skills and reducing the language conflict, but by inducing a fraction of native Spanish speakers to cross over and become “ethnically Catalan” (that is, by assimilation). In other words, it could be the case that endogamy has remained roughly unchanged, but the composition of ethnic groups has varied over time.

Our data set allows us to tentatively approach the issue of ethnic identity. In particular, we believe that ethnic or cultural assimilation should show up in those respondents who choose a language of self-identification different from their native language. That is, if an "ethnically Spaniard" (a native Spanish speaker) crosses over and becomes "ethnically Catalan", then such a switch should probably involve adopting Catalan as the language of self-identification. In fact, in our baseline sample, whereas only about 3% of the native Catalan speakers report Spanish as their language of self-identification, about 20% of native Spanish speakers report Catalan as their language of self-identification. When we eliminate these “switchers” from the sample, results remain largely unchanged (see Section 5.2 and Table 5). This can be taken as a informal test for the role played by ethnic identity formation in driving our results. Indeed, this suggestive evidence points out that language skills matter beyond ethnic identity. In particular, native Spanish speakers that keep Spanish as their language of self-identification, to the extent they improved their Catalan skills during compulsory education, are more likely to find Catalan speaking partners and use Catalan with their partner more often. This interpretation seems compatible with the latest research on language attitudes in Catalonia (Woolard, 2011 and 2008; and Newman, Trenchs-Parera and Ng, 2008). These studies suggests that the perceived link between language and ethnicity has drastically softened and that nowadays both speech communities value bilingual proficiency.

Alternative 3: Our instrumental variable may capture spurious relations due to potential cohort-specific trends in couple formation.

The historical period under consideration is highly non-stationary in many dimensions. First, it includes two antithetical political regimes (dictatorship and democracy). Second, it has witnessed huge demographic changes; specially, the huge migration inflows of the 1960’s and 1970’s. Third, the new information and communication technologies, specially during the last decades, might have affected social behavior, particularly in the marriage market. Thus, there may exist underlying trends in couple formation that can be captured by our instrumental variable.

The evidence from our placebo experiments reported in Section 5.2 suggests that the compulsory exposure variable is unlikely to be capturing spurious relations, unrelated to the policy reform.

Summarizing, in this paper we examine the non-communicative aspects of languages both theoretically and empirically. In particular, this is the first work showing that policies that promote the acquisition of language skills that appear redundant from a communicative viewpoint can significantly reduce segregation along linguistic lines. We have interpreted these results using an abstract and comprehensive notion of linguistic preferences, which is far more general than the presumed link between language and ethnic identity. We are also confident that (at least part of) the increase in the frequency of mixed couples can indeed be interpreted as a reduction in segregation.

Our empirical analysis focuses on the case of couple formation in Catalonia. Obviously, more research is needed before we can claim that linguistic preferences are relevant in most bilingual societies and in other types of social interactions. Nevertheless, we would like to comment briefly on the possible external validity of our results. Casual observation indicates that linguistic preferences do seem to be present in a wide range of bilingual societies.[38] Thus, in this respect we believe there is nothing special about Catalonia. However, we are much less convinced that our results about couple formation can be extrapolated to all types of social and economic interactions. It may well be the case that individuals (at least those who are endowed with a sufficiently broad language repertoire) are less concerned about the language of use when, for example, making a transaction on the Internet than when searching for a partner. We would not be surprised if future research finds large variations in the relevance of linguistic preferences across different types of social interactions.

7 References

Alesina, A, A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003), Fractionalization. *Journal of Economic Growth* 8, 155-194.

Artigal, J.M. (1997), The Catalan Immersion Program. In R.K. Johnson and M. Swain (editors), *Immersion Education: International Perspectives.* Cambridge University Press.

Aspachs, O., I. Clots-Figueres, J. Costa-Font, and P. Masella (2008), Compul-

[38] In particular, Villancourt (1984) report that Francophones in Quebec, even those with an excellent knowledge of English, prefer being served in French in establishments like stores and restaurants.

sory Language Educational Policies and Identity Formation. *Journal of the European Economic Association* 6(2-3): 434-444.

Bleakley, H., and A. Chin (2010), Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants. *American Economic Journal: Applied Economics* 2(1), 165-92.

Bleakley, H., and A. Chin (2008), What Holds Back the Second Generation? *Journal of Human Resources* 43(2), 267-298.

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

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

Caminal, R. and A. Di Paolo (2015), Your Language or Mine?. Barcelona GSE working paper 852.

Chen, M. K. (2013), The Effect of Language on Economic Behavior: Evidence on Savings Rates, Health Behaviors, and Retirement Assets. *American Economic Review* 103, 690-731.

Chiswick, B., and C. Houseworth (2011), Ethnic intermerriage among immigrants: Human capital and assortative mating. *Review of Economics of the Household* 9, 149-180.

Chiswick, B., and P. Miller (2007), *The Economics of Language: International Analyses.* London: Routledge.

Church, J., and I. King (1993), Bilingualism and network externalities. *Canadian Journal of Economics* 26, 337-345.

Clots-Figueras, I., and P. Masella (2013), Education, Language and Identity. *The Economic Journal* 123 (570), F332-F357.

Consell Superior d'Avaluació del Sistema Educatiu (2013), Sistema d'indicadors d'ensenyament de Catalunya, No. 17, Generalitat de Catalunya.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2012), The political economy of linguistic cleavages. *Journal of Development Economics* 97, 322-338.

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

Furtado, D., and N. Theodoropoulos (2011), Interethnic marriage: A choice between ethnic and education similarities. *Journal of Population Economics* 24, 1257-1279.

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

Galor, O., O Özak, and A. Sarid (201), Geographical Origins and Economic Consequences of Language Structures. SSRN 2820889.

Gay, V., E. Santacreu-Vasut, and A. Shoham (2013), The Grammatical Origins of Gender Roles, BEHL wp 2013-03

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

Grin, F. (1992), Towards a Threshold Theory of Minority Language Survival. *Kyklos* 45, 66-97.

Instituto de Evaluación (2011), Evaluación General de Diagnóstico 2010: Educación Secundaria Obligatoria, Segundo Curso. Informe de Resultados. Ministerio de Educación.

Jones, E. (2000), The Case for a Shared World Language. In: M. Casson and A. Goldley (eds.), *Cultural Factors in Economic Growth*, pp. 210-235, Berlin: Springer-Verlag.

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

Mèlitz, J. (2012), A Framework for Analyzing Language and Welfare. SIRE-DP-2012-89.

Meng, X., and D. Meurs (2009), Intermarriage, Language, and Economic Assimilation Process: A Case Study of France. *International Journal of Manpower* 30, 127-144.

Muñoz, C. (2005), Trilingualism in the Catalan educational system. *International Journal of the Sociology of Languages* 171, 75-93.

Myerson, R, and A. Satterwhaite (1983), Efficient Mechanisms for Bilateral Trading. *Journal of Economic Theory* 29, 265-281.

Newman, M., M. Trenchs-Parera, and S. Ng (2008), Normalizing bilingualism: The effects of the Catalonian linguistic normalization policy one generation after. *Journal of Sociolinguistics* 12(3), 306-333.

Nuhoğlu Soysal, Y., and D. Strang (1989), Construction of the First Mass Education Systems in Nineteenth-Century Europe. *Sociology of Education* 62(4), 277-288.

Selten, R., and J. Pool (1991), The distribution of foreign language skills as a game equilibrium, in R. Selten (ed.) *Game Equilibrium Models* vol. 4, pp. 64-84, Berlin: Springer-Verlag.

Siguan, M. (1991), The Catalan Language in the Educational System of Catalonia. *International Review of Education* 37 (1), 87-98.

Tucker, G.R. (2001), A Global Perspective on Bilingualism and Bilingual Education. In J.E. Alatis and A.-H. Tan (eds.), *Roundtable on Language and Linguistics.* Washington DC: Georgetown University Press.

Van Parijs, P. (2011), *Linguistic Justic for Europe and for the world.* Oxford: Oxford University Press.

Vila, F.X. (2008), Language-in-education Policies in the Catalan Language Area. *AILA Review* 21, 31-48.

Villancourt, F. (1984), The Choice of Language of Consumption. Cahier 8204, Université de Montreal.

Weber, S., J. Gabszewicz, and V. Ginsburg. (2011), Bilingualism and Communicative Benefitts. *Annals of Economics and Statistics / Annales d'Économie et de Statistique* 101-102, 271-286.

Wickström, B.-A. (2005), Can Bilingualism be Dynamically Stable? A Simple Model of Language Choice. *Rationality and Society* 17(1), 81-115.

Woolard, K.A. (2011), Is there linguistic life after high school? Longitudinal changes in the bilingual repertoire in metropolitan Barcelona. *Language and Society* 40(5), 617-648.

Woolard, K.A. (2008), Language and Identity Choice in Catalonia: The Interplay of Contrasting Ideologies of Linguistic Authority. In K. Süselbeck, U. Mühlschledgel, P. Masson (eds), *Lengua, nación e identidad. La regulación del plurinlingüismo en España y América Latina.* Madrid: Iberoamericana, 303-323.

Woolard, K.A. (1989), *Double Talk: Bilingualism and the Politics of Ethnicity in Catalonia.* Stanford University Press.

Woolard, K.A., and T.-.J. Gahng (1990), Changing Language Policies and Attitutes in Autonomous Catalonia. *Language and Society* 19 (3), 311-330.

8 Appendix

Result 1. Note that the frequencies of A, B, and N are given, respectively, by

$$\Pr(A) = F(\eta_a) + \int_{\eta_a}^{g} [1 - F(w_b - \eta_a)]\, dF(w_b)$$

$$\Pr(B) = \int_{0}^{g-\eta_a} [1 - F(w_a + \eta_a)]\, dF(w_a)$$
$$\Pr(N) = [1 - F(g)]\,[1 - F(g - \eta_a)]$$

Hence, A and N increase and B decreases with η_a.

Result 2. The expected utilities of those individuals in potential partnerships with $\eta_a < g$ are given by

$$U_a = [\Pr(A) + \Pr(B)]\, g - \int_{0}^{g-\eta_a} [1 - F(w_a + \eta_a)]\, w_a dF(w_a)$$

$$U_b = [\Pr(A) + \Pr(B)]\, g - \int_{0}^{\eta_a} w_b dF(w_b) - \int_{\eta_a}^{g} [1 - F(w_b - \eta_a)]\, dF(w_b)$$

The effect of η_a on U_a has an ambiguous sign:

$$\frac{dU_a}{d\eta_a} = -[1 - F(g)]\, f(g - \eta_a)\, g + \int_{0}^{g-\eta_a} f(w_a + \eta_a)\, w_a dF(w_a) +$$
$$+ [1 - F(g)]\,(g - \eta_a)\, f(g - \eta_a)$$

However, both U_b and $U_a + U_b$ decrease with η_a.

Figure1a: Equilibrium outcomes

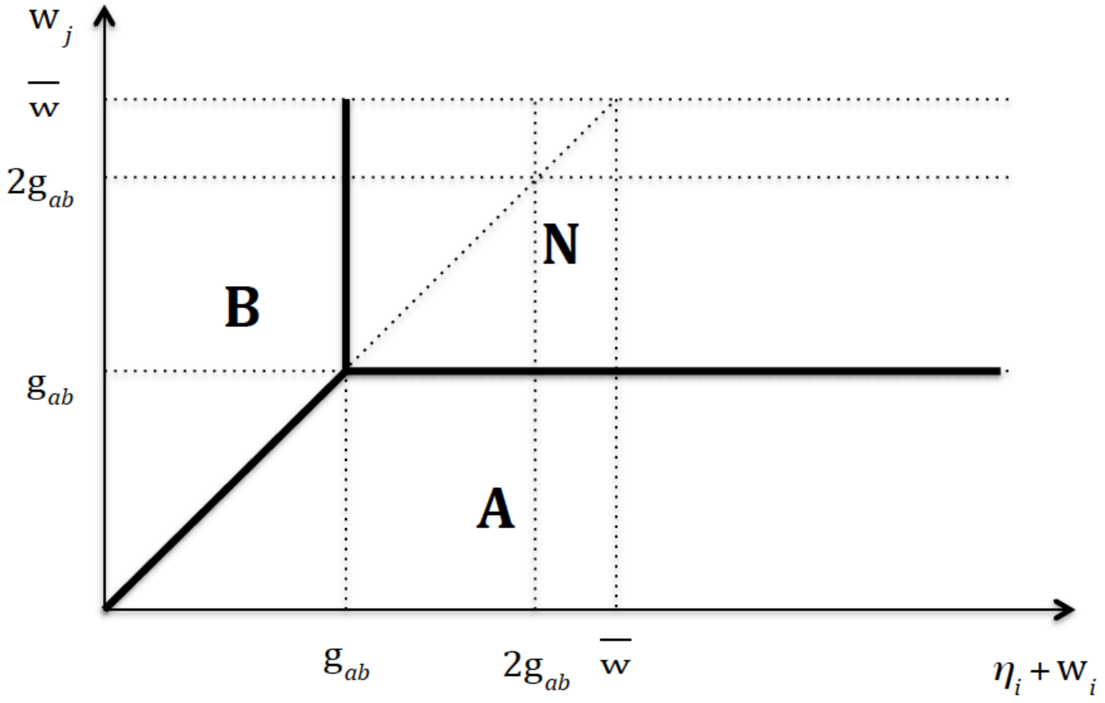

Figure1b: Efficient outcomes

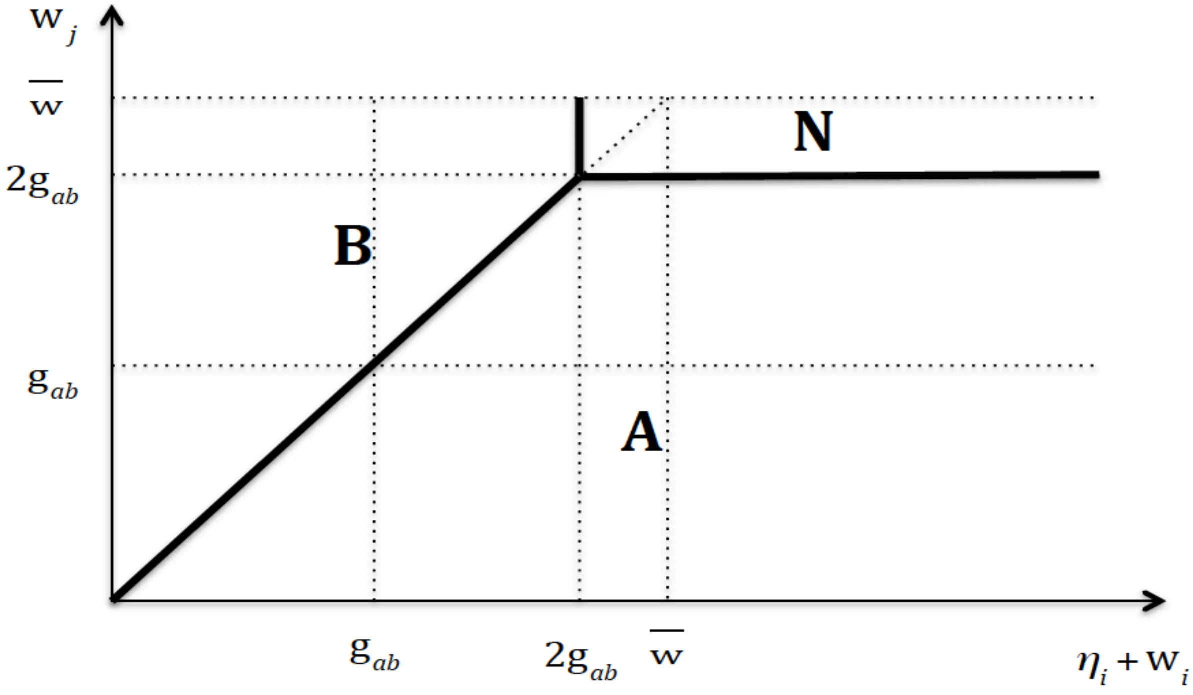

Figure 2a: Average Oral Proficiency in Catalan **Figure 2a: Average Oral Proficiency in Spanish**

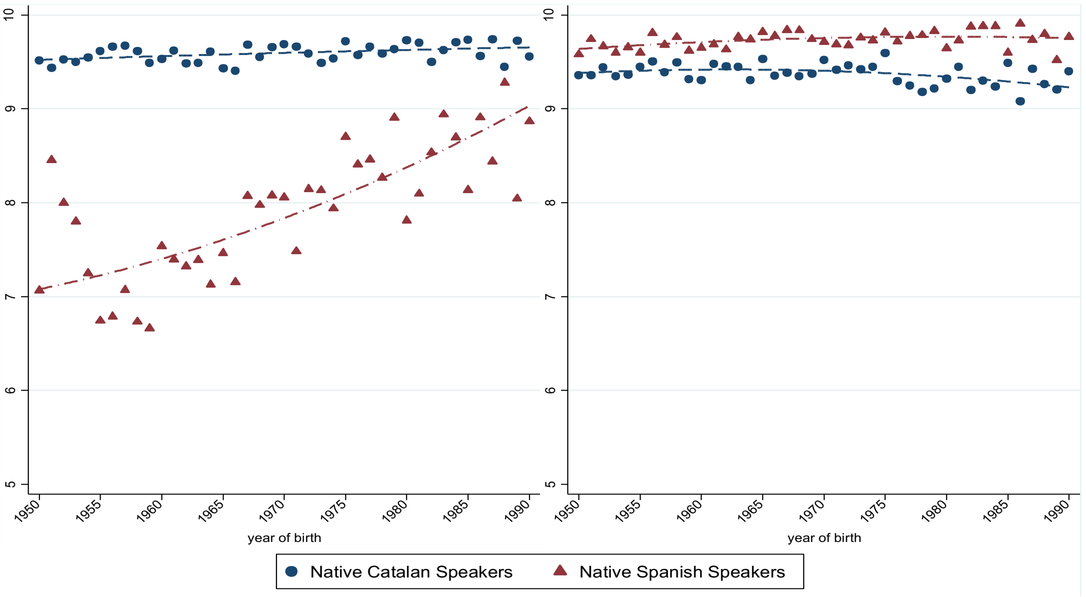

Figure 3a: Average Written Proficiency in Catalan **Figure 3b: Average Written Proficiency in Spanish**

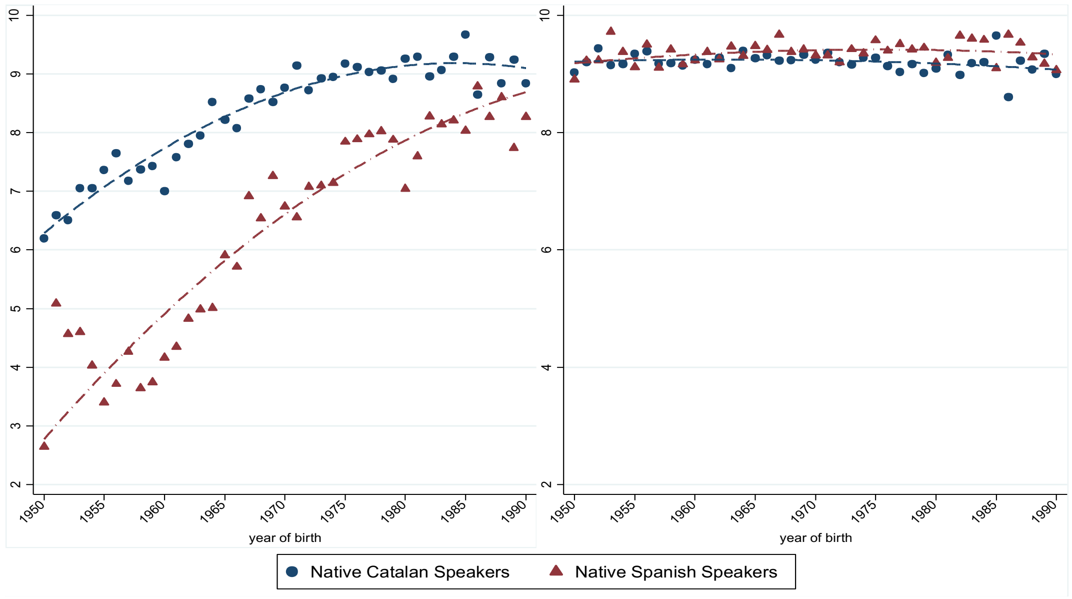

Table 1: Partner's Language and Language Use by Native Language

	% individuals with Catalan-speaking partners		% using only Catalan with the partner	
	unconditional	Proficiency in Catalan ≥ 8	unconditional	Proficiency in Catalan ≥ 8
Catalan native speakers	75.65	76.14	77.2	77.82
Spanish native speakers	35.77	42.63	15.73	22.22

Table 2: Descriptive Statistics by Language Groups

	joint sample		**native Catalan Speakers**		**native Spanish speakers**	
	mean	s.d.	mean	s.d.	mean	s.d.
partner's language = Catalan-only	0.578	0.494	0.757	0.429	0.358	0.479
language used with the partner = Catalan-only	0.497	0.500	0.772	0.420	0.157	0.364
Spanish native speaker (l = Spanish)	0.447	0.497	--	--	--	--
oral Proficiency in Catalan (Cat)	8.825	2.027	9.589	0.870	7.881	2.577
years compulsory education in Catalan (ce_l)	3.162	3.849	3.023	3.845	3.333	3.847
wave 2013	0.523	0.499	0.522	0.500	0.525	0.499
age	41.69	10.402	42.45	10.722	40.76	9.915
male	0.487	0.500	0.499	0.500	0.473	0.499
father place of birth = Barcelona	0.030	0.171	0.016	0.125	0.048	0.214
Girona	0.220	0.414	0.287	0.452	0.138	0.345
Tarragona	0.064	0.244	0.110	0.313	0.006	0.079
Southern Catalonia (Terres de l'Ebre)	0.041	0.198	0.066	0.247	0.010	0.100
Western Catalonia (Ponent)	0.056	0.230	0.095	0.293	0.008	0.086
Central Catalonia	0.071	0.257	0.120	0.324	0.011	0.104
Pyrenees and Aran Valley	0.056	0.230	0.093	0.290	0.010	0.102
Balearic Islands and Valencia	0.035	0.185	0.061	0.239	0.004	0.061
Basque Country and Galicia	0.009	0.092	0.008	0.090	0.009	0.095
other Spanish regions	0.018	0.131	0.006	0.080	0.031	0.174
other places	0.389	0.487	0.129	0.336	0.709	0.454
miss father's place of birth	0.012	0.110	0.009	0.097	0.016	0.125
mother place of birth = Barcelona	0.007	0.086	0.008	0.090	0.007	0.081
Girona	0.234	0.423	0.305	0.460	0.146	0.353
Tarragona	0.066	0.249	0.112	0.316	0.010	0.100
Southern Catalonia (Terres de l'Ebre)	0.044	0.205	0.071	0.257	0.010	0.100
Western Catalonia (Ponent)	0.058	0.234	0.100	0.300	0.006	0.076
Central Catalonia	0.068	0.251	0.113	0.316	0.012	0.109
Pyrenees and Aran Valley	0.058	0.235	0.095	0.293	0.013	0.115
Balearic Islands and Valencia	0.033	0.179	0.057	0.231	0.004	0.064
Basque Country and Galicia	0.019	0.136	0.019	0.135	0.019	0.137
other Spanish regions	0.017	0.129	0.007	0.082	0.029	0.168
other places	0.388	0.487	0.108	0.310	0.735	0.441
miss father's place of birth	0.007	0.084	0.006	0.080	0.008	0.089
Parents' native language = both Spanish	0.435	0.496	0.067	0.249	0.890	0.313
Catalan native language of father or mother	0.157	0.364	0.207	0.405	0.095	0.293
Catalan native language of father and mother	0.408	0.492	0.726	0.446	0.015	0.122
missing parents' native language	0.004	0.067	0.003	0.055	0.006	0.079
highest parental education = no education	0.029	0.168	0.027	0.162	0.031	0.174
primary	0.185	0.388	0.122	0.328	0.263	0.440
secondary	0.495	0.500	0.500	0.500	0.489	0.500
tertiary	0.201	0.401	0.238	0.426	0.155	0.362
missing parental education	0.090	0.286	0.113	0.316	0.062	0.242
number of observations	5357		2961		2396	

Table 2 (continued): Descriptive Statistics by Language Groups

	joint sample		**native Catalan Speakers**		**native Spanish speakers**	
	mean	s.d.	mean	s.d.	mean	s.d.
individual's place of birth = Barcelona	0.503	0.500	0.402	0.490	0.628	0.484
Girona	0.085	0.279	0.113	0.317	0.050	0.218
Tarragona	0.065	0.246	0.070	0.256	0.058	0.233
Southern Catalonia (Terres de l'Ebre)	0.065	0.246	0.108	0.311	0.011	0.104
Western Catalonia (Ponent)	0.097	0.296	0.129	0.336	0.056	0.231
Central Catalonia	0.082	0.274	0.104	0.305	0.054	0.226
Pyrenees and Aran Valley	0.041	0.199	0.065	0.247	0.012	0.109
Balearic Islands and Valencia	0.003	0.056	0.001	0.037	0.005	0.073
Basque Country and Galicia	0.002	0.049	0.000	0.018	0.005	0.071
other Spanish regions	0.057	0.232	0.005	0.073	0.121	0.326
individual's place of residence = Barcelona city	0.145	0.353	0.120	0.325	0.177	0.381
Barcelona's metropolitan area	0.314	0.464	0.205	0.404	0.449	0.497
Girona	0.109	0.312	0.138	0.345	0.073	0.261
Tarragona	0.078	0.269	0.079	0.269	0.078	0.268
Southern Catalonia (Terres de l'Ebre)	0.071	0.257	0.114	0.318	0.018	0.131
Western Catalonia (Ponent)	0.126	0.332	0.140	0.347	0.110	0.313
Central Catalonia	0.091	0.287	0.107	0.309	0.071	0.256
Pyrenees	0.065	0.246	0.097	0.296	0.025	0.158
completed education = primary or less	0.254	0.435	0.216	0.412	0.300	0.458
secondary	0.458	0.498	0.439	0.496	0.481	0.500
tertiary	0.267	0.443	0.323	0.468	0.199	0.399
other education levels	0.021	0.143	0.022	0.145	0.020	0.140
number of observations	5357		2961		2396	

Table 3: Linear Probability Model Estimates (selected results) — Subsample of Native Spanish Speakers

	(a)	*(b)*	*(c)*	*(d)*
OLS —Dependent Variable: Partner's Language = Catalan-Only				
Proficiency in Catalan (Cat)	0.044[a]	0.040[a]	0.036[a]	0.035[a]
	(0.003)	*(0.003)*	*(0.003)*	*(0.003)*
OLS —Dependent Variable: Language Used With the Partner = Catalan-Only				
Proficiency in Catalan (Cat)	0.040[a]	0.037[a]	0.029[a]	0.027[a]
	(0.002)	*(0.002)*	*(0.002)*	*(0.002)*
Parents' controls	*NO*	*YES*	*NO*	*YES*
Individual controls	*NO*	*NO*	*YES*	*YES*
Number of observations	2,396	2,396	2,396	2,396

Note: OLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth clusters. [a] Significant at 1%; [b] significant at 5%; [c] significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial and year of birth dummies. Regressions in column (b) contain controls for paternal and maternal place of birth (with missing indicators), dummies for Catalan as parental native language and highest parental education (with missing indicators). Regressions in column (c) include controls for individual's place of birth, place of residence and completed education (with missing indicators). Complete results are reported in Tables A1a and A1b in the online Appendix.

Table 4: 2SLS Estimates (selected results) — Joint Sample of Spanish and Catalan Speakers

	(a)	*(b)*	*(c)*	*(d)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*				
I(l = Spanish)	-2.105[a]	-1.556[a]	-1.643[a]	-1.376[a]
	(0.099)	*(0.098)*	*(0.085)*	*(0.097)*
I(l = Spanish) × ce_t	0.115[a]	0.115[a]	0.105[a]	0.104[a]
	(0.017)	*(0.017)*	*(0.014)*	*(0.015)*
F-test of excluded instruments	46.84	48.29	52.73	46.14
[p-value]	*[0.000]*	*[0.000]*	*[0.000]*	*[0.000]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*				
Proficiency in Catalan (Cat)	0.076[a]	0.068[a]	0.077[a]	0.073[a]
	(0.024)	*(0.023)*	*(0.024)*	*(0.024)*
I(l = Spanish)	-0.261[a]	-0.132[a]	-0.195[a]	-0.109[a]
	(0.043)	*(0.030)*	*(0.033)*	*(0.028)*
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*				
Proficiency in Catalan (Cat)	0.053[b]	0.035[c]	0.057[c]	0.043[c]
	(0.021)	*(0.020)*	*(0.023)*	*(0.023)*
I(l = Spanish)	-0.517[a]	-0.269[a]	-0.416[a]	-0.234[a]
	(0.040)	*(0.032)*	*(0.032)*	*(0.030)*
Parents' controls	*NO*	*YES*	*NO*	*YES*
Individual controls	*NO*	*NO*	*YES*	*YES*
Number of observations	5,357	5,357	5,357	5,357

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth clusters. [a] Significant at 1%; [b] significant at 5%; [c] significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial and year of birth dummies. Regressions in column (a) represent our baseline results. Regressions in column (b) contain controls for paternal and maternal place of birth (with missing indicators), dummies for Catalan as parental native language and highest parental education (with missing indicators). Regressions in column (c) include controls for individual's place of birth, place of residence and completed education (with missing indicators). 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 are reported in Table A3 in the online Appendix.

Table 5: Sensitivity to Language Switchers

	(a)	*(b)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*		
I(l = Spanish) × ce$_t$	0.115[a]	0.199[a]
	(0.017)	*(0.020)*
Adjusted R^2	0.203	0.338
F-test of excluded instruments	46.84	94.84
[p-value]	*[0.000]*	*[0.000]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*		
Proficiency in Catalan (Cat)	0.076[a]	0.063[a]
	(0.024)	*(0.012)*
Adjusted R^2	0.201	0.282
2SLS — *Dependent Variable: Language Used with the Partner = Catalan-Only*		
Proficiency in Catalan (Cat)	0.053[b]	0.055[a]
	(0.021)	*(0.012)*
Adjusted R^2	0.406	0.532
Number of observations	5,357	4,276

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth clusters. [a] Significant at 1%; [b] significant at 5%; [c] significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial, an indicator for being native Spanish speaker and year of birth dummies. 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. Regressions in column (a) represent our baseline results. In column (b) we repeat the estimations excluding individuals who switch from Spanish (native language) to Catalan (language of self-identification).

Table 6a: Falsification Analysis (Baseline and Placebo Reduced Form Equations)

Dependent Variable: Partner's	***Real Reform of:***				***Placebo reform in***				
Language = Catalan-Only	1983	1970	1969	1968	1967	1966	1965	1964	1963
I(l = Spanish)	-0.420[a]	-0.446[a]	-0.443[a]	-0.440[a]	-0.435[a]	-0.430[a]	-0.427[a]	-0.428[a]	-0.430[a]
	(0.016)	*(0.024)*	*(0.025)*	*(0.025)*	*(0.026)*	*(0.027)*	*(0.028)*	*(0.030)*	*(0.030)*
$I(l = Spanish) \times ce_t$	0.009[a]								
	(0.003)								
$I(l = Spanish) \times ce_t$*		0.005							
		(0.004)							
$I(l = Spanish) \times ce_t$*			0.004						
			(0.004)						
$I(l = Spanish) \times ce_t$*				0.003					
				(0.004)					
$I(l = Spanish) \times ce_t$*					0.002				
					(0.004)				
$I(l = Spanish) \times ce_t$*						0.001			
						(0.004)			
$I(l = Spanish) \times ce_t$*							0.000		
							(0.005)		
$I(l = Spanish) \times ce_t$*								0.000	
								(0.005)	
$I(l = Spanish) \times ce_t$*									0.001
									(0.005)
Adjusted R^2	0.185	0.210	0.210	0.210	0.210	0.210	0.210	0.210	0.210
Number of observations	5357	3417	3417	3417	3417	3417	3417	3417	3417

Note: OLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth. [a] Significant at 1%; [b] significant at 5%; [c] significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial, an indicator for being native Spanish speaker and 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 1944 and 1969, in Catalonia or migrated before age 6); placebo compulsory exposure (ce_t) 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).*

Table 6b: Falsification Analysis (Baseline and Placebo Reduced Form Equations)

Dependent Variable: Language Used With the Partner = Catalan-Only	**Real Reform of**	***Placebo reform in:***							
	1983	***1970***	***1969***	***1968***	***1967***	***1966***	***1965***	***1964***	***1963***
$I(l = \text{Spanish})$	-0.628^{a}	-0.615^{a}	-0.612^{a}	-0.607^{a}	-0.603^{a}	-0.598^{a}	-0.593^{a}	-0.589^{a}	-0.583^{a}
	(0.011)	*(0.019)*	*(0.020)*	*(0.021)*	*(0.023)*	*(0.024)*	*(0.025)*	*(0.026)*	*(0.027)*
$I(l = \text{Spanish}) \times ce_t$	0.006^{b}								
	(0.003)								
$I(l = \text{Spanish}) \times ce_t*$		-0.001							
		(0.003)							
$I(l = \text{Spanish}) \times ce_t*$			-0.002						
			(0.003)						
$I(l = \text{Spanish}) \times ce_t*$				-0.003					
				(0.003)					
$I(l = \text{Spanish}) \times ce_t*$					-0.003				
					(0.004)				
$I(l = \text{Spanish}) \times ce_t*$						-0.004			
						(0.004)			
$I(l = \text{Spanish}) \times ce_t*$							-0.005		
							(0.004)		
$I(l = \text{Spanish}) \times ce_t*$								-0.005	
								(0.004)	
$I(l = \text{Spanish}) \times ce_t*$									-0.006
									(0.004)
Adjusted R^{2}	0.384	0.397	0.398	0.398	0.398	0.398	0.398	0.398	0.398
Number of observations	5357	3417	3417	3417	3417	3417	3417	3417	3417

Note: OLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth. ^{a} Significant at 1%; ^{b} significant at 5%; ^{c} significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial, an indicator for being native Spanish speaker and 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 1944 and 1969, in Catalonia or migrated before age 6); placebo compulsory exposure (ce_t) 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).*

Table 7a: Sensitivity to Alternative Language Definitions and Identifying Assumptions

	baseline	*(a)*	*(b)*	*(c)*	*(d)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*					
I(l = Spanish) × ce_t	0.115^{a}		0.075^{a}	0.075^{a}	0.074^{a}
	(0.017)		*(0.017)*	*(0.017)*	*(0.016)*
I(l= parents Spanish speakers) × ce_t*		0.114^{a}	0.048^{b}	0.048^{b}	
		(0.018)	*(0.019)*	*(0.019)*	
$\phi_{l^*,t}$					*YES*
F-test of excluded instruments	46.84	38.88	22.67	19.86	20.19
[p-value]	*[0.000]*	*[0.000]*	*[0.000]*	*[0.000]*	*[0.000]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*					
Proficiency in Catalan (Cat)	0.076^{a}	0.056^{a}	0.063^{a}	0.138^{b}	0.142^{c}
	(0.024)	*(0.020)*	*(0.021)*	*(0.070)*	*(0.074)*
I(l= parents Spanish speakers) × ce_t*				-0.010	
				(0.008)	
$\theta_{l^*,t}$					*YES*
Hansen J test for overidentification			1.526		
[p-value]			*[0.211]*		
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*					
Proficiency in Catalan (Cat)	0.053^{b}	0.055^{a}	0.050^{b}	0.094	0.093
	(0.021)	*(0.018)*	*(0.021)*	*(0.061)*	*(0.064)*
I(l= parents Spanish speakers) × ce_t*				-0.006	
				(0.007)	
$\theta_{l^*,t}$					*YES*
Hansen J test for overidentification			0.718		
[p-value]			*[0.397]*		
Number of observations	5,357	5,193	5,193	5,193	5,193

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth clusters. [a] Significant at 1%; [b] significant at 5%; [c] significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial and year of birth dummies. The baseline regression and models in columns (b), (c) and (d) also contain an indicator for being native Spanish speaker; models in columns (a), (b), (c) and (d) also contain an indicator for individuals whose parents are both Spanish-only speakers. Regressions in column (d) include interactions between year of birth dummies and the indicator for individuals whose parents are Spanish-only speakers. 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 and the indicator for having Spanish-only speaking parents respectively.

Table 7b: Sensitivity to Alternative Language Definitions and Identifying Assumptions

	baseline	*(a)*	*(b)*	*(c)*	*(d)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*					
I(l = Spanish) × ce_t	0.115^{a}		0.067^{a}	0.067^{a}	0.065^{a}
	(0.017)		*(0.021)*	*(0.021)*	*(0.021)*
I(l= non-Catalan origins) × ce_t*		0.131^{a}	0.071^{a}	0.071^{a}	
		(0.019)	*(0.024)*	*(0.024)*	
$\phi_{l^*,t}$					*YES*
F-test of excluded instruments	46.84	48.38	24.73	10.35	9.69
[p-value]	*[0.000]*	*[0.000]*	*[0.000]*	*[0.001]*	*[0.003]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*					
Proficiency in Catalan (Cat)	0.076^{a}	0.085^{a}	0.070^{a}	0.081	0.081
	(0.024)	*(0.026)*	*(0.024)*	*(0.055)*	*(0.059)*
I(l= non-Catalan origins) × ce_t*				-0.002	
				(0.008)	
$\theta_{l^*,t}$					*YES*
Hansen J test for overidentification			0.025		
[p-value]			*[0.875]*		
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*					
Proficiency in Catalan (Cat)	0.053^{b}	0.089^{a}	0.051^{a}	0.014	0.010
	(0.021)	*(0.024)*	*(0.022)*	*(0.052)*	*(0.054)*
I(l= non-Catalan origins) × ce_t*				0.005	
				(0.007)	
$\theta_{l^*,t}$					*YES*
Hansen J test for overidentification			0.715		
[p-value]			*[0.398]*		
Number of observations	5,357	5,357	5,357	5,357	5,357

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth clusters. ^{a} Significant at 1%; ^{b} significant at 5%; ^{c} significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial and year of birth dummies. The baseline regression and models in columns (b), (c) and (d) also contain an indicator for being native Spanish speaker; models in columns (a), (b), (c) and (d) also contain an indicator for individuals with non-Catalan origins. Regressions in column (d) include interactions between year of birth dummies and the indicator for individuals with non-Catalan origins. 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 and the indicator for non-Catalan origins respectively.

Table 8: 2SLS Estimates (Selected Results) — Subsample of Spanish Speakers

	(a)	*(b)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*		
I(l = non-Catalan origins)*	-0.790[a]	-0.756[a]
	(0.212)	*(0.228)*
I(l = non-Catalan origins) × ce_t*	0.096[a]	0.097[a]
	(0.023)	*(0.031)*
F-test of excluded instruments	9.83	9.42
[p-value]	*[0.000]*	*[0.000]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*		
Proficiency in Catalan (Cat)	0.047	0.097[c]
	(0.051)	*(0.050)*
I(l = non-Catalan origins)*	-0.024	-0.003
	(0.022)	*(0.028)*
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*		
Proficiency in Catalan (Cat)	0.076[c]	0.079[c]
	(0.043)	*(0.047)*
I(l = non-Catalan origins)*	-0.025	-0.034
	(0.022)	*(0.024)*
Number of observations	2,396	2,396

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth clusters. [a] Significant at 1%; [b] significant at 5%; [c] significant at 10%. All regressions include dummies for wave and gender, a cubic age polynomial and year of birth dummies. Regressions in column (b) exclude observations of individuals whose partner has both Catalan and Spanish as native language. 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 having non-Catalan origins.