Your Language or Mine?

Ramon Caminal
Antonio Di Paolo

November 2015

Barcelona GSE Working Paper Series

Working Paper nº 852

Your language or mine?*

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

November 2015

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 expanded cooperation across speech communities.

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

Keywords: partnership formation, preferences, segregation, language skills, language use, language policy.

[*] We thank Lorenzo Cappellari, Ana Rute Cardoso, 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) and Generalitat de Catalunya (SGR 2014-2017) for their financial support. The usual disclaimers apply.

[†] Institut d'Anàlisi Econòmica (CSIC) and Barcelona GSE: ramon.caminal(at)iae.csic.es

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

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 have a significant influence on economic outcomes. Studies in two specific areas provide useful examples. First, it has been shown that sharing a common language promotes international trade (e.g., Frankel and Rose, 2002; Melitz, 2008). Second, 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 (either directly or through translators) 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 and misguided views. For instance, 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.

[1] They formalized the idea that learning a second language generates network externalities. Since individuals do not internalize these externalities, their incentives to learn second languages are inefficiently low. As a result there is room for public intervention. In their model, the cost of learning a second language is the same for everyone. Hence, in the social optimum, speakers of the majority language remain monolingual and communication barriers are eliminated by inducing the minorities to learn the majority language.

By and large, economists have recognized that languages are much more than neutral communication devices. A prominent example is the recent book by Ginsburgh 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.[2]

Hence, policy makers should pay attention to both the role of languages as means of communication as well as their subjective, emotional aspects. Unfortunately, we know very little about the relevance of these potential trade-offs. Even some basic questions remain largely unexplored. For example, how relevant is the subjective dimension of a language? Does it affect individual behavior in the marketplace or in other social interactions? What does it imply for the design of efficiency-oriented policies?

In this paper we address these issues both theoretically and empirically. We first present a theoretical model 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 sharing a common language is a prerequisite for some types of economic and social interactions.[3] On top of this, we introduce the notion of linguistic preferences. We assume that even fully bilingual individuals are not indifferent about the language to be used in various situations. These preferences, whose intensity may depend on the individual, may reflect an emotional attachment to the individual's native language or the language adopted as their own in later stages.[4]

The model considers a bilingual society that has initially an 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. Thus, all agents share a common language, and hence the role of

[2] In their own words: "Sustaining a high degree of societal diversity could require allocating substantial resources to creating educational institutions and developing communications and coordination between groups .... Since language is an intimate part of individual and group identity, restricting linguistic rights may alienate and disenfranchise groups of individuals whose cultural, societal, and historical values and sensibilities are perceived to be threatened" (page 11).

[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 and on the use of the minority language.

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.[5]

We then test these predictions exploiting data from two successive waves of a survey conducted in Catalonia (Spain). The survey provides detailed information on the socio-demographic and linguistic attributes of the respondents as well as the linguistic characteristics of their partners. There are two main reasons why Catalonia is a unique test field. First, it is a bilingual society (Spanish and Catalan are the two main languages) where the ability to communicate has never been at stake, at least, throughout its recent history, because of the universal knowledge of 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 legal 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 policy 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 ef-

[5] It is important to note that if we abstract from learning costs, such an improvement in language skills increases the total surplus. The reason 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 unilaterly decide to match with an inferior but linguistically homogeneous partner. This welfare result may provide an efficiency justification for public policies that promote the learning and use of weak languages (provided the learning costs are not excessive).

[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.

fect of language exposure during compulsory education allows us to generate causal estimates of the variable of interest.

The main goal of the empirical analysis is to study the influence of language skills on the formation of mixed couples and the use of the weak language (Catalan).[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. Such an exposure variable was used by Clots-Figueras and Masella (2013) in their study of the impact of the education reform on identity formation, taking a reduced-form approach.[8] Since the amount of exposure to the language-in-education reform during compulsory education is imputed on the basis of year of birth, this variable itself could affect partnership formation and language use directly, due to general cohort effects in the outcomes of interest. For this reason, and in the spirit of Bleakly and Chin (2004, 2008, 2010), we include native Catalan speakers in the analysis in order to control for these non-linguistic effects. Therefore, assuming that both native Catalan and native Spanish speakers were equally affected by common trends in the cohorts, the identifying variable is the interaction between the exposure to Catalan during compulsory schooling and a dummy variable for being a native Spanish speaker.

Our results are in line with the theoretical predictions. In particular, the education reform of the 1980's, by improving the oral Catalan skills of native Spanish speakers, raised their propensity to find a Catalan speaking partner and to speak only Catalan with the partner. Hence, these results indicate that linguistic preferences are relevant: as a result, the acquisition of language skills that appear redundant from a communicative viewpoint can significantly reduce segregation. The results turn out to be robust to a battery of sensitivity and falsification tests.

The rest of this paper is organized as follows. In the next section we lay out the theoretical framework and derive two testable hypothesis. In Section 3 we provide some historical background. Section 4 contains the empirical strategy designed to test the main hypothesis. The data set is described in Section 5. The main results

[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] In Section 9 we argue that the effect of the education reform on couple formation is largely unrelated to changes in identity patterns. On the effect of language policy on identity, see also Aspachs et al. (2008).

and robustness checks are presented in Sections 6 and 7, respectively. Section 8 discusses the empirical strategy and results concerning the second hypothesis. Finally, Section 9 summarizes the paper and discusses the results from an identity perspective.

2 The theory

Our stylized model portrays a new channel by which the distribution of language skills in a bilingual society affects the pattern of social interactions. Following the standard theory, we assume that sharing a common language is a prerequisite for most social or economic interactions. On top of this, we introduce the notion of linguistic preferences. If bilingual individuals are not indifferent about their language of use, then the universal knowledge of a *lingua franca*, although it guarantees communication, does not remove all the obstacles to cooperation between members of different speech communities. In this context, the acquisition of language skills that may appear redundant have new, non-trivial effects. We first present the simplest version of the model, and subsequently will discuss various extensions.

2.1 The benchmark model

Consider a country with two languages, called 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 abroad. In particular, A is widely known in the rest of the world and hence very useful for communicating with foreigners.[9] As a result, native B speakers have strong incentives to learn A. We assume that they do and, moreover, achieve full competence in A, so that they can perfectly communicate with domestic, monolingual native A speakers. In other words, the universal knowledge of A implies that the ability to communicate in any domestic interaction is not at stake. Such an assumption holds in many real world examples, including Catalonia, the subject of our empirical study.[10] Moreover, it allows us to emphasize the role of languages beyond their communicative value. The model can be easily extended to take into account a fraction of monolingual speakers of

[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.

[10] Other well-known examples include Wales and the Basque country.

B. But in that case, language skills affect segregation not only through linguistic preferences but also through changing the ability to communicate.

In contrast, B has no value abroad and hence native A speakers can only benefit from learning B if this additional skill facilitates domestic interactions. Hence, native A speakers have weaker incentives to learn B. We define below a parameter that represents the level of proficiency in B of native A speakers.[11]

Individuals derive utility from forming partnerships with other compatriots (e.g., trade partnerships, couples).[12] 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 suppose 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. Therefore, it is important whether or not the two members of a best match belong to the same speech community.[13] If they are, 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, we assume that 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 assume that $f(w) > 0$ for all $w \in [0, \overline{w}]$ and there are no mass points. If instead

[11] In some cases, like Wales or the Basque country, the stronger language outside the region is also the native language of the majority (α is large). However, in other cases, like Belgium, and especially Quebec, the strong international language is the native language of a minority (α is small). However, in contrast to our model, in the last two examples a significant fraction of the weaker speech community is monolingual. See also the discussion at the end of this section.

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

[13] 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. We will go back to this issue below.

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 can vary, depending on public policies and individual's learning efforts.[14] It is important to note that 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, and hence they do not need to worry about the language of use in such a contingency.

In the case of mixed matches, and depending on the values of g_{ab}, w_a, η_a, w_b, potential partners must decide whether or not to form the partnership, and the language of use in case they do. We consider the following bargaining environment. First, we assume that all these parameters are common knowledge.[15] Second, we assume that if both parties agree on forming the partnership, then they choose the language that maximizes the joint surplus. Both assumptions aim at minimizing the frequency of disagreements. The only friction that remains in place is that one party cannot compensate the other for the use of a second language (non-transferable utility).[16] Hence, 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

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

It is important to note that these participation constraints imply that individuals do not internalize the negative externality imposed on their potential partners in case they unilaterally decide not to form the partnership. For instance, suppose that $g_{ab} - \eta_a - w_a < g_{ab} - w_b = -\varepsilon$, where $\varepsilon < g_{ab}$. That is, the efficient language of use

[14] 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.

[15] Asymmetric information would only exacerbate the inefficiency of equilibria, without bringing about additional insights.

[16] Some real world examples may be better described by transferable utility (monetary compensations may be feasible). However, if there is asymmetric information on preferences, and/or if strategic considerations prevent the efficient language from always being implemented, then the main qualitative result (best matches break up too frequently) would also hold.

in this particular match is A. However, b cannot accept forming the match, using A, and incurring a loss of ε. However, the refusal to form the match is detrimental to a, who loses g_{ab}, which is higher than ε. More generally, if decisions were taken by a social planner aiming at maximizing total surplus (first best), then the best match would be formed if and only if

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

Figure 1a depicts the equilibrium outcome (i.e., when individuals are allowed to unilaterally abstein from 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.[17]

We are particularly interested in studying the impact of a general reduction in η_a: that is, an improvement in native A speakers' proficiency in B. In order to avoid uninteresting technical issues, in the rest of the exposition we will focus on the particular case that $g_{ab} = g$ and η_a is distributed on $\left[\underline{\eta}, \overline{\eta}\right]$ with a certain 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 for all a, η_a is sufficiently high (the exact condition is $\underline{\eta} \geq g$). That is, all as are essentially monolingual, or at least their proficiency in B is so low that it will never make a difference in the formation of a best match. Let us call such a distribution of language skills "Scenario 0." In this case, B will never be used in a mixed match, and hence such a match's formation will exclusively depend on b's linguistic preferences. In particular, 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., $\overline{\eta} = 0$ for all a. Let us call this "Scenario 1." In this case, both languages are in a symmetric position, and hence the outcome is also symmetric: both languages are used with a fifty percent chance. Moreover, the

[17] 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.

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. In other words, suppose we start from a situation where a positive fraction of as are willing to make the best match and use B: that is, $\underline{\eta}< g$. Let us describe an improvement in a's proficiency in B as a shift in the distribution of η_as such that the initial distribution first-order stochastically dominates the final distribution (some positive mass of as experience a reduction of their η_a), and such a shift involves at least an interval of η_a with $\eta_a < g$. Then, we obtain the following result:

Result 1 If native A speakers improve their proficiency in B, then (i) the fraction of successful mixed matches increases, and (ii) B is used more often in those matches.

See the Appendix for details.

Result 1 is 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 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, bilingual individuals are worse off than monolingual individuals.

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: $U_a^1 + U_b^1 > U_a^0 + U_b^0$. However, the as may be better off or worse off in Scenario 1: U_a^1 may be higher or lower than 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 have to bear half of the costs of using their second language.[18]. 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.[19] 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.[20] There are two reasons why Scenario 1 generates a higher total surplus than Scenario 0. First, Scenario 1 generates a higher rate of occurrence of best matches. Second, it allows a reduction in the total discomfort from using the second language, since B can now be used whenever $w_a < w_b$.[21]

In the Appendix we show that the same comparative statics hold for gradual but general changes in η_a. That is, if we start from a situation where $\underline{\eta} < g$, then we obtain the following result:

Result 2 A general reduction in η_a, abstracting from learning costs, (i) always raises native B speakers' average payoff, (ii) may or may not raise native A speakers' average payoff, and (iii) always raises aggregate payoffs.

Result 2 provides a welfare assessment of the comparative statics of Result 1. That is, an exogenous improvement in the proficiency in the weak language

[18] 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.

[19] 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.

[20] Hence, if monetary transfers across speech communities are feasible, and learning costs are not too high, then Scenario 1 dominates.

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

among native speakers of the strong language raises total welfare, at least if we abstract from learning costs. However, it may also have non-trivial distributional implications.

2.2 Discussion

Assortative matching. The previous description of the model was silent about how the best matches are determined. All we did was to show how language skills and preferences influence the intensity of the linguistic conflict and determine the rate at which a potential match translates into a realized match. A more sophisticated model would embed our bargaining problem into an environment where individuals face search and matching frictions and engage in directed search. In such a scenario, one would expect that individuals with strong preferences would be more likely to find a best match within the same speech community. Thus, inefficient matching would result not so much from the externalities associated with the bargaining failures (as in our benchmark model) but also from the complementarities in search efforts. In any case, an increase in the fraction of bilingual native speakers of the strong language is likely to reduce endogamy and raise the total surplus.

Endogenous learning. In the benchmark model, we have treated a's proficiency in B as an exogenous parameter. This is compatible with the empirical analysis, where we emphasize the exogeneity of policy actions (language-in-education policies). However, language skills can also be the result of voluntary decisions taken by individuals. In other words, causality can also work in the opposite direction: a native A speaker engaged in a mixed partnership has incentives to exert a higher effort in order to improve their command of B. Our empirical strategy focuses on one direction of causality (from language skills to couple formation), but clearly does not rule out the possibility of causality in the opposite direction. In any case, it is important to note that even if causality from mixed matches to the acquisition of additional language skills was relevant, this would not diminish the crucial role of linguistic preferences. In other words, if we observe that an individual involved in a mixed match, in spite of the fact that the ability to communicate is not at stake, is willing to undertake a costly effort in order to acquire additional language skills, then this also indicates that individuals' linguistic preferences are indeed a relevant factor in the formation of linguistically mixed matches.

Monolingual minority speakers. Another useful generalization of the benchmark model would consist in relaxing the assumption that all the members of the weak speech community are fully competent in the strong language.[22] If a fraction of the bs are monolingual, then Result 1 would still hold. More specifically, an improvement in a's proficiency in B (lower η_a) would increase the fraction of mixed matches. The argument is the same one used in deriving Result 1, but reinforced by the presence of an additional channel. A reduction in η_a increases the probability that a is willing to form the match and use language B. But now such an adjustment applies to a larger base: not only those potential matches with high values of w_b (the pure preference effect), but also to those matches with high values of η_b (better communication effect). For the same reason, such an improvement in language skills expands total welfare by a greater amount, and improves the use of B. Finally, a general improvement of b's proficiency in A (a general reduction in η_b) will have the symmetric effects (more successful best matches and more frequent use of A).

Learning costs. Our welfare results ignored learning costs. Of course, a complete discussion of the policy implications of our model would require considering not only the benefits, which has been our focus, but also the costs. In any case, two straightforward remarks are in order. First, public efforts to promote the weak language are more likely to raise the total net surplus when the sizes of the two speech communities are not very different. As α falls, starting from a level higher than one-half, the total costs associated with learning the weak language are reduced and at the same time the potential gains from cooperation between speech communities increases. Second, if the two languages are relatively close, then the learning costs are lower and hence promoting the weak language becomes cheaper.[23]

3 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. As a result of the expansion of the Catalano-Aragonese crown, it soon spread out into other

[22] Such an assumption does not fit, for instance, the cases of Belgium, Quebec, the Ukraine, or Latvia.

[23] Both conditions are met in the case of Catalonia: the two main speech communities have a similar size and, as discussed in the next section, both languages have evolved from Vulgar Latin, and hence the distance between them is relatively small.

regions, such as Valencia and the Balearic Islands (where it is still spoken) and overseas.[24] 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 (Branchadell, 2006).[25]

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. In contrast, Spanish was the only official language and the only language used in education. Towards the end of this period, repression against Catalan eased to some extent: books and magazines written in Catalan started getting published (of course, subject to censorship), but the presence of Catalan in public events was systematically precluded. 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).

During the transition from dictatorship to democracy, the reestablishment of the (provisional) regional government of Catalonia (the *Generalitat provisional*) prompted the first attempts at promoting the Catalan language. In particular, over the period 1978–1982, Catalan became a compulsory subject (three hours per week) in all schools. Also, a small fraction of primary schools (whose pupils were mostly native Catalan speakers) adopted Catalan as the medium of instruction.

The true turning point for language-in-education policies arrived a bit later, after the constitution of the permanent regional government (the Autonomous Community) in 1980. The new entity soon acquired decision powers in important areas such as education and the public media. In 1983, the regional parliament passed (unanimously) the "Language Normalization Act" (LNA), which aimed at making

[24] Catalan literature flourished during the Middle Ages with writers such as Ramon Llull and Joanot Martorell (author of the celebrated *Tirant lo Blanc*).

[25] Massive school enrollment did not take place in Spain until the twentieth century. Even though elementary education became compulsory in 1838, the percentage of the primary-school age population enrolled in school was only 42% in 1872, far below the levels prevailing in contemporary France and England (Nuhoğlu Soysal and Strang, 1989).

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 LNA set the legal framework that allowed the dramatic changes in language policy that occurred over the next two decades. However, its application was gradual. In the period 1984–1993, the two languages were taught as subjects for a similar amount of hours. Also, they were both used as the language of instruction in proportions that varied over time (the average fraction of subjects taught in Catalan increased over time) and geographycally, depending on the linguistic characteristics of the students and teachers' language skills.[26]

The LNA also introduced "language immersion programs" in the primary schools and preschools, inspired by those implemented earlier in Quebec. These programs were aimed at schools in predominantly Spanish-speaking neighborhoods. Schools in immersion programs used Catalan as the only language of instruction for the first years of education and followed a specific methodology to stimulate the acquisition of Catalan as a second language. Spanish was introduced as an additional language of instruction only at a later stage (normally grade 3). Immersion programs were tried in 1984 in a handful of schools, but they expanded very quickly. In 1990 they involved about one-fourth of all public schools.

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).[27] Since 1994, the authorities (after a favorable ruling by the Spanish Constitutional Court) 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).[28] 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

[26] Some minimum mandatory requirements were set. Catalan had to be used as the language of instruction in at least one area of study (out of eight) from grades 3 to 5, and in two areas from grade 6 onwards, while Spanish had to be used as the language of instruction in at least one area throughout the entire curriculum.

[27] The share of primary schools using Catalan as the main medium of instruction rose from 42% in 1986 to 73% in 1992, while those employing both Catalan and Spanish decreased from 33% to 24% over the same period (Vila-i-Moreno, 2000, and Vila-i-Moreno and Galindo-Solé, 2009).

[28] Unfortunately, we do not have more precise information on the evolution of the average number of hours taught in each language, or their geographical variation. All we know is that the use of Catalan increased over time and that differences in language policies between schools did not cause a significant reallocation of students.

instruction, at least in compulsory education.[29]

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 speakers and native Spanish speakers), with basically no effect on the Spanish skills of either speech community.[30]

The regional authorities also sought to promote the knowledge and use of Catalan using a variety of means. For instance, in 1983 the regional government launched a Catalan-only TV channel (and a radio station) that managed to capture about 20% of the audience on average. The same year, it started the first catalanization campaign (Norma), which was followed by successive campaigns with different goals and formats. Finally, the "Language Policy Law" of 1998 also introduced several requirements that raised the value of Catalan skills in the labor market.[31]

4 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 repeated 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), habitual language (most frequently used), the language of

[29] 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.

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

[31] For instance, it established proficiency requirements in Catalan to access public sector jobs, and introduced various regulations affecting the language choices of retailers, restaurants, and hotels.

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.[32] We pay special attention to the partner's language[33] 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. We also exclude individuals born before 1950, after 1990, and those who were students at the time of the survey. 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 the habitual and self-identification language) are classified into three categories: (1) only Catalan, (2) both Catalan and Spanish, and (3) only Spanish. Thus, an immediate question is how to classify the middle group: that is, where to draw the line between these two speech communities. In the baseline analysis we choose a strict definition of the Spanish speaking community (the main focus of the analysis) and hence allocate the middle group to the Catalan speaking community. In other words, we say a respondent is a native Spanish speaker if they reported option (3), only Spanish, as their native language. According to this definition, native Spanish speakers amount to about 45% of the restricted sample (2,396 individuals). The rest (individuals selecting options (1) and (2), 2,961 observations) are classified as native Catalan speakers. Of course,

[32] 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).

[33] Unfortunately, this information is collected in a slightly different way in the two waves of the survey. In 2008, the survey inquired about the partner's habitual language, whereas in 2013, about the native language. Since the results are virtually identical when we estimate the model(s) separately for each wave (results available upon request), such a small inconsistency in data collection is not likely to have a sizeable impact on the results.

we checked that the main results are robust to alternative definitions.

The language proficiency variables are coded with a 0–10 scale, with 0 being the lowest, and 10 the highest, level of proficiency. We claim that oral skills (and in particular, the ability to speak) are crucial in couple formation, whereas writing skills are much less important. Moreover, our empirical strategy can be more easily justified if we restrict attention to oral skills. 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 stable across the cohorts and ranges around 9.5 for both speech communities. Thus, native Catalan speakers are largely bilingual (with an excellent command of both languages), whereas earlier generations of native Spanish speakers had a limited command of Catalan, and younger generations are becoming increasingly bilingual, possibly due to the language-in-education reform.

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. As in the case of oral skills, the level of written Spanish proficiency (Figure 3b) is uniformly high and virtually identical for both speech communities. 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.[34]

The partner's language is also classified into the same three categories as the respondent's native language. In the baseline analysis, we define a respondent's partner as a Catalan speaker if either option (1) or (2), Catalan-only or Catalan and Spanish, is reported, which is consistent with the above definition of the respondents' native languages. 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.[35] In the baseline analysis we choose a strict definition of the use of Catalan: we say a

[34] This evidence is also compatible with the results of the systematic tests mentioned in Section 3 conducted by the national educational authorities.

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

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.

These results are compatible with the predictions of the theoretical model: there is a positive association between additional language skills that are redundant from the communicational viewpoint, and linguistic fragmentation in partnership formation. In the rest of this paper we investigate in more detail the empirical relevance of the theoretical results. First, we present a simple regression setup in which we take into account the observable covariates and provide conditional correlations. Second, we apply an instrumental variable strategy that allows a causal interpretation of the results. Finally, we present several robustness checks and sensitivity analyses. The full set of control variables that are used in the rest of the paper is presented in Table 2, together with basic descriptive statistics by language group.

5 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):

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

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.[36]

We next include several controls for parental background (parents' place of birth, education, habitual language) and for individual attributes (place of birth, place of residence, and completed education), in order to check whether the coefficients that relate language skills with partnership formation and language use are robust to the inclusion of a more demanding set of controls.

Given that the testable hypothesis refers to native speakers of the strong language, 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 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 the weak language on the part of native speakers of the strong language 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 and individual characteristics (columns b, c, and d).[37]

Essentially, the only statistically significant parental characteristic seems to be the parents' habitual language. Having at least one parent whose habitual language was Catalan is positively related to the two outcomes. Conditional on the parents' habitual language, their place of birth is not significantly related with the probability of choosing a Catalan-speaking partner.

In contrast, several individual controls are significantly correlated with the two outcomes. If the place of birth, and even more importantly, the place of residence, is an area with a relatively high fraction of Catalan speakers, the propensity to

[36] 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.

[37] 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 parental characteristics that are potentially endogenous with respect to the two outcomes. Therefore, the evidence regarding these control variables must be interpreted with caution.

choose a Catalan-speaking partner and/or to use only Catalan with the partner is higher. Having completed tertiary education is also positively related with both outcomes.

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 4. 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.

Overall, the evidence using observational data is consistent with the predictions of the theoretical model. Namely, in a bilingual territory where virtually everyone is fully proficient in the strong language, better skills in the weak language on the part of native speakers of the strong language is associated with a higher frequency of mixed couples and a more intensive use of the minority language with the partner. Nevertheless, these conditional correlations might not represent the causal mechanism portrayed by the theoretical model.

Indeed, using observational data to estimate the causal relation between language proficiency and couple formation, or language use, is challenging for three main reasons. 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. Overall, OLS estimates of the relations between language proficiency and couple formation or language use are unlikely to reflect the causal parameters of interest. Therefore, in the next section we present the identification strategy that enables us to obtain (plausibly) causal estimates, and hence a more powerful test of our theory.

6 Causal evidence: Identification strategy and IV estimates

6.1 Empirical framework

This section describes the empirical strategy to generate consistent estimates that can be interpreted as causal relations rather than conditional correlations. 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. The LNA aimed specifically at turning the entire Catalan population bilingual, regardless of linguistic origins and family background. Hence, the exposure to the new language-in-education policy generated an exogenous improvement in the Catalan skills of younger cohorts, especially of native Spanish speakers.

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.[38] More specifically, Clots-Figueras 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

[38] 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 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.

were subject to eight years of compulsory schooling, and those born in 1983 or after to ten years. [39]

Thus, the variable capturing compulsory exposure to Catalan at school, ce_t, can be conveniently 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 & 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 a very young age assimilate the language more easily). Their identifying variable is the interaction between age at arrival with 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.[40]

[39] 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.

[40] Basically the same strategy has been applied by Miranda and Zhu (2013a, 2013b), and by Isphording and Sinning (2012) to estimate the earnings penalty associated to limited English

In our case, we exploit the fact that oral language skills (the most relevant language attribute affecting partnership formation) 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 in Catalan of native Catalan speakers or the oral proficiency in Spanish of native Spanish 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 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 skills in oral 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):

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

The second-stage equation explains the two outcomes of interest (having a Catalan-speaking partner and language use with the partner) and includes proficiency in oral Catalan as an endogenously determined covariate:

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

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 in the UK and the US, respectively, and by Isphording (2013) to estimate the return to foreign language skills in Spain.

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).[41]

It is important to emphasize that the assumptions under which our identification approach is valid are not trivial, which is the reason why in Section 7 we present a battery of robustness checks and falsification analyses.

6.2 Estimation results

Selected 2SLS estimates of Equations (3) and (4) are displayed in Table 5 (the complete results of the first-stage regressions can be found in Table A2 in the Appendix). Overall, the results obtained from our identification strategy are in line with those obtained by OLS, and hence 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 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). As we add parental controls, the point estimate drops slightly, but still remains positive and highly significant. 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)), 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 4–4.3 percentage points.

The effect of 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. The difference between the OLS and 2SLS estimates could be

[41] In the empirical analysis, we take into account clustering of the standard errors at the year of birth–native language level, which is the level of variation of our instrument.

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, has the expected sign and is strongly significant. Thus, native Spanish speakers affected by the language policy did improve their oral proficiency in Catalan. The corresponding estimates 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. As mentioned above, some of the assumptions used in our estimation strategy require further scrutiny, especially because a causal nature depends on the overall validity of our identification strategy. The next section presents the results of several robustness checks and falsification analyses.

6.3 Sensitivity analyses

In this section we present several sensitivity analyses of our baseline specification.

Gender. So far we have assumed that gender differences in the outcomes of interest are well captured by a shift in the intercept, but it could be the case that language skills have a different effect on partnership formation and language use by males and than by females. Therefore, we run separate estimations of Equations (3) and (4) for males and females, for each of the two outcomes. The results for partnership formation (second panel of Table A3 in the Appendix) reveal that the effect of Catalan skills on the probability of finding a Catalan-speaking partner is somewhat higher for females, albeit not statistically different from the males' coefficient. Females' Catalan use with their partner is also more sensitive to increased proficiency due to language exposure, as shown in the third column of Table A3. However, for the male subsample the coefficient of interest is imprecisely estimated.

Age polynomial. As mention above, the use of two repeated cross-sections breaks the perfect collinearity between the year of birth (which is used to identify the cohorts that are affected by the education reform) and age. Even so, a flexible functional form of the age effect in both Catalan skills and partnership formation is important since it helps in partialling-out the impact of age-related confounders that are not fully picked up by year of birth fixed effects. In the baseline specification, we included a third-order polynomial of age, which was the best compromise between parsimony and strength of the identifying variable. Nevertheless, this choice is far from affecting the results of our model. As shown in Table A4 in the Appendix, the results are insensitive to the choice of the age polynomials (ranging from a linear specification to a fourth-order polynomial), and remain virtually unchanged even when we saturate the model with the full set of age dummies as control variables (column (d)).

Language groups. One source of concern is the role of individual or group identity in explaining our baseline results. In the next section we discuss these issues in more detail. For now, it is sufficient to note that when we exclude from the sample respondents with Spanish as their native language and Catalan as their language of self-identification ("language switchers"), then the main results (column (a)) are very similar to the baseline estimation (compare columns (a) and (b) in Table 6). In particular, the effect of oral skills in Catalan on partnership formation is slightly smaller, and the effect on the use of Catalan, slightly higher. Moreover, the instrument becomes much stronger and the coefficients are estimated more precisely.

One delicate decision in the definition of linguistic communities is the allocation of respondents and partners who appear to lie in the intersection: those that have both Spanish and Catalan as their native/habitual language. We check for robustness by excluding these intermediate responses. The results are presented in columns (c)–(e) of Table 6 and are in line with those obtained from the whole sample. If we exclude intermediate respondents, then we observe a small reduction in the effect of Catalan proficiency on both outcomes, whereas dropping individuals having a partner who speaks both Catalan and Spanish barely affects the results.

Couple stability and commitment. If we exclude respondents who do not have a partner at the time of the survey, but did in the past (column (f) of Table 6), then the results remain unaffected. A probably more interesting exercise is to exclude individuals who do not live with their partner at the time of the interview. That is, we restrict the analysis to couples with a higher degree of stability and commitment. The results indicate (column (g) of Table 6) that the causal effect of oral fluency in Catalan on partnership formation and language use is then much larger, which clearly reinforces our main message.

Alternative specification of the identifying variable. The exclusion restriction in our baseline specification is the interaction between the Spanish-speaker indicator and a linear function of compulsory exposure (ce_t), which appeared to be the best alternative. In fact, as shown in Table 7, the results obtained with linear exposure are "better" in terms of the strength of the instrument. In column (a) we show the results obtained considering dummies for partial/full language exposure during compulsory education, which yields a slightly higher estimate of the causal parameter of interest for partnership formation and the same coefficient as the baseline for language use. Adopting a quadratic specification for years of exposure to Catalan during compulsory education (column (b)) has virtually no impact on the parameter of interest (relative to our baseline estimation), while the first stage estimate indicates that the quadratic term is not relevant. Lastly, we considered dummies for each possible year of exposure, ranging from one to ten. The results are presented in column (c) and are qualitatively similar to the baseline estimation, with the exception of a modest increase in the language proficiency coefficient for the second outcome. However, as can be seen, our preferred specification provides a significantly higher weak identification test, which is the rationale for our final choice.

6.4 Falsification and identification checks

In this section we discuss several checks concerning the validity of our identification strategy.

Two placebo experiments. One component of the identifying variable, exposure to Catalan during compulsory schooling, is defined purely as a function of year of birth. The rationale for relying only on variation produced by year of birth was to eliminate any endogenous component (such as school attainment) in the variable capturing potential language exposure at school. However, such a decision comes at a cost, since compulsory exposure could capture spurious relations due to potential cohort-specific trends in (language-related) couple formation and/or language use. We have run two alternative placebo experiments, which aim at providing evidence that our identifying variable is not contaminated by any spurious effects.

In the first placebo test, we consider an additional subsample of individuals who were born in the same cohort as our main sample but migrated to Catalonia from other Spanish regions only after completing compulsory education (i.e., they migrated at 14 or later). We label them Mig. Therefore, based on their birth cohort, we impute compulsory language exposure to this placebo cohort "as if" they had received compulsory schooling in Catalonia (ce_t^*). We then use 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. We then extend the reduced form equation (5) to include a dummy that identifies this never-treated sample and its interaction with placebo compulsory exposure (ce_t^*) and other control variables (included in X), that is:

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

If there exists a contemporaneous trend in the outcome(s) across the cohorts, that is common to both our main sample and to the auxiliary sample of migrants, this would be picked up by the coefficient η. This would suggest that our exposure variable is driven by such a trend and not by the language proficiency effect. The reduced form estimates are presented in Table 8, where columns (a) and (b) correspond to Equations (5) and (6), respectively. The results confirm the positive effect of exposure during compulsory schooling among Spanish speakers on the two outcomes, whereas the coefficient for placebo exposure among Spanish migrants is not significant, is small in size, and negative.[42] Similar evidence is obtained when we include years since migration to Catalonia (and its square) among the controls (column (c)).

[42] Virtually all individuals belonging to the auxiliary sample of migrants are native Spanish speakers. The results remain the same dropping the 30 observations of migrants who are native Catalan speakers.

The second placebo subsample consists of 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). That is, we impute years of (pseudo) exposure to Catalan at school "as if" the reform had been applied 15 years before, in 1968 instead of 1983 (ce_t^{**}). We estimate the reduced form model (5), but using the subsample of Catalonian-born individuals (or migrated from the rest of Spain, before age 6) who were never affected by the compulsory component of the reform:

$$Y_{itl} = \alpha + \beta' X_i + \pi I\left(l = Spanish\right) + \eta I\left(l = Spanish\right) \times ce_t^{**} + \theta_t + \varepsilon_{itl} \qquad (7)$$

Again, 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, this is not the case, as suggested by the corresponding coefficients in column (d) of Table 8, which are also 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 that are unrelated with the introduction of Catalan at school with the 1983 reform.

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 that the self-reported native language might not be an accurate representation of such an exogenous characteristic; for example, 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 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 habitual/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 9a and 9b, 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.[43]

These results suggest that the baseline specification is generally robust to the use of alternative proxies of native language. Although parental language, which is also reported by respondents, could suffer from the same problems than the native language, parental origins is unlikely to be affected by misreporting or other kinds of errors, and is plausibly exogenous. Moreover, the fact that the estimates obtained using this last proxy are higher than in the baseline is consistent with the idea that the new sub-population of compliers 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.

The availability of two alternative proxies to define language groups opens the possibility of relaxing the main identifying assumptions in our model. Indeed, 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 9a and 9b for Spanish speaking parents, and parents of non-Catalan origin, respectively. In both cases, the estimates of the coefficients 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 restriction can be reasonably excluded from the outcome equation(s). Moreover, 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 speakerss 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 equation(s). In this case, we obtain higher point estimates for Catalan skills when we consider the first proxy, which also lose precision (and strength of

[43] 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 habitual/native languages. However, the results are virtually the same when these observations are excluded (detailed results available upon request).

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. Nevertheless, 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

$$Cat_{itl} = \mu + \lambda' X_i + \pi I\left(l = Spanish\right) + \gamma I\left(l = Spanish\right) \times ce_t + \varphi_{l^*t} + u_{itl} \quad (8)$$

$$Y_{itl} = \alpha + \beta' X_i + \pi I\left(l = Spanish\right) + \delta_{IV} Cat_{itl} + \theta_{l^*t} + \varepsilon_{it} \quad (9)$$

where l^* is one of the two proxies of native language, and the terms φ_{l^*t} and θ_{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 9a and 9b, 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.[44]

[44] The heterogeneous effect of exposure to Catalan by parental language cannot be used as an

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 10. 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.

7 Concluding remarks

We have presented empirical evidence that endorses the idea that languages are much more than neutral communication devices. In particular, we have shown 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 notion of linguistic preferences. Bilingual individuals are not indifferent about their language of use. Hence, any form of social cooperation between members of different speech communities must solve a potential conflict of interest over the choice of language. As more native speakers of the strong language become bilingual, the intensity of the conflict decreases and mixed partnerships become more likely.

Our notion of linguistic preferences is abstract in the sense that we have not made any attempt to explain how these preferences are formed or are related to more specific social phenomena. In particular, it is well known that language is a key symbol of ethnic, national, or class identity. For the case of Catalonia, Woolard (1989) has emphasized that ethnicity is critical to understanding language choices. Thus, one may wonder if our results may simply reflect the dynamics of ethnic politics in Catalonia. More specifically, individuals may tend to look for partners in their same ethnic group, and each group identity is signaled by a different language. Perhaps the educational reform affected the frequency of mixed couples (according to our definition) not so much by changing language skills and reducing the lan-

exclusion restriction, since virtually all Spanish speakers have both parents who have only Spanish as habitual language.

guage conflict, but by inducing a fraction of native Spanish speakers to cross over and become "ethnically Catalan" (that is, by assimilation). The main difference from our interpretation would be that according to such an ethnic perspective, the acquisition of new language skills per se might not change ethnic barriers (after all, native Catalan speakers have been bilingual for generations). However, ethnic or cultural assimilation may be prompted by certain policies. 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, respondents report not only their native language but also their habitual language, and the language of self-identification. We believe that ethnic issues should be reflected in these responses. In our baseline sample, only about 3% of the native Catalan speakers report Spanish as their language of self-identification. In contrast, about 20% of native Spanish speakers report Catalan as their language of self-identification. When we eliminate these "switchers" from the sample, the results remain largely unchanged.[45] Such a test suggests that language skills matter independently of ethnic identity. Thus, we feel quite comfortable with our notion of linguistic preferences, more general than the presumed link between language and ethnic identity. Moreover, the fact that a fraction of native Spanish speakers who maintain Spanish as their language of self-identification, and nevertheless use Catalan (for instance, within the couple) is very important: it suggests that Catalan is being perceived, at least by these group, as "anonymous," that is, everyone's language, like most hegemonic languages, and not necessarily identified with a specific ethnic group ("native Catalans"), like most minoritized languages.[46]

8 References

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-

[45] The estimated coefficient of the Catalan skills variable in the couple formation equation is only slightly smaller after excluding the "switchers." However, the coefficient of the instrumental variable in the first equation almost doubles when we exclude the "switchers," which suggests that the reform had a bigger impact on the Catalan skills of those native Spanish speakers who kept Spanish as their identity language.

[46] See also Woolard (2008) for a discussion of these issues in the case of Catalonia.

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.

Branchadell, A. (2006), *L'aventura del català: De les Homilies d'Organyà al nou Estaut.* Barcelona: L'esfera dels llibres.

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

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.

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.

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.

Isphording, I. (2013), Returns to Foreign Language Skills of Immigrants in Spain. *Labour* 27(4), 443-461.

Isphording, I., and M. Sinning (2012), The Returns to Language Skills in the US Labor Market. CReAM Discussion Paper Series 1236, University College, London.

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.

Miranda, A., and Y. Zhu (2013a), English deficiency and the native–immigrant wage gap. *Economics Letters* 118(1), 38-41.

Miranda, A., and Y. Zhu (2013b), The Causal Effect of Deficiency at English on Female Immigrants' Labour Market Outcomes in the UK. Studies in Economics 1301, University of Kent.

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

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.

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

Vila-i-Moreno, X. (2000), Les polítiques lingüístiques als sistemes educatius dels territoris de llengua catalana. *Revista de Llengua i Dret* 34, 169-208.

Vila-i-Moreno, X., and Galindo-Solé, M. (2009), El sistema de conjunció en català en l'educació primària a Catalunya: impacte sobre els usos. *Treballs de Sociolingüística Catalana* 20, 21-69.

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. (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.

9 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

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

$$U_b = [\Pr(A) + \Pr(B)]\, g - \int_0^{\eta_a} w_b dF\left(w_b\right) - \int_{\eta_a}^{g} \left[1 - F\left(w_b - \eta_a\right)\right] dF\left(w_b\right)$$

The effect of η_a on U_a has an ambiguous sign:

$$\frac{dU_a}{d\eta_a} = -\left[1 - F\left(g\right)\right] f\left(g - \eta_a\right) g + \int_0^{g-\eta_a} f\left(w_a + \eta_a\right) w_a dF\left(w_a\right) + \\ + \left[1 - F\left(g\right)\right]\left(g - \eta_a\right) f\left(g - \eta_a\right)$$

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

Figure1a: Equilibrium outcomes

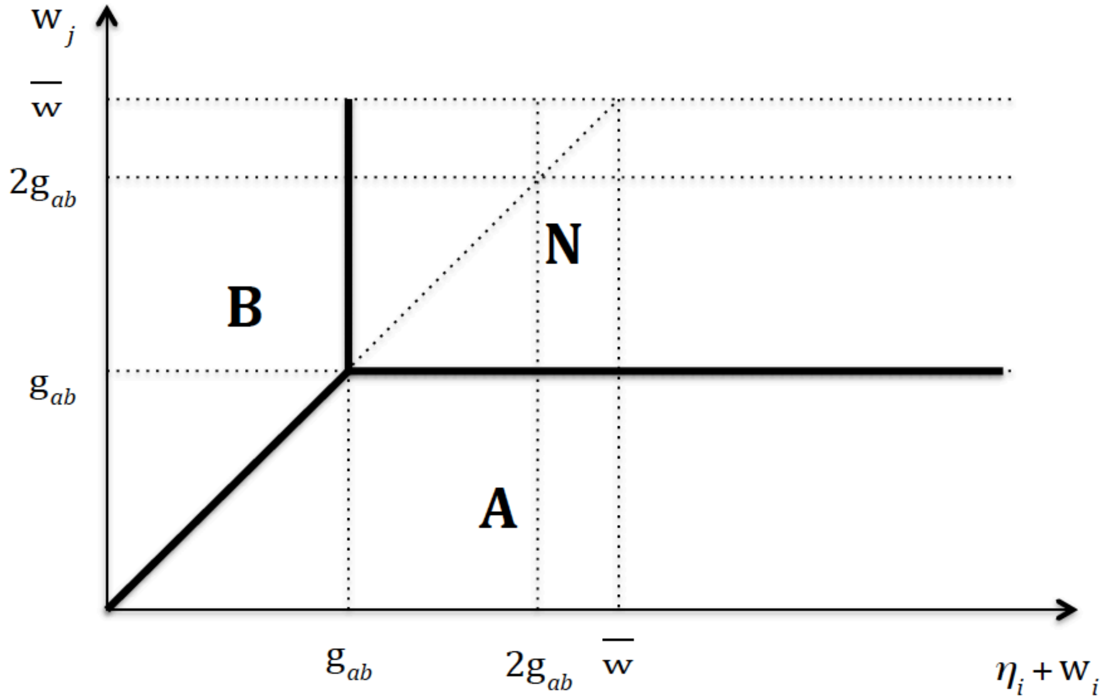

Figure1b: Efficient outcomes

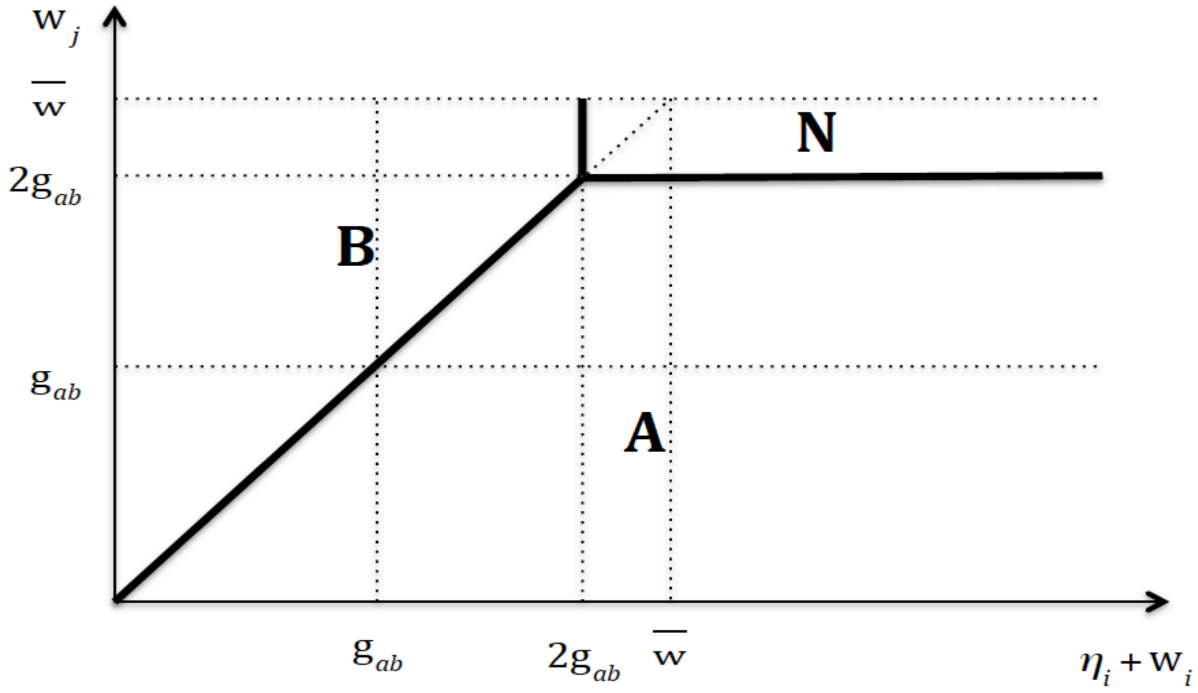

Figure 2a: Average Oral Proficiency in Catalan **Figure 2a: Average Oral Proficiency in Spanish**

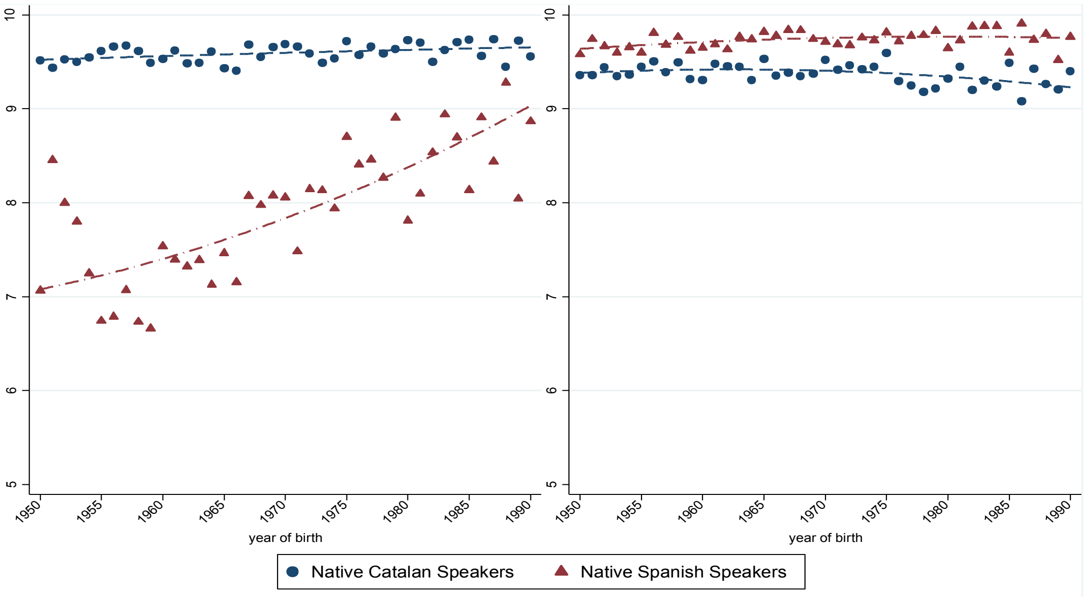

Figure 3a: Average Written Proficiency in Catalan **Figure 3b: Average Written Proficiency in Spanish**

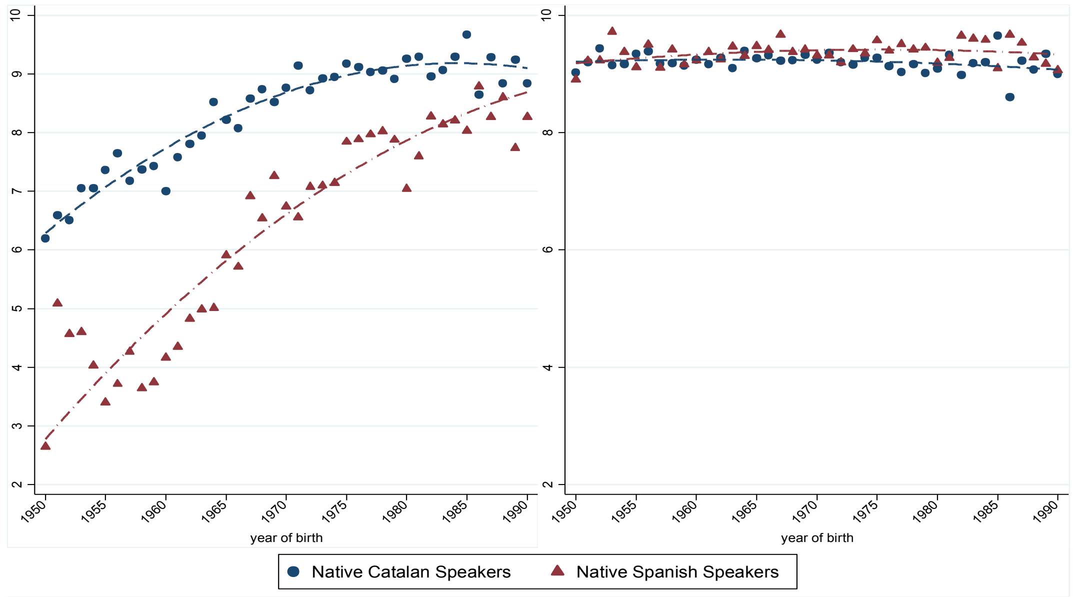

Table 1: Partner's Language and Language Use by Native Language

	% 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	41.695	10.402	42.453	10.722	40.758	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
Catalan used by parents = no Catalan	0.435	0.496	0.067	0.249	0.890	0.313
Catalan used by father or mother	0.157	0.364	0.207	0.405	0.095	0.293
Catalan used by father and mother	0.408	0.492	0.726	0.446	0.015	0.122
missing parents' 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. Regression in column (b) contains controls for paternal and maternal place of birth (with missing indicators), dummies for Catalan as parental habitual language and highest parental education (with missing indicators). Regression in column (c) includes controls for individual's place of birth, place of residence and completed education (with missing indicator). Complete results are reported in Tables A1a and A1b in the Appendix.

Table 4: Linear Probability Model Estimates (selected results) — Joint Sample of Spanish and Catalan Speakers

	(a)	*(b)*	*(c)*	*(d)*
OLS — Dependent Variable: Partner's Language = Catalan-Only				
Proficiency in Catalan (Cat)	0.045[a]	0.041[a]	0.037[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)	0.043[a]	0.037[a]	0.031[a]	0.029[a]
	(0.002)	*(0.002)*	*(0.002)*	*(0.002)*
Parents' controls	*NO*	*YES*	*NO*	*YES*
Individual controls	*NO*	*NO*	*YES*	*YES*
Number of observations	5,357	5,357	5,357	5,357

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, an indicator for being native Spanish speaker and year of birth dummies. Regression in column (b) contains controls for paternal and maternal place of birth (with missing indicators), dummies for Catalan as parental habitual language and highest parental education (with missing indicators). Regression in column (c) includes controls for individual's place of birth, place of residence and completed education (with missing indicator). Complete results available upon request.

Table 5: 2SLS Estimates (selected results) — Joint Sample of Spanish and Catalan Speakers

	baseline	*(a)*	*(b)*	*(c)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*				
I(l = Spanish) × ce$_t$	0.115^a	0.115^a	0.105^a	0.104^a
	(0.012)	*(0.012)*	*(0.011)*	*(0.012)*
F-test of excluded instruments	91.66	93.41	96.84	85.01
[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
	(0.017)	*(0.017)*	*(0.018)*	*(0.018)*
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*				
Proficiency in Catalan (Cat)	0.053^a	0.035^a	0.057^a	0.043^a
	(0.015)	*(0.015)*	*(0.017)*	*(0.017)*
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-native language 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. Regression in column (a) contains controls for paternal and maternal place of birth (with missing indicators), dummies for Catalan as parental habitual language and highest parental education (with missing indicators). Regression in column (b) includes controls for individual's place of birth, place of residence and completed education (with missing indicator). The F-test on excluded instruments refers to the Angrist-Pischke multivariate F-test on the interactions between years of exposure to Catalan at compulsory schooling and the indicator for Spanish as initial language. Complete results of the first-stage regressions are reported in Table A2 in the Appendix.*

Table 6: Sensitivity to Mixed Languages and Partnership Status

	baseline	*(a)*	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*	*(g)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*								
I(l = Spanish) × ce$_t$	0.115^{a}	0.199^{a}	0.203^{a}	0.116^{a}	0.113^{a}	0.114^{a}	0.116^{a}	0.114^{a}
	(0.012)	*(0.015)*	*(0.015)*	*(0.012)*	*(0.013)*	*(0.013)*	*(0.014)*	*(0.016)*
Adjusted R^2	0.203	0.338	0.363	0.204	0.204	0.205	0.203	0.202
F-test of excluded instruments	91.66	176.11	173.70	91.91	81.72	81.65	69.47	53.10
[p-value]	*[0.000]*	*[0.000]*	*[0.000]*	*[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}	0.063^{a}	0.067^{a}	0.066^{a}	0.082^{a}	0.072^{a}	0.075^{a}	0.118^{a}
	(0.017)	*(0.009)*	*(0.011)*	*(0.018)*	*(0.018)*	*(0.018)*	*(0.018)*	*(0.026)*
Adjusted R^2	0.201	0.282	0.302	0.219	0.220	0.246	0.201	0.135
2SLS — *Dependent Variable: Language Used with the Partner = Catalan-Only*								
Proficiency in Catalan (Cat)	0.053^{a}	0.055^{a}	0.062^{a}	0.044^{a}	0.057^{a}	0.045^{a}	0.047^{a}	0.090^{a}
	(0.015)	*(0.009)*	*(0.009)*	*(0.015)*	*(0.015)*	*(0.015)*	*(0.016)*	*(0.020)*
Adjusted R^2	0.406	0.532	0.562	0.453	0.421	0.456	0.413	0.381
Number of observations	5,357	4,276	4,175	5,117	5,019	4,829	4,654	4,157

Note: regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth-initial language 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 initial language. Robustness checks for language switches: (a) excluding individuals who switch from Spanish (native language) to Catalan (language of self-identification); (b) excluding individuals who switch from Spanish (native language) to Catalan (habitual language). Mixed languages: (c) excluding individuals who had Spanish and Catalan as native language; (d) excluding individuals whose partner has Spanish and Catalan as habitual/native language; (e) excluding individuals who had Spanish and Catalan as native language and those whose partner has Spanish and Catalan as habitual/native language. Partnership status: (f) excluding individuals who do not have a partner at the time of the survey; (g) excluding individuals who do not live with their partner at the time of the survey.

Table 7: Sensitivity to Alternative Specifications of the Identifying Variable

	baseline	*(a)*	*(b)*	*(c)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*				
I(l = Spanish) × ce_t	0.115^{a}		0.162^{a}	
	(0.012)		*(0.042)*	
I(l = Spanish) × $(ce_t)^2$			-0.005	
			(0.004)	
I(l = Spanish) × I(no exposure)		*ref. cat.*		
I(l = Spanish) × I(partial exposure)		0.639^{a}		
		(0.121)		
I(l = Spanish) × I(full exposure)		0.981^{a}		
		(0.119)		
I(l = Spanish) × I(ce_t = 0)				*ref. cat.*
I(l = Spanish) × I(ce_t = 1)				0.490^{a}
				(0.080)
I(l = Spanish) × I(ce_t = 2)				-0.051
				(0.079)
I(l = Spanish) × I(ce_t = 3)				0.695^{a}
				(0.079)
I(l = Spanish) × I(ce_t = 4)				0.775^{a}
				(0.079)
I(l = Spanish) × I(ce_t = 5)				0.532^{a}
				(0.079)
I(l = Spanish) × I(ce_t = 6)				1.098^{a}
				(0.081)
I(l = Spanish) × I(ce_t = 7)				0.962^{a}
				(0.080)
I(l = Spanish) × I(ce_t = 8)				0.888^{a}
				(0.140)
I(l = Spanish) × I(ce_t = 9)				1.149^{a}
				(0.143)
I(l = Spanish) × I(ce_t = 10)				0.490^{a}
				(0.080)
F-test of excluded instruments	91.66	35.60	45.59	44.93
[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}	0.090^{a}	0.076^{a}	0.076^{a}
	(0.017)	*(0.019)*	*(0.017)*	*(0.017)*
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*				
Proficiency in Catalan (Cat)	0.053^{a}	0.057^{a}	0.053^{a}	0.061^{a}
	(0.015)	*(0.015)*	*(0.015)*	*(0.016)*
Number of observations	5,357	5,357	5,357	5,357

Note: regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth-initial language 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 indicator for "full exposure" takes the value of 1 for individuals born in 1997 or after; the indicator for "partial exposure" takes the value of 1 for individuals born between 1970 and 1976 (included). 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 initial language.

Table 8: Falsification Analysis (Baseline and Placebo Reduced Form Equations)

	(a)	*(b)*	*(c)*	*(d)*
OLS (REDUCED FORM) — Dependent Variable: Partner's Language = Catalan-Only				
$I(l = Spanish) \times ce_t$	0.009^{a}	0.009^{a}	0.007^{a}	
	(0.002)	*(0.002)*	*(0.002)*	
$I(l = Spanish) \times ce_t*$		-0.002	-0.005	
		(0.017)	*(0.017)*	
$I(l = Spanish) \times ce_t**$				0.003
				(0.003)
Years since migration (squared)			*YES*	
Adjusted R^2	0.185	0.194	0.195	0.209
OLS (REDUCED FORM) — Dependent Variable: Language Used With the Partner = Catalan-Only				
$I(l = Spanish) \times ce_t$	0.006^{a}	0.006^{a}	0.005^{a}	
	(0.002)	*(0.002)*	*(0.002)*	
$I(l = Spanish) \times ce_t*$		0.017	0.014	
		(0.011)	*(0.012)*	
$I(l = Spanish) \times ce_t**$				-0.003
				(0.003)
Years since migration (squared)			*YES*	
Adjusted R^2	0.384	0.421	0.421	0.398
Number of observations	5357	5934	5884	3417

Note: regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth-initial language 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. (a) reduced form equation for the baseline sample; regressions in columns (b) and (c) are obtained from the pool of the baseline and the placebo samples, where the latter contains individuals who migrated from other Spanish regions after completing compulsory education. Placebo compulsory exposure (ce_t*) in columns (b) and (c) is imputed "as if" they received compulsory education in Catalonia. Regressions in columns (b) and (c) also include an indicator for belonging to the placebo sample and the respective interactions with control variables. Regressions in column (d) are based on a subsample of never-treated individuals (born between 1944 and 1969, in Catalonia or migrated before age 6); compulsory exposure (ce_t**) in column (d) is imputed "as if" the reform was applied 15 years before (i.e. in 1968 instead of 1983).*

Table 9a: 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
	(0.012)		*(0.018)*	*(0.018)*	*(0.017)*
I(l= parents Spanish speakers) × ce$_t$*		0.114^a	0.048^b	0.048^b	
		(0.013)	*(0.022)*	*(0.022)*	
$\varphi_{l*,t}$					*YES*
F-test of excluded instruments	91.66	73.95	44.33	17.45	19.45
[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^b
	(0.017)	*(0.015)*	*(0.016)*	*(0.069)*	*(0.069)*
I(l= parents Spanish speakers) × ce$_t$*				-0.010	
				(0.008)	
$\theta_{l*,t}$					*YES*
Hansen J test for overidentification			1.302		
[p-value]			*[0.254]*		
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*					
Proficiency in Catalan (Cat)	0.053^a	0.055^a	0.050^a	0.094	0.093
	(0.015)	*(0.013)*	*(0.016)*	*(0.061)*	*(0.063)*
I(l= parents Spanish speakers) × ce$_t$*				-0.006	
				(0.008)	
$\theta_{l*,t}$					*YES*
Hansen J test for overidentification			0.551		
[p-value]			*[0.458]*		
Number of observations	5,357	5,193	5,193	5,193	5,193

Note: ^{a} significant 1%; ^{b} significant 5%; ^{c} significant 10%. Standard errors (within parenthesis in italic) for the baseline model and in columns (b), (c) and (d) are adjuster for year of birth-initial language clusters, while standard errors in column (a) are adjusted for clusters at the year of birth-parental language (i.e. both parents are Spanish-only speakers) level. 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 initial language and the indicator for having Spanish-only speaking parents respectively.

Table 9b: 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
	(0.012)		*(0.019)*	*(0.019)*	*(0.018)*
I(l= non-Catalan origins) × ce$_t$*		0.131^a	0.071^a	0.071^a	
		(0.014)	*(0.026)*	*(0.026)*	
$\varphi_{l*,t}$					*YES*
F-test of excluded instruments	91.66	85.51	45.78	12.39	12.51
[p-value]	*[0.000]*	*[0.000]*	*[0.000]*	*[0.001]*	*[0.006]*
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
	(0.017)	*(0.020)*	*(0.019)*	*(0.048)*	*(0.048)*
I(l= non-Catalan origins) × ce$_t$*				-0.002	
				(0.007)	
$\theta_{l*,t}$					*YES*
Hansen J test for overidentification			0.042		
[p-value]			*[0.839]*		
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*					
Proficiency in Catalan (Cat)	0.053^a	0.089^a	0.051^a	0.014	0.010
	(0.015)	*(0.019)*	*(0.017)*	*(0.044)*	*(0.045)*
I(l= non-Catalan origins) × ce$_t$*				0.005	
				(0.007)	
$\theta_{l*,t}$					*YES*
Hansen J test for overidentification			0.663		
[p-value]			*[0.416]*		
Number of observations	5,357	5,357	5,357	5,357	5,357

Note: ^{a} significant 1%; ^{b} significant at 5%; ^{c} significant 10%. Standard errors (within parenthesis in italic) for the baseline model and in columns (b), (c) and (d) are adjuster for year of birth-initial language clusters, while standard errors in column (a) are adjusted for clusters at the year of birth-non-Catalan origins (i.e. both parents born outside Catalonia) level. 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 initial language and the indicator for non-Catalan origins respectively.

Table 10: 2SLS Estimates (Selected Results) — Subsample of Spanish Speakers

	(a)	*(b)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*		
I(l = non-Catalan origins) × ce_t*	0.096^a	0.097^a
	(0.023)	*(0.024)*
F-test of excluded instruments	16.84	16.06
[p-value]	*[0.000]*	*[0.000]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*		
Proficiency in Catalan (Cat)	0.047	0.097^b
	(0.038)	*(0.038)*
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*		
Proficiency in Catalan (Cat)	0.076^b	0.079^b
	(0.032)	*(0.036)*
Number of observations	2,396	2,396

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth-non-Catalan origins (i.e. both parents born outside Catalonia) 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 individuals whose parents are both Spanish-only speakers and year of birth dummies. Regression in column (b) excludes observations of individuals whose partner has both Catalan and Spanish as habitual/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.

APPENDIX

Table A1a: Linear Probability Model Estimates— Subsample of Native Spanish Speakers

	(a)		*(b)*		*(c)*		*(d)*	
OLS — Dependent Variable: Partner's Language = Catalan-Only								
constant	-1.463^{b}	*(0.583)*	-1.453^{a}	*(0.524)*	-1.156^{b}	*(0.564)*	-1.267^{b}	*(0.515)*
proficiency in Catalan (Cat)	0.044^{a}	*(0.003)*	0.040^{a}	*(0.003)*	0.036^{a}	*(0.003)*	0.035^{a}	*(0.003)*
wave 2013	-0.080^{a}	*(0.019)*	-0.077^{a}	*(0.019)*	-0.086^{a}	*(0.019)*	-0.080^{a}	*(0.019)*
age	0.133^{a}	*(0.047)*	0.125^{a}	*(0.042)*	0.108^{b}	*(0.046)*	0.112^{b}	*(0.042)*
age^{2}	-0.004^{a}	*(0.001)*	-0.004^{a}	*(0.001)*	-0.003^{a}	*(0.001)*	-0.003^{a}	*(0.001)*
age^{3}	0.000^{a}	*(0.000)*	0.000^{a}	*(0.000)*	0.000^{a}	*(0.000)*	0.000^{a}	*(0.000)*
male	0.046^{b}	*(0.019)*	0.048^{b}	*(0.020)*	0.047^{b}	*(0.019)*	0.048^{b}	*(0.020)*
father place of birth = Barcelona				*reference category*				
Girona			0.084	*(0.148)*			-0.002	*(0.148)*
Tarragona			0.051	*(0.103)*			0.031	*(0.109)*
Southern Catalonia (Terres de l'Ebre)			0.058	*(0.111)*			-0.013	*(0.100)*
Western Catalonia (Ponent)			0.013	*(0.100)*			-0.052	*(0.100)*
Central Catalonia			0.115	*(0.114)*			0.047	*(0.119)*
Pyrenees and Aran Valley			-0.169	*(0.126)*			-0.288^{b}	*(0.124)*
Balearic Islands and Valencia			-0.008	*(0.127)*			-0.043	*(0.121)*
Basque Country and Galicia			-0.029	*(0.068)*			-0.068	*(0.073)*
other Spanish regions			0.022	*(0.036)*			0.005	*(0.037)*
other places			0.125^{c}	*(0.074)*			0.100	*(0.077)*
miss father's place of birth			-0.069	*(0.043)*			-0.074	*(0.046)*
mother place of birth = Barcelona				*reference category*				
Girona			-0.064	*(0.134)*			-0.136	*(0.131)*
Tarragona			-0.100	*(0.087)*			-0.111	*(0.093)*
Southern Catalonia (Terres de l'Ebre)			0.213	*(0.141)*			0.144	*(0.125)*
Western Catalonia (Ponent)			-0.040	*(0.101)*			-0.104	*(0.101)*
Central Catalonia			0.079	*(0.111)*			0.013	*(0.113)*
Pyrenees and Aran Valley			0.128	*(0.178)*			0.087	*(0.178)*
Balearic Islands and Valencia			0.073	*(0.094)*			0.054	*(0.095)*
Basque Country and Galicia			0.067	*(0.064)*			0.038	*(0.063)*
other Spanish regions			0.040	*(0.029)*			0.024	*(0.028)*
other places			0.155	*(0.123)*			0.122	*(0.128)*
miss father's place of birth			-0.005	*(0.104)*			-0.005	*(0.107)*
Catalan used by parents = no Catalan				*reference category*				
Catalan used by father or mother			0.151^{a}	*(0.041)*			0.144^{a}	*(0.041)*
Catalan used by father and mother			0.171^{b}	*(0.082)*			0.167^{c}	*(0.084)*
missing parents' language			0.098	*(0.126)*			0.106	*(0.124)*
highest parental education = no education				*reference category*				
primary			0.042^{c}	*(0.022)*			0.044^{c}	*(0.024)*
secondary			0.055	*(0.034)*			0.045	*(0.038)*
tertiary			0.094^{c}	*(0.049)*			0.081	*(0.056)*
missing parental education			0.073	*(0.057)*			0.063	*(0.056)*
year of birth fixed effects (θ_t)	*YES*		*YES*		*YES*		*YES*	
adjusted R^{2}	0.071		0.081		0.092		0.098	
number of observations	2396		2396		2396		2396	

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%.

Table A1a (continued): Linear Probability Model Estimates— Subsample of Native Spanish Speakers

	(a)	(b)	(c)		(d)	
OLS — Dependent Variable: Partner's Language = Catalan-Only						
individual's place of birth = Barcelona			*reference category*			
Girona			-0.035	*(0.069)*	-0.018	*(0.066)*
Tarragona			0.003	*(0.047)*	0.003	*(0.048)*
Southern Catalonia (Terres de l'Ebre)			-0.177	*(0.130)*	-0.223^{c}	*(0.116)*
Western Catalonia (Ponent)			0.092	*(0.057)*	0.117^{b}	*(0.057)*
Central Catalonia			0.039	*(0.066)*	0.046	*(0.068)*
Pyrenees and Aran Valley			-0.063	*(0.124)*	-0.040	*(0.127)*
Balearic Islands and Valencia			0.239^{b}	*(0.105)*	0.218^{c}	*(0.109)*
Basque Country and Galicia			0.313^{b}	*(0.133)*	0.328^{b}	*(0.146)*
other Spanish regions			-0.034	*(0.038)*	-0.017	*(0.039)*
individual's place of residence = Barcelona city			*reference category*			
Barcelona's metropolitan area			-0.006	*(0.029)*	-0.000	*(0.027)*
Girona			0.147^{b}	*(0.062)*	0.147^{b}	*(0.063)*
Tarragona			0.010	*(0.050)*	0.016	*(0.052)*
Southern Catalonia (Terres de l'Ebre)			0.399^{a}	*(0.109)*	0.392^{a}	*(0.106)*
Western Catalonia (Ponent)			0.004	*(0.050)*	0.005	*(0.051)*
Central Catalonia			0.095	*(0.073)*	0.098	*(0.072)*
Pyrenees			0.254^{a}	*(0.087)*	0.252^{a}	*(0.089)*
completed education = primary or less			*reference category*			
secondary			0.028	*(0.025)*	0.020	*(0.026)*
tertiary			0.090^{a}	*(0.029)*	0.067^{b}	*(0.031)*
other education levels			-0.002	*(0.072)*	-0.023	*(0.075)*
year of birth fixed effects (θ_t)	*YES*	*YES*	*YES*		*YES*	
adjusted R^2	0.071	0.081	0.092		0.098	
number of observations	2396	2396	2396		2396	

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%.

Table A1b: Linear Probability Model Estimates— Subsample of Native Spanish Speakers

	(a)		*(b)*		*(c)*		*(d)*	
OLS — Dependent Variable: Language Used With the Partner = Catalan-Only								
constant	-1.141^{b}	*(0.445)*	-1.002^{b}	*(0.436)*	-0.608	*(0.402)*	-0.634	*(0.408)*
proficiency in Catalan (Cat)	0.040^{a}	*(0.002)*	0.037^{a}	*(0.002)*	0.029^{a}	*(0.002)*	0.027^{a}	*(0.002)*
wave 2013	-0.028^{c}	*(0.016)*	-0.031^{c}	*(0.016)*	-0.036^{b}	*(0.015)*	-0.034^{b}	*(0.016)*
age	0.073^{c}	*(0.037)*	0.056	*(0.036)*	0.029	*(0.034)*	0.028	*(0.033)*
age^2	-0.002^{c}	*(0.001)*	-0.001	*(0.001)*	-0.001	*(0.001)*	-0.001	*(0.001)*
age^3	0.000^{b}	*(0.000)*	0.000^{c}	*(0.000)*	0.000	*(0.000)*	0.000	*(0.000)*
male	0.006	*(0.015)*	0.012	*(0.014)*	0.009	*(0.013)*	0.011	*(0.013)*
father place of birth = Barcelona				*reference category*				
Girona			0.096	*(0.103)*			-0.059	*(0.102)*
Tarragona			0.014	*(0.076)*			-0.010	*(0.076)*
Southern Catalonia (Terres de l'Ebre)			0.073	*(0.143)*			-0.006	*(0.121)*
Western Catalonia (Ponent)			0.204^{b}	*(0.093)*			0.105	*(0.084)*
Central Catalonia			0.213^{b}	*(0.083)*			0.116	*(0.089)*
Pyrenees and Aran Valley			0.107	*(0.119)*			-0.015	*(0.112)*
Balearic Islands and Valencia			0.042	*(0.075)*			0.012	*(0.074)*
Basque Country and Galicia			0.105^{b}	*(0.044)*			0.058	*(0.044)*
other Spanish regions			0.048^{b}	*(0.021)*			0.018	*(0.022)*
other places			0.092	*(0.063)*			0.057	*(0.073)*
miss father's place of birth			-0.008	*(0.037)*			-0.027	*(0.040)*
mother place of birth = Barcelona				*reference category*				
Girona			-0.053	*(0.102)*			-0.193^{b}	*(0.094)*
Tarragona			-0.042	*(0.072)*			-0.062	*(0.071)*
Southern Catalonia (Terres de l'Ebre)			0.306^{b}	*(0.138)*			0.174	*(0.123)*
Western Catalonia (Ponent)			0.035	*(0.080)*			-0.064	*(0.087)*
Central Catalonia			0.018	*(0.083)*			-0.071	*(0.075)*
Pyrenees and Aran Valley			0.043	*(0.165)*			0.017	*(0.164)*
Balearic Islands and Valencia			0.016	*(0.058)*			0.018	*(0.059)*
Basque Country and Galicia			-0.086^{c}	*(0.046)*			-0.080	*(0.049)*
other Spanish regions			0.015	*(0.020)*			-0.013	*(0.019)*
other places			0.057	*(0.081)*			0.010	*(0.091)*
miss father's place of birth			0.006	*(0.062)*			-0.000	*(0.066)*
Catalan used by parents = no Catalan				*reference category*				
Catalan used by father or mother			0.139^{a}	*(0.030)*			0.126^{a}	*(0.030)*
Catalan used by father and mother			0.340^{a}	*(0.088)*			0.335^{a}	*(0.083)*
missing parents' language			-0.069^{c}	*(0.039)*			-0.047	*(0.038)*
highest parental education = no education				*reference category*				
primary			0.011	*(0.018)*			0.011	*(0.020)*
secondary			0.014	*(0.023)*			-0.002	*(0.026)*
tertiary			0.031	*(0.040)*			0.011	*(0.048)*
missing parental education			0.060^{c}	*(0.032)*			0.054^{c}	*(0.030)*
year of birth fixed effects (θ_t)	*YES*		*YES*		*YES*		*YES*	
adjusted R^2	0.086		0.118		0.154		0.173	
number of observations	2396		2396		2396		2396	

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%.

Table A1b (continued): Linear Probability Model Estimates— Subsample of Native Spanish Speakers

	(a)	*(b)*	*(c)*		*(d)*	
OLS — Dependent Variable: Language Used With the Partner = Catalan-Only						
individual's place of birth = Barcelona			*reference category*			
Girona			0.089^{b}	*(0.043)*	0.117^{b}	*(0.045)*
Tarragona			0.020	*(0.033)*	0.022	*(0.032)*
Southern Catalonia (Terres de l'Ebre)			-0.031	*(0.132)*	-0.126	*(0.125)*
Western Catalonia (Ponent)			0.137^{a}	*(0.038)*	0.139^{a}	*(0.041)*
Central Catalonia			0.104^{b}	*(0.051)*	0.113^{b}	*(0.048)*
Pyrenees and Aran Valley			-0.081	*(0.115)*	-0.088	*(0.119)*
Balearic Islands and Valencia			0.049	*(0.115)*	0.025	*(0.118)*
Basque Country and Galicia			-0.143^{a}	*(0.048)*	-0.105^{b}	*(0.047)*
other Spanish regions			0.001	*(0.020)*	0.019	*(0.021)*
individual's place of residence = Barcelona city			*reference category*			
Barcelona's metropolitan area			0.011	*(0.018)*	0.014	*(0.018)*
Girona			0.132^{b}	*(0.050)*	0.128^{b}	*(0.052)*
Tarragona			0.026	*(0.032)*	0.025	*(0.034)*
Southern Catalonia (Terres de l'Ebre)			0.428^{a}	*(0.096)*	0.412^{a}	*(0.093)*
Western Catalonia (Ponent)			0.061^{b}	*(0.030)*	0.054^{c}	*(0.031)*
Central Catalonia			0.093	*(0.064)*	0.085	*(0.062)*
Pyrenees			0.332^{a}	*(0.091)*	0.316^{a}	*(0.088)*
completed education = primary or less			*reference category*			
secondary			0.043^{b}	*(0.017)*	0.047^{b}	*(0.019)*
tertiary			0.143^{a}	*(0.024)*	0.135^{a}	*(0.026)*
other education levels			0.048	*(0.053)*	0.029	*(0.056)*
year of birth fixed effects (θ_t)	*YES*	*YES*	*YES*		*YES*	
adjusted R^2	0.086	0.118	0.154		0.173	
number of observations	2396	2396	2396		2396	

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%.*

Table A2: First-Stage Regressions

	(a)		*(b)*		*(c)*		*(d)*	
OLS — Dependent Variable: Proficiency in Catalan (Cat)								
constant	5.353^{a}	*(1.658)*	4.725^{a}	*(1.665)*	6.779^{a}	*(1.655)*	6.267^{a}	*(1.656)*
native Spanish speaker *(l = Spanish)*	-2.105^{a}	*(0.071)*	-1.556^{a}	*(0.103)*	-1.643^{a}	*(0.069)*	-1.376^{a}	*(0.102)*
$I(l = Spanish) \times ce_t$	0.115^{a}	*(0.012)*	0.115^{a}	*(0.012)*	0.105^{a}	*(0.011)*	0.104^{a}	*(0.011)*
wave 2013	0.126^{b}	*(0.060)*	0.072	*(0.062)*	0.046	*(0.059)*	0.026	*(0.062)*
age	0.367^{b}	*(0.141)*	0.320^{b}	*(0.142)*	0.161	*(0.142)*	0.159	*(0.143)*
age^{2}	-0.009^{b}	*(0.004)*	-0.008^{b}	*(0.004)*	-0.004	*(0.004)*	-0.004	*(0.004)*
age^{3}	0.000^{b}	*(0.000)*	0.000^{c}	*(0.000)*	0.000	*(0.000)*	0.000	*(0.000)*
male	-0.161^{a}	*(0.051)*	-0.164^{a}	*(0.050)*	-0.098^{b}	*(0.049)*	-0.111^{b}	*(0.049)*
father place of birth = Barcelona				*reference category*				
Girona			0.170^{c}	*(0.096)*			-0.157	*(0.128)*
Tarragona			0.151	*(0.108)*			-0.057	*(0.120)*
Southern Catalonia (Terres de l'Ebre)			0.165	*(0.109)*			-0.182	*(0.121)*
Western Catalonia (Ponent)			0.195^{b}	*(0.091)*			-0.110	*(0.108)*
Central Catalonia			0.209^{c}	*(0.107)*			-0.204	*(0.128)*
Pyrenees and Aran Valley			0.333^{b}	*(0.131)*			-0.052	*(0.144)*
Balearic Islands and Valencia			-0.282	*(0.344)*			-0.425	*(0.378)*
Basque Country and Galicia			-0.055	*(0.289)*			-0.285	*(0.272)*
other Spanish regions			0.063	*(0.094)*			-0.047	*(0.095)*
other places			0.250	*(0.230)*			0.189	*(0.230)*
miss father's place of birth			0.134	*(0.189)*			0.113	*(0.177)*
mother place of birth = Barcelona				*reference category*				
Girona			0.085	*(0.086)*			-0.255^{a}	*(0.096)*
Tarragona			0.178^{c}	*(0.101)*			-0.158	*(0.126)*
Southern Catalonia (Terres de l'Ebre)			0.170	*(0.104)*			-0.243^{c}	*(0.125)*
Western Catalonia (Ponent)			0.134^{c}	*(0.073)*			-0.217^{a}	*(0.079)*
Central Catalonia			0.007	*(0.101)*			-0.312^{a}	*(0.111)*
Pyrenees and Aran Valley			-0.101	*(0.153)*			-0.501^{b}	*(0.202)*
Balearic Islands and Valencia			0.050	*(0.175)*			0.061	*(0.168)*
Basque Country and Galicia			0.282	*(0.180)*			0.170	*(0.170)*
other Spanish regions			0.011	*(0.101)*			-0.119	*(0.099)*
other places			-0.211	*(0.281)*			-0.424	*(0.281)*
miss father's place of birth			-0.107	*(0.329)*			-0.210	*(0.326)*
Catalan used by parents = no Catalan				*reference category*				
Catalan used by father or mother			0.363^{a}	*(0.100)*			0.241^{b}	*(0.096)*
Catalan used by father and mother			0.426^{a}	*(0.108)*			0.302^{a}	*(0.108)*
missing parents' language			-0.094	*(0.437)*			-0.086	*(0.402)*
highest parental education = no education				*reference category*				
primary			0.466^{a}	*(0.096)*			0.332^{a}	*(0.088)*
secondary			0.716^{a}	*(0.117)*			0.427^{a}	*(0.101)*
tertiary			0.869^{a}	*(0.122)*			0.464^{a}	*(0.098)*
missing parental education			0.215	*(0.184)*			0.178	*(0.187)*
year of birth fixed effects (θ_t)	*YES*		*YES*		*YES*		*YES*	
adjusted R^{2}	0.204		0.221		0.264		0.270	
number of observations	2396		2396		2396		2396	

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%.

Table A2 (continued): First-Stage Regressions

	(a)	*(b)*	*(c)*		*(d)*	
OLS — ***Dependent Variable: Proficiency in Catalan (Cat)***						
individual's place of birth = Barcelona			*reference category*			
Girona			0.289^{b}	*(0.127)*	0.496^{a}	*(0.158)*
Tarragona			0.617^{a}	*(0.166)*	0.681^{a}	*(0.185)*
Southern Catalonia (Terres de l'Ebre)			0.255	*(0.163)*	0.438^{b}	*(0.185)*
Western Catalonia (Ponent)			0.332^{a}	*(0.120)*	0.489^{a}	*(0.139)*
Central Catalonia			0.222	*(0.177)*	0.483^{b}	*(0.202)*
Pyrenees and Aran Valley			0.124	*(0.141)*	0.444^{b}	*(0.206)*
Balearic Islands and Valencia			-0.753	*(0.502)*	-0.730	*(0.503)*
Basque Country and Galicia			-0.292	*(0.725)*	-0.253	*(0.658)*
other Spanish regions			-0.580^{a}	*(0.172)*	-0.444^{b}	*(0.182)*
individual's place of residence = Barcelona city			*reference category*			
Barcelona's metropolitan area			-0.169^{b}	*(0.084)*	-0.140^{c}	*(0.084)*
Girona			0.160	*(0.146)*	0.189	*(0.152)*
Tarragona			-0.215	*(0.171)*	-0.144	*(0.174)*
Southern Catalonia (Terres de l'Ebre)			0.329^{b}	*(0.159)*	0.441^{b}	*(0.175)*
Western Catalonia (Ponent)			0.144	*(0.120)*	0.186	*(0.121)*
Central Catalonia			0.267	*(0.164)*	0.318^{c}	*(0.161)*
Pyrenees			0.365^{a}	*(0.112)*	0.388^{a}	*(0.120)*
completed education = primary or less			*reference category*			
secondary			0.686^{a}	*(0.106)*	0.615^{a}	*(0.098)*
tertiary			1.206^{a}	*(0.139)*	1.089^{a}	*(0.133)*
other education levels			0.740^{a}	*(0.202)*	0.660^{a}	*(0.203)*
year of birth fixed effects (θ_t)	*YES*	*YES*	*YES*		*YES*	
adjusted R^2	0.204	0.221	0.264		0.270	
number of observations	2396	2396	2396		2396	

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%.

Table A3: 2SLS Estimates (selected results) — Separate Estimations by Gender

	baseline	*males*	*females*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*			
I(l = Spanish) × ce_t	0.115^a	0.098^a	0.128^a
	(0.012)	*(0.015)*	*(0.015)*
F-test of excluded instruments	91.66	41.07	77.91
[p-value]	*[0.000]*	*[0.000]*	*[0.000]*
2SLS — *Dependent Variable: Partner's Language = Catalan-Only*			
Proficiency in Catalan (Cat)	0.076^a	0.070^c	0.081^a
	(0.017)	*(0.036)*	*(0.019)*
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*			
Proficiency in Catalan (Cat)	0.053^a	0.034	0.069^a
	(0.015)	*(0.029)*	*(0.022)*
Number of observations	5,357	2,611	2,746

Note: 2SLS regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth-native language clusters. ^{a} Significant at 1%; ^{b} significant at 5%; ^{c} significant at 10%. All regressions include a dummy for wave, a cubic age polynomial, an indicator for being native Spanish speaker and year of birth dummies; baseline regressions also contain a gender indicator. The F-test on excluded instruments refers to the Angrist-Pischke multivariate F-test on the interactions between years of exposure to Catalan at compulsory schooling and the indicator for Spanish as initial language.

Table A4: Sensitivity to Age Polynomial's Order

	(a)	*(b)*	*baseline*	*(c)*	*(d)*
FIRST STAGE — *Dependent Variable: Proficiency in Catalan (Cat)*					
$I(l = Spanish) \times ce_t$	0.114^{a}	0.115^{a}	0.115^{a}	0.116^{a}	0.116^{a}
	(0.012)	*(0.012)*	*(0.012)*	*(0.012)*	*(0.012)*
Age	-0.007^{c}	0.053	0.367^{b}	0.651	
	(0.004)	*(0.044)*	*(0.141)*	*(0.452)*	
Age^{2}		-0.001	-0.009^{b}	-0.020	
		(0.001)	*(0.004)*	*(0.019)*	
Age^{3}			0.000^{b}	0.000	
			(0.000)	*(0.000)*	
Age^{4}				-0.000	
				(0.000)	
Age dummies	*NO*	*NO*	*NO*	*NO*	*YES*
Adjusted R^{2}	0.203	0.203	0.203	0.203	0.204
F-test of excluded instruments	86.95	89.03	91.66	91.90	92.31
[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}	0.075^{a}	0.076^{a}	0.075^{a}	0.076^{a}
	(0.018)	*(0.017)*	*(0.017)*	*(0.017)*	*(0.018)*
Age	-0.024^{a}	-0.009	0.039	-0.019	
	(0.002)	*(0.010)*	*(0.035)*	*(0.113)*	
Age^{2}		0.000	-0.001	0.001	
		(0.000)	*(0.001)*	*(0.004)*	
Age^{3}			0.000	-0.000	
			(0.000)	*(0.000)*	
Age^{4}				0.000	
				(0.000)	
Age dummies	*NO*	*NO*	*NO*	*NO*	*YES*
Adjusted R^{2}	0.201	0.202	0.201	0.202	0.200
2SLS — *Dependent Variable: Language Used With the Partner = Catalan-Only*					
Proficiency in Catalan (Cat)	0.053^{a}	0.053^{a}	0.053^{a}	0.053^{a}	0.053^{a}
	(0.015)	*(0.015)*	*(0.015)*	*(0.015)*	*(0.015)*
Age	-0.000	-0.004	-0.002	0.013	
	(0.002)	*(0.010)*	*(0.032)*	*(0.106)*	
Age^{2}		0.000	0.000	-0.001	
		(0.000)	*(0.001)*	*(0.004)*	
Age^{3}			0.000	0.000	
			(0.000)	*(0.000)*	
Age^{4}				-0.000	
				(0.000)	
Age dummies	*NO*	*NO*	*NO*	*NO*	*YES*
Adjusted R^{2}	0.406	0.406	0.406	0.406	0.407
Number of observations	5,357	5,357	5,357	5,357	5,357

Note: regression estimates with standard errors (within parenthesis in italic) adjusted for year of birth-initial language clusters. ^{a} *Significant at 1%;* ^{b} *significant at 5%;* ^{c} *significant at 10%. All regressions include dummies for wave and gender, 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 initial language.*