Departament d'Economia Aplicada

Language knowledge and earnings in Catalonia

Antonio Di Paolo,
Josep Lluís Raymond-Bara

DOCUMENT DE TREBALL

10.01

Universitat Autònoma de Barcelona

Facultat de Ciències Econòmiques i Empresarials

Aquest document pertany al Departament d'Economia Aplicada.

Data de publicació : **Febrer 2010**

Departament d'Economia Aplicada
Edifici B
Campus de Bellaterra
08193 Bellaterra

Telèfon: (93) 581 1680
Fax:(93) 581 2292
E-mail: d.econ.aplicada@uab.es
http://www.ecap.uab.es

LANGUAGE KNOWLEDGE AND EARNINGS IN CATALONIA

Antonio Di Paolo, Josep Lluís Raymond

Universitat Autònoma de Barcelona & Institut d'Economia de Barcelona (IEB)

Abstract

This paper investigates the economic value of Catalan knowledge for national and foreign first- and second-generation immigrants in Catalonia. Specifically, drawing on data from the "Survey on Living Conditions and Habits of the Catalan Population (2006)", we want to quantify the expected earnings differential between individuals who are proficient in Catalan and those who are not, taking into account the potential endogeneity between knowledge of Catalan and earnings. The results indicate the existence of a positive return to knowledge of Catalan, with a 7.5% increase in earnings estimated by OLS; however, when we account for the presence of endogeneity, monthly earnings are around 18% higher for individuals who are able to speak and write Catalan. However, we also find that language and education are complementary inputs for generating earnings in Catalonia, given that knowledge of Catalan increases monthly earnings only for more educated individuals.

Keywords: Language, Earnings, Immigrants, Endogeneity, Complementarity

JEL classifications: J79, J24, J61, C31

Corresponding author: Antonio Di Paolo; e-mail: antonio.dipaolo@uab.cat, tel.: +34935813415; fax: +34935812292. Departament d'Economia Aplicada, Universitat Autònoma de Barcelona (UAB); Campus de Bellaterra, Edifici B 08193 Bellaterra (Cerdanyola), Spain. Institut d'Economia de Barcelona, Universitat de Barcelona.

Josep Lluís Raymond; e-mail: josep.raymond@uab.cat. Departament de Fonaments de l'Anàlisi Econòmic, Universitat Autònoma de Barcelona (Spain). Institut d'Economia de Barcelona, Universitat de Barcelona.

Acknowledgments: we thank Marta Masats (IDESCAT), Josep-María Martínez (IDESCAT) and Elisenda Vila (IDESCAT) for their special effort in supporting us with the data. We also thank Ada Ferrer and the participants in the applied economic seminar and in the IEB seminar for useful comments and suggestions. Antonio Di Paolo gratefully acknowledges the financial support from the Institut d'Estudis Catalans (IEC) grant "Borsa d'Estudi Generalitat de Catalunya"; Josep Lluís Raymond acknowledges the financial support from the MEC grant ECO 2009-12234. Any view expressed herein and remaining errors are complete responsibility of the authors.

1. Introduction

Since the late 1970s, after the end of the dictatorship (1939-1975), Catalan has been a co-official language in Catalonia, together with Castilian (more commonly known as Spanish). The Franco regime prohibited the use of Catalan in public and strongly disapproved of its use in private. Nevertheless, among the native population (of Catalan origin) the Catalan language was transmitted and used within the family, and this intergenerational transmission has guaranteed the survival of the language up to the present day. Moreover, the permanence of Catalan as the language of identification

among the native population has supported the implementation of public linguistic policies in Catalonia.

The range of public linguistic policies implemented by the Catalan government (the *Generalitat*) aimed to re-establish the use of Catalan in public, and to stimulate its study and its use in private. The first important attempt in this direction was the "Linguistic Normalization Act" of 1983 (*Llei de Normalització Lingüística*). This new legislation established that Catalan was not only to be the official language of the Catalan government and of the local public administrations, but also the main language used in primary and secondary education[1]. More relevant to the economic value of knowledge of Catalan was the "Linguistic Policy Act" of 1998 (*Llei de Política Lingüística*), which attempted to reassert the presence of Catalan (versus Castilian), by a) increasing fluency requirements for public sector employees, and b) introducing major incentives (and in some case requirements) for increasing the use of Catalan in private business and other socioeconomic and cultural domains[2] (Solé and Alarcón 2001).

However, despite these public policies, knowledge of Catalan remains limited in a significant proportion of the Catalan population today[3]. In fact, given the coexistence of Catalan and Castilian and the absence of any legal requirement with respect to proficiency in Catalan (except with respect to work-related issues), many individuals have been able to maintain or use Castilian as their habitual language. Figure 1 illustrates the evolution of the Catalan population and the percentage of the population that speak the language (i.e., have basic knowledge) and speak and write it (i.e., have advanced knowledge). This information offers an intuitive picture of the sociolinguistic context of Catalonia and justifies the claim that there may exist an economic return to knowledge of Catalan.

Even before the "normalization" phase (that is, after the legislation introduced in 1983), Catalonia already had a large immigrant population, made up of people from other Spanish regions who had arrived since the 1960s. Given the subsequent

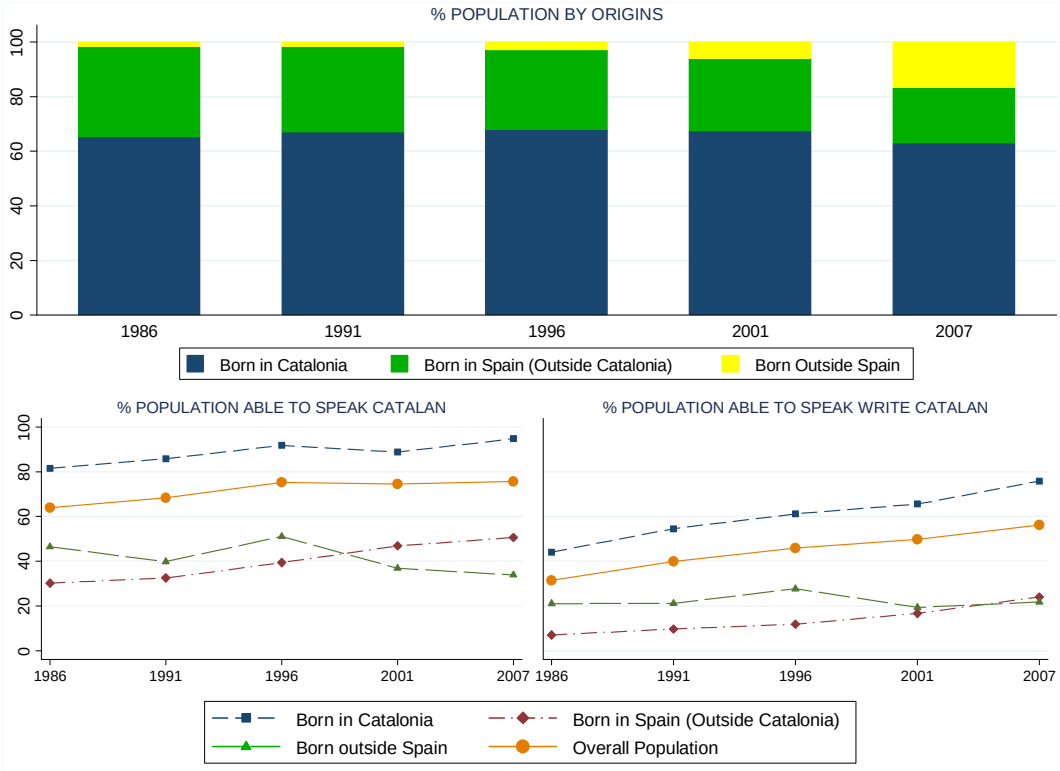
[1] In fact, Spanish (or Castilian) is taught as a second language in pre-university education. At university the language used is not determined by law, and is established by the professor.

[2] With the 1998 Act, private suppliers of public services have been subjected to similar linguistic requirements as the public sector. Moreover, the Catalan government has introduced economic incentives for "normalizing" Catalan in private firms, and stimulating active learning of this language for workers; finally, the government also promoted the language through the mass media by introducing incentives in the use of Catalan on radio, TV and also in the newspapers and written publications in general.

[3] In general, all the natives are fully proficient in Spanish; moreover, Catalan and Spanish are relatively close, though not enough to enable people to communicate if the Spanish speaker does not know Catalan. In addition, it is reasonable to assume that foreign immigrants (that is, those that do not come from Spanish-speaking countries) prioritize Spanish over Catalan, because of its greater international value; this means that the vast majority of foreign immigrants who know Catalan are also fluent in Spanish.

coexistence of Spanish (their mother-tongue) and Catalan, many of these immigrants may not have achieved a high level of fluency in the latter language (due to the lack of incentives), by the time it was recognized as co-official. In fact, the proportion of individuals coming from other Spanish regions with advanced knowledge of Catalan is extremely low throughout the entire period.

Figure 1: Linguistic Census over time[a]



[a] *Data for the period 1986-2001 are taken by the Statistical Census of the population; the information for 2007 is taken by the Demographic Survey, carried out jointly by the National and the Catalan Statistical Institutes (INE and IDESCAT respectively).*

Moreover, a large proportion of the current native-born population are second-generation immigrants from elsewhere in Spain. In general, their mother-tongue is Castilian, and their knowledge of Catalan may be limited, especially in the case of those who received their schooling before the 1983 legislation. This explains why the percentage of individuals born in Catalonia who speak and write Catalan is appreciably lower than one hundred (even though it increases steadily over the entire period)[4]. Finally, in the last decade Catalonia has experienced new immigration flows from other countries (Fernández-Huerta and Ferrer-i-Carbonell 2007). Many immigrants are from

[4] In fact, as we illustrate below with the descriptive statistics, almost 90% of native-born individuals of Catalan origin (that is, native-born who are not second-generation immigrants) are able to speak and write Catalan, mainly because, as commented above, even before "normalization" the language was transmitted and used within the family.

Spanish-speaking countries in Latin America, but Catalan is a new language for all the foreign immigrants and fluency in Catalan is often considered as a possible source of cultural and economic assimilation for them (Pujolar 2009).

Given the institutional setting, it seems fair to say that for individuals who are not of Catalan origin (first- and second-generation immigrants) learning and using Catalan is much more a choice rather than a requirement or a need. The decision to learn the language choice depends on several factors such as individual characteristics — origin, human capital, ability, etc. — but also on many other aspects related to language use and attitudes in daily life (i.e. the neighbourhood/environment, the family, and other habits related to exposure to Catalan). Moreover, economic expectations concerning the attainment of linguistic proficiency may represent an important incentive to learn Catalan. In other words, non-Catalans may view learning Catalan as an investment, given that being proficient in this language in Catalonia should improve their labour market opportunities — that is, it may provide better chances of finding work, wider occupational choice and, in all likelihood, the prospect of higher earnings.

In this paper we aim to quantify the return to this linguistic investment, specifically, by measuring the earnings return to proficiency in Catalan (that is, being able to speak and write the language) for first- and second-generation immigrants. However, to obtain an estimate that reflects the true economic value of Catalan knowledge (expressed in terms of the earning premium), we must take into account the potential endogeneity of proficiency in Catalan in the earnings equation: that is, we need to rule out the potential effect of unobserved factors which simultaneously affect the propensity to speak and write Catalan and the unexplained component of the earnings equation. As discussed in detail below, we deal with this potential endogeneity by applying two alternative and parallel methodologies (self-selection methods and Two-Stage Least Square).

Finally, we also consider the potential complementarity between level of schooling and knowledge of Catalan. In fact, in order to obtain the language premium, an individual may first need to have achieved a given level of formal education, which is the fundamental mechanism for gaining access to the occupations where knowledge of the language might be considered a valuable asset. We therefore repeat the estimations separating the sample into low- and high-educated individuals, in order to allow for a different endogeneity process in the two groups. Obtaining a positive return only for high-educated individuals would represent an important message for Catalan policy-makers. Indeed, obtaining a positive earnings return would mean that the linguistic

policies have significant socio-economic implications (which may not have been explicitly contemplated). However, if this return exists only for more educated individuals, the linguistic policies should be reconsidered, and perhaps accompanied or complemented by educational policies aimed at the least advantaged sector of the population.

This paper proceeds as follows: in the next section we briefly review some relevant contributions in the literature in order to contextualise our research. In section 3 we describe the data used in the estimations, and discuss the empirical specification and the identification issues. Section 4 presents the basic results (complemented by sensitivity analyses), followed by evidence of the language-education complementarity. Finally, section 5 concludes.

2. Related Research

As a general framework for analysing the value of knowledge of Catalan in the labour market, we refer to research on the earnings returns to language knowledge, with the pioneer studies of Chiswick (1991), and Chiswick and Miller (1995, 1999, 2007). These authors consider that linguistic proficiency constitutes a form of human capital alongside education and labour market experience. This is because language knowledge is likely to be person-embodied; it is productive in the labour market, and it is obtained via individual effort and out-of-pocket time and economic resources. But what is still more important is that Chiswick and Miller explicitly consider the potential endogeneity between earnings and language proficiency: specifically, they claim that the (positive) correlation between language fluency and earnings (among immigrants) may only indicate that more able workers are more likely to earn more and speak their host-country language better. They try to correct for this potential endogeneity using instrumental variables and sample selection correction methods, finding positive (but unstable) effects of language proficiency among (first-generation) immigrants.

The same framework can be applied to Catalonia. However, for the reasons explained above, we consider that proficiency in Catalan may have an economic return for first-generation national and non-national immigrants, but also for second-

generation immigrants (native-born population with non-Catalan parents)[5]. The first study focusing on the case of Catalonia is the paper by Rendon (2007), which analyzes the potential premium of knowing Catalan on the probability of being employed. Using Census data from 1991 and 1996, Rendon found that being proficient in Catalan (considering speaking and writing capabilities separately) has a positive effect on the probability of being employed among native-born and national immigrants. Rendon's analysis takes into account the presence of self-selection into Catalan knowledge, using as instruments the externality effects of the place of residence, origin variables, and two other variables indicating whether the individual arrived in Catalonia at age 10 or younger and whether he/she was affected by the 1983 legislation (that is, whether he/she may have been educated in Catalan). Unfortunately, as the author recognizes, the study is limited to language return on employment probabilities, since the Census data do not contain information on earnings. Moreover, the modest impact that he found (between 2 and 5% in the case of speaking, and between 3 and 6% in the case of writing), may be due to the fact that the data are limited to 1996, that is, before the implementation of the Linguistic Policy Act in 1998. Conceivably, data compiled more recently may reflect that this legislation has increased the economic value of knowledge of Catalan.

Many other recent studies focus on obtaining an (unbiased) estimate of the effect of language proficiency on earnings. As noted above, there is general concern that an OLS estimate of the effect of language knowledge on earnings could be biased, due to the presence of unobserved individual heterogeneity (which may affect both language fluency and economic outcomes). Moreover, several authors suggest that misclassification/measurement error represents another potential source of bias in the earnings equation, given that (as in this case) language proficiency measures are usually self-reported on a categorical scale.

For example, in order to account for misclassification error and unobserved heterogeneity, Dustmann and Van Soest (2001) used a panel structure of the GSOEP, and in a later study (Dustmann and Van Soest 2002), combined a panel estimation with

5 Chiswick and Miller (2007), among others, consider the issue of limited language knowledge of immigrants of the second generation (children of immigrant parents). Other studies (see Lindley 2002, for instance) consider of bilingualism (with respect to local or disglossic language), which seems to have a negative effect on earnings. Nevertheless, it is not clear whether (within the current institutional framework) Catalan might be still considered a disglossic language (that is, a language used only in informal communications).

IV techniques[6]. These authors suggest that language knowledge has a positive earning effect among immigrants in Germany, which tends to be underestimated by OLS; in fact, their simplest OLS estimates suggest a language return on wages of around 5%, whereas their estimated return range between 11% and 14% when they try to correct the biases in the estimates. Moreover, they also point out that the negative bias of measurement error overcomes the positive bias of unobserved heterogeneity. In a subsequent study, Dustmann and Fabbri (2003) combine the use of the interview language as an instrument for the measurement error in language proficiency, and the propensity score matching for correcting the potential individual heterogeneity bias. They confirm that the two sources of bias operate in opposite directions, but their final results are inconclusive.

Bleakley and Chin (2004) they propose an interesting IV approach, based on the idea that younger immigrants learn their host-country language (English) easily, but that immigrants from English-speaking countries do not need to learn a new language. They suggest that the OLS estimates are slightly upward biased by the endogeneity between earnings and language knowledge, and downward biased by measurement error. Moreover, they argue that an important component of language proficiency effects on earnings is mediated by education; specifically, young immigrants who are more likely to be proficient in English will probably reach higher levels of education attainment, thus increasing their earning opportunities.

In this paper we use data from the 2006 wave of the "Survey on Living Conditions and Habits of the Catalan Population" (*Enquesta de Condicions de Vida i Hàbits de la Població*, ECVHP06), carried out by the Statistical Institute of Catalonia (IDESCAT)[7]. This database is very attractive for our purposes, because it contains socioeconomic information (including individual monthly earnings), and also includes information related to language knowledge and use. Nevertheless, its cross-sectional structure entails the absence of longitudinal information and reduces the possibilities of accounting for misclassification into the self-reported knowledge of Catalan variable — categorized into four ordinal levels: unable to understand the language, understand it but not be able to speak it, able to speak it, and able to speak and write it. In order to limit the potential misclassification/measurement error bias, we are forced to adopt a

[6] Specifically, Dustmann and Van Soest (2002) exploit lagged self-reported language fluency (for time-independent measurement error), partner and household characteristics (for unobserved heterogeneity) and parental education (for time-persistent measurement error).

[7] We are very grateful to IDESCAT for providing the data. IDESCAT bears no liability for any mistakes or opinions expressed in this paper, which are the sole responsibility of the authors.

stringent definition of language proficiency, considering only individuals who are able to speak and write Catalan as proficient[8].

In contrast, we explicitly deal with the endogeneity bias by comparing the results obtained using two alternative methodologies. First, we compute the earnings return to knowledge of Catalan, accounting for the determinants of self-selection into Catalan proficiency using the methodology proposed by Lee (1978) and Heckman (1979), also known as the Endogenous Switching Model. We specify the first stage equation as a reduced-form equation for explaining the propensity to being able to speak and write Catalan. The variables included represent a satisfactory approximation of the observable determinants of language knowledge (see the next section for details). On the one hand, some of these variables are clearly shared with the second-stage earnings equation (i.e. education, origins, gender, etc.). On the other hand, other variables are specific determinants of Catalan proficiency (i.e. language use within the family, neighbourhood effects, and other elements that capture exposure to the language).

However, we need to identify which of these variables can be reasonably used as exclusion restrictions, in order to achieve the full identification of the language premium. Therefore, we also compute the language return from the Two-Stage Least Square estimation, and we check for the validity of the exclusion restrictions by implementing the formal tests for instruments' validity. We are able to obtain almost the same estimated return from the Endogenous Switching Model (exploiting the entire set of language determinants) as from the TSLS regression, using as instruments only the sub-set of language determinants that pass (with a conservative level of statistical significance) the overidentification and weak instruments tests. This result can be taken as evidence of the reliability of our findings, and it also makes more reasonable the untestable assumption that a linear combination of our selected instruments affects individual's earnings only through Catalan knowledge — i.e. orthogonal with the error term of the earnings equation. Moreover, we also check for the sensitivity of the return to proficiency in Catalan to different sample specifications.

Finally, as briefly mentioned above, we suspect that there are significant complementarities between fluency in Catalan and individual skills. For example, more educated individuals may face lower costs of achieving proficiency. Moreover, given the Catalan institutional setting, only more educated individuals may obtain

[8] The justification for this strategy is that, given that individuals are likely to over-report their true language fluency, we can reasonably assume that individuals who claim to be able to speak and write Catalan have at least good or satisfactory speaking ability, even if their writing competence is only basic.

occupations where full proficiency in Catalan is really important. Therefore, more skilled workers may also have higher benefits, which imply higher returns to language fluency. Several studies confirm this intuition in other contexts; Carliner (1996) and Mora and Davila (1998) suggested the idea of non-constant language returns across skill levels, finding that the wage penalty of limited language proficiency is higher for more educated workers.

Berman et al. (2003) ruled out unobserved heterogeneity bias from the language augmented earning regression through differentiation with panel data. They suggest that individual fixed-effects (interpreted as unobserved individual ability) have a small effect on language returns in high-skilled occupations. However, the small positive cross-sectional return for low-skilled workers turns out to be insignificant when individual fixed-effects are introduced (that is, the estimated language effect for low-skilled workers is accounted for entirely by the ability bias). Moreover, the recent study by Lang and Siniver (2009) confirms the previous findings of complementarity between language and skills, obtaining a higher return to language knowledge among high-educated workers (and an insignificant return for low-educated workers), in both cross-sectional and longitudinal estimates.

Therefore, following on from this recent literature, in this paper we also check for potential complementarities between knowledge of Catalan and completed schooling. Specifically, we repeat the estimations dividing the sample into low-educated (individuals with eight years of schooling or less), and high-educated (individuals with more than eight years of schooling), in order to capture differences both in the earnings return and in the endogeneity process for low- and high-educated individuals.

3. Data, Descriptive Statistics and Empirical Specification

The empirical analysis is based on the data from the 2006 “Survey of Living Conditions and Habits of the Catalan Population (ECVHP06)”, carried out by the Statistical Institute of Catalonia (IDESCAT). The original sample comprised 10,358 observations of individuals aged 16 or more and residing in Catalonia. The survey compiled socioeconomic and demographic information on the overall population, at individual and family level. The data were collected between the fourth trimester of 2005 and the third trimester of 2006; therefore, the information about individual labour

market status and monthly earnings reflects the situation in 2005-2006. Analyzing this period is very appealing for our purposes, since the unemployment rate was exceptionally low (6.6%)[9]; this means that we can focus only on the employed population, as we consider that neglecting for potential self-selection into employment should not be problematic during this period/[10]. Moreover, as the female employment rate was sufficiently high (52.3%), women could be included in the analysis. The final sample used consisted in 2,582 observations of the whole of individuals aged 16 to 65 in regular employment, with valid information on earnings (collected in brackets). Also, in order to consider only first- and second-generation immigrants, we dropped the observation of individuals with at least one parent born in Catalonia (see Table 1A in the Appendix for details).

The information about knowledge of Catalan is reported in four categories: namely, an individual may claim he/she "does not understand", "understand but is unable to speak", "is able to speak but not to write", and "is able to speak and write" Catalan. Table 1 contains the relative frequencies of each category of knowledge of Catalan in the selected sample; we report and compare the expected frequencies of language knowledge for first- and second-generation immigrants (distinguishing between national and foreign immigrants), and for the working native-born population of Catalan origin (excluded from the main analysis).

Table 1: Knowledge of Catalan

PERCENTAGES	*CATALAN ORIGINS*	***SELECTED SAMPLE***	***IMMIGRANTS (SEC. GEN.)***	***NATIONAL IMMIGRANTS (FIRST GEN.)***	***FOREIGNER IMMIGRANTS (FIRST GEN.)***
Do not understand	—	5.39	0.38	2.12	11.58
Understand but unable to speak	*0.93*	31.71	11.93	35.06	13.34
Able to speak but not to write	*12.19*	20.50	17.37	30.12	57.92
Able to speak and to write	*86.88*	42.40	70.32	32.71	17.16
TOTAL (#)	*2,912*	2,582	1,048	850	684

Source: ECVHP06.

[9] Note that the value reported is the mean unemployment rate between the fourth trimester of 2005 and the third trimester of 2006. The information is taken from the EPA (*Encuesta de la Población Activa*, Active Population Survey (INE)) for Catalonia. This was the lowest unemployment rate in Catalonia since 1978.

[10] We also tried to estimate a Probit model for employment, but it performed very badly due to the extremely low number of zeros (unemployed individuals).

This descriptive evidence shows that, as expected, almost all the native-born individuals of Catalan origin are able to speak and write Catalan[11], probably due to intergenerational transmission of the language within the family. However, the information from the "Linguistic Census" presented above shows that knowledge of Catalan is far from being widespread among the immigrant population. In fact, only 42% of the selected sample (first- and second-generation immigrants) claimed to be able to speak and write the language. More specifically, children of immigrants who were born in Catalonia show higher linguistic integration than the first-generation immigrants, since 70% of them claim to be able to speak and write the local language; only 32% of national immigrants report the same degree of linguistic competence, a proportion that falls to 17% for foreign first-generation immigrants.

We now move on to analyse the differences in the expected monthly earnings between individuals who are proficient in Catalan and those who are not. As noted above, individuals tend to over-report their linguistic ability when responding to survey questions. Therefore, in order to minimize the potential misclassification error, we adopt a stringent definition of proficiency in Catalan, considering only individuals who claim to be able to speak and write Catalan to be fully proficient. In Table 2, we report the mean monthly earnings[12] according to proficiency in Catalan; this simple descriptive picture indicates that being proficient in Catalan is associated with higher monthly earnings. Moreover, it seems that the earnings penalty associated with a limited knowledge of Catalan is higher for first-generation immigrants, especially for those from other Spanish regions.

Table 2: Mean monthly earnings and Catalan proficiency

	MEAN MONTHLY EARNINGS			
	OVERALL SAMPLE	***SECOND GENERATION***	***NATIONAL FIRST GEN.***	***FOREIGNER FIRST GEN.***
Proficient	1181.61	1128.21	1368.82	1052.84
Not Proficient	1057.54	1112.67	1047.37	947.31

Source: ECVHP06. Proficient: individuals who declare to be able to speak and write Catalan.

However, we need to take earnings covariates into account, given that our main purpose is to obtain the *ceteris-paribus* impact of knowledge of Catalan on monthly

[11] The 13% of individuals of Catalan origin who state that they cannot write Catalan can be entirely accounted for by an age-cohort effect; that is, they are individuals who were schooled in Spanish and, for some reason, may not have achieved full writing competence in Catalan.

[12] Given that the information about individuals' monthly earnings is shown in brackets (see table 2A in the appendix for details), the mean earnings are computed using the Interval Regression. However, in the rest of the paper we use a linear model for explaining monthly earnings, generating a continuous dependent variable as the mean point of each interval.

earnings. Therefore, we start the analysis with a simple OLS estimation of a Mincer-type equation to explain an individual's monthly earnings (in logs). Our basic empirical model, like many others used in the literature, contains socio-demographic information (gender, marital status, indicators of origin and years since migration for those born abroad), years of education, current job-tenure (in months), previous potential experience and the number of hours of work per month; we also include three indicators related to mode of employment – self-employment, membership of a labour union, or employment in a firm with more that 500 workers – and another indicator for living in the Metropolitan Area of Barcelona[13]. We then add to the model the linguistic knowledge indicator, whose coefficient indicates the percentage of increase in earnings associated with proficiency in Catalan. In general, the OLS coefficient of the language return can be taken as a correct estimation of the economic value of knowledge of Catalan only under the restrictive hypothesis that the individual's propensity to know Catalan is uncorrelated with the unexplained earnings determinants.

However, as shown in the literature, language knowledge is likely to be an endogenous variable, correlated with the error term of the earnings equation. We need to deal with this potential endogeneity in order to estimate correctly the return to knowledge of Catalan; we try to rule out the endogeneity bias in the estimation with two alternative methods. First, we specify a reduced-form model for explaining the decision of achieving speaking and writing fluency in Catalan, which enables us to account for the potential self-selection into Catalan proficiency with the method proposed by Lee (1978) and Heckman (1979) (also known as the Endogenous Switching Model).

Following Chiswick and Miller (1995, 2001, 2007), the Probit model for explaining proficiency in Catalan includes as regressors a gender indicator, individuals' age, origin variables[14] and the duration of the stay in Catalonia (represented by the years since migration). Moreover, we consider that the likelihood of being able to speak and write Catalan depends on schooling, but also on the cultural resources at home (number of books at home and weekly frequency of reading). As in Rendon's study (2007), our model also includes indicators for young age on arrival (arriving before the age of 10) and for exposure to the Linguistic Normalization of 1983 (that is, being schooled in

[13] Table 3A in the Appendix reports the full definition of the explanatory variables and some descriptive statistics; the Appendix also describes how some of these variables have been constructed.

[14] We consider four main origins' groups for foreigners immigrants, which is the only information provided in the database: namely, we include dummies for those born in Europe, South America, Africa and other countries (mainly Asia). Moreover, we consider five different territorial groups for national immigrants (see the Appendix for details).

Catalan) distinguishing between full exposure (i.e. those educated entirely under the new system) and partial exposure (i.e. those already in school when the reform was implemented). We also include in the model the percentage of individuals who speak and write Catalan in the neighbourhood of residence[15], in order to capture the territorial externalities in language knowledge. Finally, the information contained in the survey enable us to consider the role of language usage in the family (speaking Catalan with the parents, at home, or with the children) and of the exposure to local media (reading the newspaper in Catalan and watching the newscast on Catalan television) as determinants of knowledge of Catalan (capturing the intensity of the exposure to the language).

The joint Maximum Likelihood of the earnings equation with the endogenous indicator for language knowledge and the equation that explains proficiency in Catalan provides an estimate of the correlation coefficient between the error terms of the two equations. Its statistical significance indicates whether, and in which direction, the unobservable determinants of proficiency in Catalan are related to the unexplained earnings' determinants; in other words, it indicates if the selection on unobservable matters. If this is the case, a simple OLS estimation of the language return would be biased, and accounting for the self-selection process behind knowledge of Catalan (and its relationship with earnings) is essential for obtaining a correct estimation of the economic value of knowledge of Catalan.

Second, we try to control for the same endogeneity problem with an Instrumental Variables estimation, which should rule out the endogeneity bias exploiting the exogenous variation in Catalan proficiency generated by the instruments used — variables that are highly correlated with knowledge of Catalan but not with the error term of the earnings equation[16]. In the next section we show that the Two-Stage Least Square estimation of the return to Catalan proficiency yields exactly the same result as the Endogenous Switching Model, even using as instruments only the sub-set of determinants of Catalan proficiency which pass the formal test for weak instruments and overidentification. As commented before, obtaining the same results with two alternative methodologies supports the validity of our results; even so, in this particular

[15] The zone of residence is defined as the district for those who reside in the city of Barcelona, the municipality when identifiable (mainly big municipalities) and the *comarca* (a wider territorial area than the unicipality, of which there are 41 in Catalonia) for small units.

[16] Note that also the Endogenous Switching Model requires an exclusion restriction, in order to guarantee that identification of the language premium is not only produced by the non-linearity of the self-selection correction term. However, the validity of this method also relies on the joint-normality distributional assumption of the error terms of the two equations, whereas the TSLS method is not subject to this non-trivial assumption.

case the Endogenous Switching Model seems to provide a more precise estimation of the language return. Moreover, as we show below, its first-stage equation offers the opportunity of analysing the reduced-form model for the determinants of Catalan proficiency, which is of independent interest for Catalan policymakers.

4. Estimation Results

4.1 The determinants of knowledge of Catalan

The ML estimation of the first-stage equation explaining the propensity to be able to speak and write Catalan is reported in Table 3. The Pseudo-R^2 obtained from an independent Probit model is 0.47, which indicates that the estimation performs very well, confirming that the variables included in the model have a high explanatory power for the probability of being proficient in Catalan. The results indicate that males and females do not show any significant difference in the likelihood of knowing Catalan. As expected, the propensity to be proficient in Catalan decreases with age, indicating that older individuals have more difficulty in assimilating this language.

Individuals born outside Catalonia are clearly penalized, except for those who come from the East of Spain (Valencia and Balearic Islands); this result is to be expected, since Catalan is also spoken in these regions of Spain (even though it is less institutionalized). Moreover, the disadvantage is higher for those who are born outside Spain, especially for Latin American immigrants; in all likelihood, this happens because their mother-tongue is Spanish, and the incentives for learning Catalan are lower for them (*ceteris paribus*). However, the positive and statistically significant coefficient for the duration in Catalonia (the years since migration) indicates that a longer exposure to the local language favours its assimilation[17]. Schooling is clearly one of the most important determinants of the probability of speaking and writing Catalan; but individuals who read more frequently and individuals with more than 100 books at home are more likely to know the language. As Rendon (2007) found, linguistic assimilation is easier for individuals who arrived at a young age; moreover,

[17] However, the negative sign of the interaction (Born in Spain)×YSM indicates that national immigrants are less likely to be proficient in Catalan as the length of their stay increases; therefore, the advantage of individuals who were born in Spain with respect to foreigners decreases with time spent in Catalonia. This shows that individuals who came from the rest of Spain in the past may have had fewer incentives to learn Catalan, since they may well have arrived when the use of this language was still restricted to private oral communication.

individuals affected by the normalization legislation of 1983 are more likely to being able to speak and write Catalan, especially those who received all their schooling in this language.

Table 3: The Determinants of Catalan Proficiency

Dependent Variable: Being Proficient in Catalan (0-1)	***Coefficients***	***z-Statistic***
Constant	-1.649	*-5.07*
Sex (=1 if female)	0.080	*1.14*
Age/10	-0.353	*-6.53*
Native-(Immigrant Sec. Gen.)	*ref. cat.*	—
East of Spain	0.045	*0.15*
South of Spain	-0.820	*-4.94*
Central Spain	-0.766	*-4.79*
North-West of Spain	-0.533	*-2.21*
North-East of Spain	-0.636	*-3.58*
Europe	-1.561	*-7.66*
Africa	-1.464	*-7.32*
Latin America	-2.056	*-11.60*
Asia and other countries	-1.899	*-10.65*
YSM/10 (= 0 for natives)	0.268	*4.78*
(Born in Spain)×YSM	-0.023	*-4.06*
Years of education	0.116	*11.30*
Reads frequently	0.390	*5.19*
More than 100 books at home	0.189	*2.33*
Arrived at 10 or younger	0.426	*3.53*
Partial Normalization	0.573	*5.36*
Complete Normalization	1.003	*6.20*
% Speak and write Catalan	2.485	*7.10*
Speaks Catalan with parents	0.374	*2.43*
Speaks Catalan at home	0.501	*4.42*
Speaks Catalan with children	0.501	*4.42*
News on Catalan TV	0.155	*1.91*
Catalan newspaper	0.218	*2.25*

z-Statistics in Italics

Our results also point out that the individual's environment plays an important role in explaining the chances of achieving language proficiency. In fact, the positive and highly significant coefficient associated with the proportion of individuals who speak and write Catalan in the neighbourhood of residence suggests that living in a (linguistically) more stimulating environment facilitates language assimilation. In addition, the use of Catalan within the family (with the parents, at home and with the children[18]) significantly increases the probability of speaking and writing the language. Finally, the local media also seem to play some role, given that the coefficient

[18] We also tried to include the number of children, but the coefficient was not statistically different from zero (common result in the literature); moreover, including this variable does not modify the overall results obtained.

associated with watching Catalan television and reading newspapers in Catalan increases the chances of being fully proficient.

4.2 Catalan Proficiency and Earnings

In this sub-section we analyze the results from the earnings regressions[19] and the effect of Catalan proficiency, which are shown in Table 4. Specifically, the first column of the table contains the estimates from a standard OLS Mincer-type equation. In addition, the second column contains the same earnings equation augmented by our indicator of language proficiency (being able to speak and write Catalan). In general, the model performs very well, and the coefficient estimates have the expected sign[20].

The first interesting finding is that foreigners' earnings penalty (with respect to native-born children of immigrants) is strongly reduced when controlling for knowledge of Catalan; as in previous studies (Amuedo-Dorantes and De La Rica 2007, Izquierdo et al. 2009), African immigrants seem the most disadvantaged, whereas European immigrants do not earn significantly less than second-generation immigrants. Moreover, once Catalan proficiency is taken into account, national immigrants earn significantly more than second-generation immigrants. The coefficient for years since migration has a negative sign, and it is statistically significant when controlling for language knowledge. This evidence at first sight seems contradictory; however, it may reflect the fact that individuals who arrived in Catalonia many years ago are less skilled than more recent immigrants (and also less likely to learn Catalan).

Females earn significantly less than males, and this difference is not modified when accounting for Catalan proficiency. Moreover, married individuals tend to have higher earnings than those who are not married in both equations. As expected, an increase in the years of schooling is associated with higher monthly earnings, and the return to education falls slightly when the regression includes the language proficiency indicator.

[19] Since the information regarding monthly earnings is coded in intervals, we construct a linear dependent variable from the mid-points of each interval (see the table 2A Appendix for details); the results do not change using the Interval Regression method.

[20] The R^2 reaches the value of 0.43, which suggests that the variables included in the model have a significant explanatory power for explaining the log of monthly earnings, and that the earnings equation is correctly specified.

Table 4: Catalan knowledge and Earnings

Dependent Variable: ***Ln(Earnings)***	**OLS**	**OLS + CATALAN PROFICIENCY**	**ENDOGENOUS SWITCHING ML**	**TSLS — SELECTED INSTRUMENTS**
Constant	5.888***	5.855***	5.802***	5.788***
	(0.056)	*(0.056)*	*(0.051)*	*(0.066)*
Native-(Immigrant Sec. Gen.)	*Ref. Cat.*	*Ref. Cat.*	*Ref. Cat.*	*Ref. Cat*
Spain	0.044	0.065**	0.092***	0.089***
	(0.028)	*(0.028)*	*(0.028)*	*(0.032)*
Europe	-0.059	-0.024	0.024	0.019
	(0.046)	*(0.046)*	*(0.040)*	*(0.052)*
Africa	-0.173***	-0.140***	-0.095*	-0.097**
	(0.031)	*(0.033)*	*(0.037)*	*(0.041)*
Latin America	-0.102***	-0.056**	0.004	0.001
	(0.025)	*(0.028)*	*(0.037)*	*(0.042)*
Asia and other countries	-0.125***	-0.085*	-0.032	-0.032
	(0.043)	*(0.044)*	*(0.039)*	*(0.052)*
YMS/10	-0.010**	-0.012**	-0.015**	-0.015***
	(0.005)	*(0.005)*	*(0.005)*	*(0.005)*
Sex (=1 if female)	-0.303***	-0.305***	-0.310***	-0.306***
	(0.016)	*(0.016)*	*(0.017)*	*(0.016)*
Married	0.064***	0.066***	0.067***	0.069***
	(0.016)	*(0.016)*	*(0.017)*	*(0.016)*
Years of schooling	0.044***	0.043***	0.040***	0.041***
	(0.002)	*(0.003)*	*(0.003)*	*(0.003)*
Job Tenure (in months)/10	0.011***	0.011***	0.012***	0.012***
	(0.001)	*(0.001)*	*(0.001)*	*(0.001)*
(Previous) Experience/10	0.103***	0.114***	0.125***	0.129***
	(0.024)	*(0.024)*	*(0.025)*	*(0.025)*
Experience^2/100	-0.018***	-0.019***	-0.019**	-0.020***
	(0.006)	*(0.006)*	*(0.006)*	*(0.006)*
Hours of work (per month)/10	0.026***	0.026***	0.026***	0.026***
	(0.002)	*(0.002)*	*(0.001)*	*(0.002)*
Union member	0.094***	0.090***	0.091***	0.086***
	(0.018)	*(0.018)*	*(0.018)*	*(0.018)*
#Workers>500	0.050**	0.047**	0.045*	0.041**
	(0.020)	*(0.020)*	*(0.020)*	*(0.020)*
Living in Barcelona	0.006	0.005	0.008	0.005
	(0.017)	*(0.017)*	*(0.017)*	*(0.017)*
Proficiency in Catalan	—	0.072***	0.164***	0.164***
		(0.019)	*(0.038)*	*(0.055)*
$\hat{\sigma}_{\varepsilon}$	0.371	0.370	0.371	0.371
$\hat{\rho}_{\varepsilon u}$	—	—	-0.18 (χ^2 = 7.16)	—

Standard Errors in Italics; Robust Standard Errors for OLS and TSLS Estimations.
Selected Instruments: Arrived at age 10 or younger, % Speak and write Catalan, Speaks Catalan at home, Speaks Catalan with children and Reads Frequently.

Additionally, increasing current job-tenure (linearly) raises individual earnings (one additional month increases monthly earnings by 0.011%), while the positive effect of previous potential labour market experience seems to be convex (given the negative and significant coefficient of the quadratic term). However, the return to these elements of human capital decreases when we take knowledge of Catalan into account. An

increase in one hour of work per month increases monthly earnings, and the estimated coefficient does not change with or without accounting for Catalan proficiency. Finally, individuals who are union members and those who work in large firm earn somewhat more than the others.

Above all, the main result shown in the second column of Table 6 is that the OLS estimate of the effect of Catalan proficiency is positive and statistically significant, suggesting that knowing Catalan is associated with monthly earning about 7.5% higher than the mean (that is, $\exp(0.072)-1\cong 0.075$). Nevertheless, as we explained above, this simple estimate of the return to language proficiency could be biased. Indeed, it may also contain the effect of elements which increase the likelihood of knowing Catalan, which are also potentially correlated with the individual's earnings capacity[21]. As explained above, we deal with this potential endogeneity with two alternative methodologies.

Specifically, in the third column of Table 3 we report the Maximum Likelihood estimates obtained from the Endogenous Switching Model, the first stage of which was illustrated in the previous section. The most important differences with respect to the previous estimates consist in *i*) the further reduction of earnings penalties for foreign workers when accounting for both language knowledge and the self-selection behind its relationship with earnings. Consistently, when controlling for selection into knowledge of Catalan, individuals who are born in other Spanish regions earn even more than natives (second-generation immigrants). Moreover, when we take into account those elements which improve the likelihood of being proficient in Catalan, *ii*) there is an increase in the negative effect of the years since migration, *iii*) additional reduction of the returns to education, and *iv*) an increase in the returns to current and previous experience.

Finally, the estimated earnings return to knowledge of Catalan is significantly higher when the self-selection process is taken into account, showing that individuals who are able to speak and write Catalan earn approximately 18% more than the rest ($\exp(0.164)-1\cong 0.18$). Note also that the correlation coefficient between the earnings equation's error term and the unobservable determinants of knowledge of Catalan is statistically significant and negative. The χ^2 test for the statistical significance of this

[21] Moreover, measurement error in knowledge of Catalan may also represent an additional source of bias. However, due to data limitation, we are not able to explicitly control for measurement error — we only use a stringent definition of Catalan proficiency. Therefore, the estimate that we obtain represents the lower bound of the true value of Catalan proficiency (under the standard hypothesis of the measurement error model).

correlation coefficient ($\hat{\rho}_{\varepsilon u}$) is 7.16 (P-value 0.0074), rejecting the null hypothesis of independence of the random terms of the two equations. This finding indicates that a) in order to obtain an unbiased estimate of the effect of proficiency in Catalan it is important to correct for self-selection (or, in general, for endogeneity), given that the earnings equation and the knowledge of Catalan equation are not independent. In addition, b) the negative coefficient of the correction term indicates that the unobservable elements that increase the propensity to speak and write in Catalan (maintaining the observable determinants fixed), are negatively correlated with the unobservable determinants of earnings.

The second strategy for obtaining an unbiased estimate of the economic value of knowledge of Catalan consists in estimating the language return through Two-Stage Least Square (TSLS). As suggested by Cameron and Trivedi (2009, pg. 192-194), the Endogenous Switching Model and the TSLS estimator are alternative and similar methods for dealing with the same endogeneity problem — an endogenous dummy variable. However, TSLS would also help us to understand which of the determinants of Catalan proficiency can be correctly used as exclusion restrictions. The presence of valid instruments is needed in both TSLS and in the Endogenous Switching Model to achieve the full identification of the language premium. Specifically, through the implementation of the formal tests for weak instruments and overidentification, we are able to verify which of the specific determinants of knowledge of Catalan (that is, variables that only appear in the first stage of the Endogenous Switching Model) can be reasonably considered as valid instruments.

In the fourth column in Table 4 we report the results obtained with the TSLS estimation, computed by using as instruments the sub-set of language determinants that pass the two tests for instrument validity (with a conservative level of statistical significance). The selected instruments are the young arrival age, language use (at home and with the children, but not with the parents), the weakly frequency of reading and the percentage of proficient individuals in the community of residence[22]. With respect to the potential weakness of the instruments, we obtain a Partial R^2 from the

[22] Running a TSLS regression using the full set of determinants of proficiency in Catalan as instruments yields the same estimated language return, but it fails the overidentification test. This result indicates that (regional) origin variables for national immigrants, the number of books at home, and the exposure to local media cannot be considered valid instruments, because these variables are potentially correlated with individual's earnings capacity. The same happens for the "normalization" indicators and speaking Catalan; with respect to the former, it is possible that the change of the language of instruction may had modified the earnings generation process — i.e. affecting the return to education, as suggested by Angrist and Lavy (1997). Moreover, speaking Catalan with the parents could be correlated with the error term of the earnings equation through the effect of parental networking.

first-stage of 0.13, with the associated Angrist-Pischke F-Statistic for the weak identification test of 72.1 (exceeding the Stock-Yogo critical values for weak identification test: see Cameron and Trivedi 2009 for details); with these values, we firmly reject the null hypothesis of weak instruments. Moreover, the Hansen's J test of overidentifying restrictions over the selected instruments is 2.8 (with a P-Value of 0.6), which suggests that the instruments are likely to be valid — they are uncorrelated with the structural error terms and are correctly excluded from the estimated earnings equation. In addition, as reported in Table 4, the return to knowledge of Catalan (such as the other estimated coefficients) is identical to the estimate obtained from the Endogenous Switching Model; this result is a clear sign in favour of the validity of our results. The obtained evidence suggests that we can reasonably assume that the linear combination of (at least) our selected instruments affects monthly earnings only through Catalan proficiency — i.e. it produces the exogenous source of variation of language knowledge that is necessary to rule out unobserved heterogeneity and to identify the true language premium[23]. However, the reader must bear in mind that the validity of our results is strictly subject to the untestable hypothesis of orthogonalithy between the linear combination of instruments and the earnings equation's error term. In any case, under the validity of the estimated premium, the language return estimated through the Endogenous Switching Model yields a more precise estimate[24]; moreover, its first stage is based on a compelling reduced-form equation for explaining the decision or not to learn Catalan, which also provides additional valuable information for policy making.

4.3 Sensitivity Analysis

Before considering the potential complementarity between knowledge of Catalan and educational attainment, we briefly analyse the sensitivity of the results obtained to our sample selection. In Table 5 we report the language premium estimated by OLS and by the Endogenous Switching Model for the selected sample and for different sub-

[23] With respect to the strength of the instruments, we compare the results for five just-identified specifications, for each one of the selected instruments (complete results are available upon request). The estimated premium ranges between 0.11 and 0.31, though considering the confidence interval of the estimates we cannot reject our final estimate (which is also much more precise). That is, the premium is sensitive to the included instruments, but not too much as to consider the selected instruments as invalid. Similar results are also obtained with different combination of the selected instruments.

[24] Moreover, additional results indicate that estimating the Endogenous Switching Model using only the selected exclusion restrictions also yields the same results, though somewhat less precise.

samples[25]. First of all, we separately compute the return for males and females, allowing also for a gender-specific self-selection process; in fact, males and females may face different costs, but also different returns to knowledge of Catalan. The results indicate that the positive effect of Catalan proficiency is higher for females than for males (as found by Rendon 2007); however, given that the obtained estimates are not statistically different, we do not reject the joint estimation for both sexes.

Table 5: Sensitivity Analysis

CHANGING SAMPLE SELECTION	$\exp\left(\hat{\beta}\right)-1$	***STANDARD ERROR***
Selected Sample — OLS	0.075	0.021
Selected Sample — Endogenous Switching Model	0.179	0.044
Only Males — OLS	0.039	0.026
Only Males — Endogenous Switching Model	0.144	0.055
Only Females — OLS	0.112	0.034
Only Females — Endogenous Switching Model	0.234	0.073
Excluding Foreigners — OLS	0.067	0.023
Excluding Foreigners — Endogenous Switching Model	0.143	0.054

Moreover, we also try to exclude non-national immigrants from the sample, because knowledge of Catalan might represent additional advantages for immigrants in the labour market (for example, a better adaptation to the local institutions)[26]. As expected, the estimated return to Catalan proficiency computed excluding foreigners is lower than the return obtained from the estimation with the overall (selected) sample. This means that the earning premium for being proficient in Catalan is somewhat higher for foreign immigrants than for national first-generation immigrants and native-born second-generation immigrants. Again, taking into account the standard error of the estimate, the distribution of the estimated language return with the overall selected sample overlaps with the one obtained excluding foreigners; therefore, we cannot reject the joint estimation including foreign individuals.

4.4 Language-Education Complementarity

Following on from the recent literature, in this section we check the existence of complementarity between language knowledge and individual skills in Catalonia. More specifically, we focus on the role of completed education, as we suspect that being

[25] Complete estimations are not shown here, but are available upon request from the authors. Notice that, also for these different sub-samples, the results from the TSLS estimation are almost identical.
[26] Because of the reduced number of observations we are not able to carry out the estimation for the foreigner sub-sample .

proficient in Catalan might have an economic reward only for medium/high-educated workers. In fact, knowledge of Catalan might represent a valuable asset only in occupations that require higher levels of formal education. Therefore, we estimate the return to knowledge of Catalan separately for individuals with low education (that is, individuals with eight or less years of schooling) and for individuals with high education (individuals with more than eight years of schooling); the estimated return to proficiency in Catalan for both sub-samples are reported in Table 6[27].

Table 6: Language-Education Complementarity

DIVIDING HIGH AND LOW EDUCATED INDIVIDUALS	$\exp\left(\hat{\beta}\right)-1$	***STANDARD ERROR***
Selected Sample — OLS	0.075	0.021
Selected Sample — Endogenous Switching Model	0.179	0.044
High Education — OLS	0.072	0.029
High Education — Endogenous Switching Model	0.248	0.064
Low Education — OLS	0.036	0.029
Low Education — Endogenous Switching Model	-0.002	0.061

Note: High Education = Individuals with more than 8 years of schooling.
Low Education = Individuals with 8 years of schooling or less.

The results indicate strong complementarity between knowledge of Catalan and individual's schooling. The return estimated with the simple OLS for the full sample is almost completely identified by the return to Catalan proficiency for more educated individuals; indeed, the estimated return to language knowledge for less educated individuals is not statistically different from zero. Moreover, we obtain similar results when accounting for endogeneity, estimating the return to knowledge of Catalan through the Endogenous Switching Model. There exists a positive and statistically significant return for high-educated individuals (somewhat higher than the estimate for the full sample), but for low-educated individuals the return is almost zero. Definitively, it seems that being proficient in Catalan increases expected monthly earnings for first- and second-generation immigrants, but only for those who have achieved a certain level of skill through the educational system[28]. The evidence that we

[27] Table 4A in the appendix contains the complete OLS earnings equations, and the second-stage equations from the Endogenous Switching Model; for the sake of brevity, the results from the first-stage are not shown here, but are available upon request.

[28] However, we must take into account the results obtained by González (2005). Estimating non-parametric bounds to language returns, she suggests that the higher returns to language for high-educated workers are only due to a strong selection ability bias in higher education levels. We should bear in mind that our results, which are obtained by separating according to educational level, may only capture (or may only capture in part) the process of self-selection into higher levels of education produced by individual unobserved ability. In any case, we also estimated a TSLS model with both Catalan proficiency and schooling as endogenous variables, instrumenting an individual's schooling with parental education and father's occupation, but the estimated return to Catalan knowledge was only slightly reduced (from 18% to 16%), and the return to education remained virtually unchanged.

have obtained raises several questions with regard to the planning of the linguistic and socioeconomic policy agenda in Catalonia.

5. Summary and Conclusion

This paper is the first attempt to quantify the economic value of Catalan knowledge in terms of earnings, for first- and second-generation immigrants in Catalonia. More specifically, we compute the expected earnings increase for those individuals who are able to speak and write Catalan, which is our definition of language proficiency. From a simple OLS estimation, we obtain that the monthly earnings of proficient individuals are roughly 7.5% higher. In addition, Catalan proficiency accounts for a significant part of the estimated by-origin differences in earnings. However, we argue that these OLS estimates could be seriously biased by the endogeneity of knowledge of Catalan in the earnings equation; we deal with this potential endogeneity with two alternative methodologies.

First, we estimate the return to knowledge of Catalan with an Endogenous Switching Model; we model the decision process related to the acquisition of speaking and writing competences with a reduced-form Probit equation, which is jointly estimated with the earnings equation. The results indicate that taking into account the self-selection process behind knowledge of Catalan, the earnings premium for first- and second-generation immigrants amounts to an 18% increase in earnings (significantly surpassing the OLS estimate). Second, we also address the endogeneity by implementing a TSLS estimation, using as instruments the sub-set of determinants of knowledge of Catalan that pass the formal tests for weak instruments and overidentifications (with a conservative level of statistical significance). The return to knowledge of Catalan estimated by TSLS is identical to the estimate from the Endogenous Switching Model. So the robustness of the results to different methodologies and to different exclusion restrictions validates our general findings.

In addition, from the Endogenous Switching Model, we also obtain a negative correlation coefficient between the unobservable determinants of knowledge of Catalan and the error term in the earnings equation; this result, in principle, is not consistent with the unobserved ability argument. In general, learning the host-country's language is a "need" for immigrants; therefore, unexplained differences between them in

language fluency may be principally related to an unobserved ability. However, as mentioned above, given the co-officiality with Castilian and the absence of legal requirements, knowing Catalan in Catalonia is more a choice than a necessity. In all likelihood, the most important unobserved determinant of knowledge of Catalan could be represented by the economic expectations related to achieving complete proficiency. Possibly, individuals who (*ceteris paribus*) learn Catalan because they expect that this is the only way to improve their situation in the labour market are more likely to end up in less-paid occupations. This tendency could account for the negative correlation coefficient between the error terms of the two equations, and it represents an interesting issue to be investigated in future research[29].

These results have an important bearing on the future planning of linguistic policies in Catalonia, given that stimulating the use of Catalan through public policies — beyond the explicit cultural and linguistic purposes — has far-reaching socioeconomic implications as well. What is still more relevant for the design of linguistic policies in Catalonia in the medium-long term is the strong complementarity between knowledge of Catalan and individual schooling. Specifically, our results suggest that only more educated individuals (with more than eight years of schooling) receive an earnings return to language knowledge. This means that for low-educated individuals, Catalan proficiency may have only a very limited impact on their labour market outcomes (especially if defined with respect to earnings). In fact, the target of linguistic policies (as of other socioeconomic policies) is often the foreign population, which currently includes those who are more likely to be low-educated. Therefore, policy-makers might consider that in order to achieve a complete economic and social assimilation (maybe a step beyond cultural integration), fostering knowledge of Catalan among immigrants may not be sufficient. An effective public policy would consist in combining Catalan learning with promotion of human capital and other skills for the most disadvantaged part of the Catalan population.

[29] In fact, an important caveat of this work is that we neglect the potential occupation selection effect of knowledge of Catalan, in the spirit of Chiswick and Miller (2010). However, the inclusion of information about individual occupation requires a special treatment whether individuals self-select into occupation according to their linguistic proficiency (see Aldashev et al, 2009), which is out of the purposes of this work and represents an interesting issue for future research.

References

Alarcón, A. & Solé, C. (2001). "*Llengua i Economia a Catalunya*", Institut d'Estudis Catalans.
Aldashev, A., Gernandt, J. & Thomsen, S.L., (2009). "Language usage, participation, employment and earnings: Evidence for foreigners in West Germany with multiple sources of selection," Labour Economics, vol. 16(3), pp. 330-341.

Amuedo-Dorantes, C. & de la Rica, S. (2007). "Labour Market Assimilation of Recent Immigrants in Spain," British Journal of Industrial Relations, vol. 45(2), pp. 257-284, 06.

Angrist J.D. & Lavy, V. (1997) "The effect of a change in language of instruction on the returns to schooling in Morocco" Journal of Labor Economics, vol. 15, pp. 48-76.

Berman, E., Lang, K. & Siniver, E. (2003). "Language-skill complementarity: returns to immigrant language acquisition," Labour Economics, vol. 10(3), pp. 265-290.

Bleakley, H. & Chin, A. (2004). "Language Skills and Earnings: Evidence from Childhood Immigrants," The Review of Economics and Statistics, vol. 86(2), pp. 481-496, 05.

Cameron, A. C., & Trivedi, P. K., (2009). "*Microeconometrics Using STATA*", STATA press.

Carliner G. (1996). "The wages and language skills of US immigrants". NBER Working Paper W5763.

Chiswick, B. (1991). "Speaking, Reading, and Earnings among Low-Skilled Immigrants," Journal of Labor Economics, vol. 9(2), pp. 149-70.

Chiswick, B. & Miller, P.W. (1995). "The Endogeneity Between Language and Earnings: International Analyses"; Journal of Labor Economics, vol. 13(2), pp. 245-287.

Chiswick, B. & Miller, P.W. (1999). "Language skills and earnings among legalized aliens"; Journal of Population Economics, vol. 12(1), pp. 63-89.

Chiswick, B. & Miller, P.W. (2001). "A Model of Destination Language Acquisition: Application to Male Immigrants in Canada"; Demography, 38(3), pp. 391-409.

Chiswick, B. & Miller, P.W. (2007). "*The Economics of the Language: International Analyses*"; Routledge editors.

Chiswick, B. & Miller, P.W. (2010). "Occupational language requirements and the value of English in the US labor market". Journal of Population Economics, vol. 23(1), pp. 353-372.

Dustmann, C. & van Soest, A. (2001). "Language Fluency And Earnings: Estimation With Misclassified Language Indicators"; The Review of Economics and Statistics, 83(4), pp. 663-674.

Dustmann, C. & van Soest, A. (2002). "Language and the earnings of immigrants," Industrial and Labour Relations Review, vol. 55(3), pp. 473-492.

Dustmann, C. & Fabbri, F. (2003). "Language proficiency and labour market performance of immigrants in the UK"; Economic Journal, 113(489), pp. 695-717, 07.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error"; Econometrica, vol. 47, pp.153-162.

Fernández-Huertas Moraga, J & Ferrer-i-Carbonell, A. (2008). Immigration in Catalonia. El Centre d'Estudis de Temes Contemporanis (CETC).

González, L. (2005). "Nonparametric bounds on the returns to language skills," Journal of Applied Econometrics, vol. 20(6), pp. 771-795.

Izquierdo, M., Lacuesta, A. & Vegas, R. (2009). "Assimilation of Immigrants in Spain: a Longitudinal Analysis" Labour Economics, 16(6) pp. 669-678.

Lang, K. & Siniver, E. (2009). "The Return to English in a Non-English Speaking Country: Russian Immigrants and Native Israelis in Israel," The B.E. Journal of Economic Analysis & Policy, vol. 9(1), 50.

Lee, L., (1978). "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables." *International Economic Review*, vol.19, pp. 415-433.

Lindley, J. (2002). "The English Language Fluency and Earnings of Ethnic Minorities in Britain," Scottish Journal of Political Economy, vol. 49(4), pp. 467-87.

Mora M.T., & Davila, A. (1998). “Gender, Earnings, and the English Skill Acquisition of Hispanic Workers in the U.S.” Economic Inquiry, 26, pp. 631-644.
Pujolar, J. (2009) “Immigration and language education in Catalonia: Between national and social agendas”. Linguistics and Education, *i-first.*

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

APPENDIX

Table 1A: Sample Selection Criteria

Total sample	10,358
Only Individuals aged 16 to 65	-1,719
Only individuals who were regularly working when interviewed	-2,790
Excluding natives with at least one parent born in Catalonia	-2,912
Only individuals with valid information about earnings	-355
Final sample	2,582

Table 2A: Monthly Earnings in Brackets (in Euros) — Selected Sample

EARNING INTERVALS (in Euros)	FREQ.	% SAMPLE
Less than 450	111	4.30
Between 451 and 600	168	6.51
Between 601 and 900	672	26.03
Between 901 and 1050	372	14.41
Between 1051 and 1200	417	16.15
Between 1201 and 1500	373	14.35
Between 1501 and 1800	212	8.21
Between 1801 and 3000	221	8.56
More than 3000	36	1.49

Definition of constructed variables:

Age = *mid points of the original age variable collected in intervals (16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65).*

YSM = *2005 minus year of arrival in Catalonia; year of arrival in Catalonia collected in intervals (before 1940, between 1941 and 1945, 1946-1950, 1951-1955, 1956-1960, 1961-1965, 1966-1970, 1971-1975, 1975-1980, 1981-1985, 1986-1990, 1991-1995, 1996-2000, 2001-2005).*

Arrived at 10 or younger* = *1 if the individual arrived in Catalonia at age 10 or younger, 0 otherwise.*

Complete Normalization* = *1 if the individual was born after 1977, and arrived at age 6 or younger if born outside Catalonia, 0 otherwise.*

Partial Normalization* = *1 if the individual was born between 1969 and 1977 and arrived younger than age 16 if born outside Catalonia, 0 otherwise*.

% Speak and Write Catalan* = *% of individuals who claim to be able to speak and write Catalan (original sample): by district for those who reside in the city of Barcelona; by municipality when identifiable; by comarca for small units.*

Years of Schooling = *4 if the individuals has no education, or hif e/she has completed only primary education (grouped in the original database); 7 for lower-secondary uncompleted; 8 for lower-secondary completed; 12 for vocational education; 13 for general upper-secondary education; 18 for completed tertiary education.*

Job Tenure (in months) = *mid point of the original variable collected in intervals (< 2 years, 2 - 5 years, 5 - 10 years, 10 - 15 years, 15 - 20 years, > 20 years).*

Previous Experience = *potential work experience, previous to the current work (age-schooling-(job-tenure (in months))/12 -6).*

Hours of work (per month) = *mid point of the original variable (hours per week) collected in intervals times 4 (0-20, 20-30, 30-35, 35-40, 40-45, 45-55, 55-100).*

*Variables constructed by IDESCAT staff, from the original registers of the survey (maximum desegregation).

Table 3A: Variables Description and Descriptive Statistics

VARIABLE	DESCRIPTION	MEAN	S.D
Native-(Immigrant Sec. Gen.)	*= 1 if born in Catalonia (parents born outside Catalonia), 0 otherwise*	0.406	0.491
Spain	*= 1 if born in other Spanish regions, 0 otherwise*	0.329	0.470
East of Spain	*= 1 if born in the East of Spain, 0 otherwise*	0.018	0.132
South of Spain	*= 1 if born in the South of Spain, 0 otherwise*	0.163	0.369
Central Spain	*= 1 if born in Central Spain, 0 otherwise*	0.092	0.289
North-West of Spain	*= 1 if born in the North-West of Spain , 0 otherwise*	0.022	0.146
North-East of Spain	*= 1 if born in the North-East of Spain, 0 otherwise*	0.035	0.183
Europe	*= 1 if born in Europe, 0 otherwise*	0.041	0.198
Africa	*= 1 if born in Africa, 0 otherwise*	0.066	0.248
Latin America	*= 1 if born in Latin America, 0 otherwise*	0.112	0.315
Asia and Other Countries	*= 1 if born in Asia or Other Countries, 0 otherwise*	0.046	0.211
Sex =1 if Female	*= 1 if the individual is Female, 0 otherwise*	0.423	0.494
Age	*= Individual's age in years*	39.86	10.66
YSM	*= Years since migration to Catalonia (0 for native-born individuals)*	15.92	23.99
Years of Education	*= Years of completed schooling (imputed)*	11.15	3.91
Speak Catalan with Parents	*= 1 if the individual speaks Catalan with at least one parent, 0 otherwise*	0.057	0.232
Speak Catalan at Home	*= 1 if the individual speaks Catalan at home, 0 otherwise*	0.176	0.381
Speak Catalan with Children	*= 1 if the individual speaks Catalan with his/her children, 0 otherwise*	0.184	0.387
More than 100 books at home	*= 1 if the individual has more than 100 books at home, 0 otherwise*	0.311	0.463
Reads Frequently	*= 1 if the individual reads every day or many days at week, 0 otherwise*	0.364	0.481
News Catalan TV	*= 1 if the individual watches the newscast on the Catalan television, 0 otherwise*	0.254	0.436
Catalan Newspaper	*= 1 if the individual normally reads the newspaper in Catalan, 0 otherwise*	0.123	0.329
Arrived Younger than 10	*= 1 if the individual arrived at Catalonia at 10 or younger, 0 otherwise*	0.156	0.363
Complete Normalization	*= 1 if the individual was completely affected by the Linguistic Normalization of 1983 (completely schooled in Catalan) , 0 otherwise*	0.183	0.387
Partial Normalization	*= 1 if the individual was only partially affected by the Linguistic Normalization of 1983 (partially schooled in Catalan) , 0 otherwise*	0.111	0.314
% Write Catalan (*Comarca*)	*= Percentage of individuals able to speak and write in Catalan in the neighbourhood of residence*	0.516	0.099
Married	*= 1 if the individual is married, 0 otherwise*	0.622	0.485
Job Tenure (in Months)	*= Job tenure in months (current job)*	105.7	116.35
(Previous) Experience	*= Previous potential experience in years*	14.24	10.86
Hours of work (per month)	*= Number of hours worked per month*	155.87	49.89
Unionized	*= 1 if the individual is affiliated to a labour union, 0 otherwise*	0.417	0.493
#Workers>500	*= 1 if the individual works in a firm with more than 500 workers, 0 otherwise*	0.275	0.446
Living in Barcelona	*= 1 if the individual lives in the Barcelona Metropolitan Area, 0 otherwise*	0.280	0.449

Table 4: Catalan knowledge and Earnings by Completed Education

Dependent Variable: ***Ln(Earnings)***	**OLS High-Education**	**OLS Low-Education**	**ML High-Education**	**ML Low-Education**
Constant	5.888***	5.855***	5.802***	5.788***
	(0.056)	*(0.056)*	*(0.051)*	*(0.066)*
Native-(Immigrant Sec. Gen.)	*Ref. Cat.*	*Ref. Cat.*	*Ref. Cat.*	*Ref. Cat*
Spain	0.044	0.065**	0.092***	0.089***
	(0.028)	*(0.028)*	*(0.028)*	*(0.032)*
Europe	-0.059	-0.024	0.024	0.019
	(0.046)	*(0.046)*	*(0.040)*	*(0.052)*
Africa	-0.173***	-0.140***	-0.095*	-0.097**
	(0.031)	*(0.033)*	*(0.037)*	*(0.041)*
Latin America	-0.102***	-0.056**	0.004	0.001
	(0.025)	*(0.028)*	*(0.037)*	*(0.042)*
Asia and other countries	-0.125***	-0.085*	-0.032	-0.032
	(0.043)	*(0.044)*	*(0.039)*	*(0.052)*
YMS/10	-0.010**	-0.012**	-0.015**	-0.015***
	(0.005)	*(0.005)*	*(0.005)*	*(0.005)*
Sex (=1 if female)	-0.303***	-0.305***	-0.310***	-0.306***
	(0.016)	*(0.016)*	*(0.017)*	*(0.016)*
Married	0.064***	0.066***	0.067***	0.069***
	(0.016)	*(0.016)*	*(0.017)*	*(0.016)*
Years of schooling	0.044***	0.043***	0.040***	0.041***
	(0.002)	*(0.003)*	*(0.003)*	*(0.003)*
Job Tenure (in months)/10	0.011***	0.011***	0.012***	0.012***
	(0.001)	*(0.001)*	*(0.001)*	*(0.001)*
(Previous) Experience/10	0.103***	0.114***	0.125***	0.129***
	(0.024)	*(0.024)*	*(0.025)*	*(0.025)*
Experience^2/100	-0.018***	-0.019***	-0.019**	-0.020***
	(0.006)	*(0.006)*	*(0.006)*	*(0.006)*
Hours of work (per month)/10	0.026***	0.026***	0.026***	0.026***
	(0.002)	*(0.002)*	*(0.001)*	*(0.002)*
Union member	0.094***	0.090***	0.091***	0.086***
	(0.018)	*(0.018)*	*(0.018)*	*(0.018)*
#Workers>500	0.050**	0.047**	0.045*	0.041**
	(0.020)	*(0.020)*	*(0.020)*	*(0.020)*
Living in Barcelona	0.006	0.005	0.008	0.005
	(0.017)	*(0.017)*	*(0.017)*	*(0.017)*
Proficiency in Catalan	—	0.072***	0.164***	0.164***
		(0.019)	*(0.038)*	*(0.055)*
$\hat{\sigma}_{\varepsilon}$	0.371	0.370	0.377	0.35
$\hat{\rho}_{\varepsilon u}$	—	—	-0.29 (χ^2 = 8.54)	0.077 (χ^2 = 0.5)

Standard Errors in Italics; Robust Standard Errors for OLS Estimations.
High Education = Individuals with more than 8 years of schooling.
Low Education = Individuals with 8 years of schooling or less.

Últims documents de treball publicats

NUM	TÍTOL	AUTOR	DATA
10.01	Language knowledge and earnings in Catalonia	Antonio Di Paolo, Josep Lluís Raymond-Bara	Febrer 2010
09.12	Inflation dynamics and the New Keynesian Phillips curve in EU-4	Borek Vasicek	Desembre 2009
09.11	Venezuelan Economic Laboratory The Case of the Altruistic Economy of Felipe Pérez Martí	Alejandro Agafonow	Novembre 2009
09.10	Determinantes del crecimiento de las emisiones de gases de efecto invernadero en España (1990-2007)	Vicent Alcántara Escolano, Emilio Padilla Rosa	Novembre 2009
09.09	Heterogeneity across Immigrants in the Spanish Labour Market: Advantage and Disadvantage	Catia Nicodemo	Novembre 2009
09.08	A sensitivity analysis of poverty definitions	Nicholas T. Longford, Catia Nicodemo	Novembre 2009
09.07	Emissions distribution in postKyoto international negotiations: a policy perspective	Nicola Cantore, Emilio Padilla	Setembre 2009
09.06	Selection Bias and Unobservable Heterogeneity applied at the Wage Equation of European Married Women	Catia Nicodemo	Juliol 2009
09.05	La desigualdad en las intensidades energéticas y la composición de la producción. Un análisis para los países de la OCDE	Juan Antonio Duro Moreno, Vicent Alcantara Escolano, Emilio Padilla Rosa	Maig 2009
09.04	Measuring intergenerational earnings mobility in Spain: A selection-bias-free approach	María Cervini Pla	Maig 2009
09.03	The monetary policy rules and the inflation process in open emerging economies: evidence for 12 new EU members	Borek Vasicek	Maig 2009
09.02	Spanish Pension System: Population Aging and Immigration Policy	Javier Vázquez Grenno	Abril 2009
09.01	Sobre los subsistemas input-output en el análisis de emisiones contaminantes. Una aplicación a las emisiones de CH4 en Cataluña	Francisco M. Navarro Gálvez, Vicent Alcántara Escolano	Març 2009
08.10	The monetary policy rules in EU-15: before and after the euro	Borek Vasicek	Desembre 2008
08.09	Agglomeration and inequality across space: What can we learn from the European experience?	Rosella Nicolini	Desembre 2008