Budría et al. IZA Journal of Development and Migration (2017) 7:12
DOI 10.1186/s40176-017-0094-2

IZA Journal of Development and Migration

ORIGINAL ARTICLE Open Access

The impact of host language proficiency across the immigrants' earning distribution in Spain

Santiago Budría[1*], Carlos Martinez de Ibarreta[1] and Pablo Swedberg[2]

* Correspondence: srbudria@comillas.edu
[1]Department of Quantitative Methods, Universidad Pontificia Comillas, C/Alberto Aguilera 23, s/n, 28015 Madrid, Spain
Full list of author information is available at the end of the article

Abstract

This paper explores the impact of Spanish language proficiency on immigrant earnings in Spain using an instrumental variable quantile regression approach. The impact is on average roughly 17.2% but varies substantially across the earning distribution. The return to destination language proficiency actually ranges from zero at the bottom quantiles to 30% at the top quantile of the earning distribution. These findings suggest that the benefits derived from host language knowledge are particularly important among individuals with stronger unobserved abilities and marketable skills and that language training policies targeted at specific immigrant population categories may be ineffective from a labor market earning perspective.
JEL Classification: F22, J24, J61

Keywords: Immigration, Spanish language proficiency, Earnings, Instrumental variable quantile regression (IVQR)

1 Introduction

There is a significant literature examining how immigrant's host language proficiency affects earnings. Most of this research has been conducted in English-speaking countries (Rivera-Batiz 1992; Dustmann and Fabbri 2003; Lui 2007; Chiswick and Miller 1999, 2010; Bleakley and Chin 2004; Zhen 2013). More recently, researchers have further focused on non-English-speaking countries, including Germany (Dustmann and van Soest 2002), Israel (Chiswick 1998; Chiswick and Repetto 2001; Berman et al. 2003), and Norway (Hayfron 2001), among others. The common finding is that greater destination language fluency significantly raises immigrants' earnings. In Spain, most research conducted has focused on Catalonia, and its regional language, Catalan (Rendón 2007; Di Paolo 2011; Di Paolo and Raymond 2012), while efforts to assess the impact of Castilian Spanish language proficiency on immigrant earnings at a country level are much relatively recent (Budría and Swedberg 2014). There is also evidence to suggest that returns to foreign languages are sizable in Spain (Isphording 2013).

This paper uses the Spanish National Immigrant Survey (NISS), a large-scale immigration survey released by the Spanish National Statistics Institute, to calculate quantile returns to host language proficiency. Most papers to date have estimated returns only at the average of the (conditional) wage distribution. Averages, however, fail to describe the full distributional impact of the variable under scrutiny unless this

variable affects both the central and the tail deciles in the same way. In many cases, interest focuses on the impact of the covariate on points other than the center of the conditional distribution. This seems particularly relevant in the present context, as the effects of host language proficiency among the low-wage immigrants will probably be more relevant for public policy than the effects among the high-earning population. Recent evidence shown for other countries suggests significant differences across the earning distribution, albeit the extent and pattern of fluctuations differs between studies (Boyd and Cao 2009; Wang and Wang 2011; Ginsburgh and Prieto-Rodriguez 2011). To our knowledge, this is the first paper to examine the extent of heterogeneity across segments of the conditional wage distribution for the returns to Spanish language proficiency.

The paper relies on Chernozhukov and Hansen's (2008) instrumental variable quantile regression (IVQR) approach. The main advantage of this method is that it allows us to account for the potential endogeneity of the language variable. This refinement is crucial in the present context. Simple OLS and quantile estimates are likely to be upward biased if language ability depends on unobservable individual characteristics that are potentially related to unmeasurable earning determinants. At the same time, self-reported measures of language proficiency are subject to measurement error, an issue that has drawn the attention of researchers (Dustmann and van Soest 2002, 2004; Bleakley and Chin 2004). While classical measurement error leads to attenuation bias whereby OLS and QR estimates are below the true returns to Spanish proficiency, the bias under non-systematic errors is more ambiguous and complex.

To partially address these issues, we search for instruments that account for exogenous variations in Spanish language proficiency. Instrumental variables can provide consistent estimates under ability-bias and classical measurement error. While exploring the extent of non-classical measurement error is beyond the scope of the present paper and can be still a concern using IV, there is evidence that suggests that classical dominates non-classical measurement error as a source of bias in OLS and IV estimates.[1] The first instrument used in this paper is based on Bleakley and Chin (2004) and exploits the fact that younger children learn languages more easily than older children. The second instrument exploits the notion that parents' exposure to communication with their children in the destination country's language acts as a transmission mechanism. This mechanism is more likely to operate among immigrants who live with children in school, for school enrolment and attendance contribute notably and more rapidly to the children's host language proficiency.

There is some debate in the policy arena on whether language proficiency is associated with unobserved ability. Non-proficient immigrants may be, in some ways, less capable and therefore lack essential abilities and skills that are required to perform a high-paying job. If this was the case, their lower wages are a mere statistical illusion that reflects an omitted variable problem rather than a causal relationship between language ability and earnings. Under IVQR, the estimates at different quantiles represent the impact of a given covariate for individuals that have the same observable characteristics but that due to unobserved earning capacity are located at different points of the earning distribution. By "unobserved earnings capacity," we are referring to all the unmeasured characteristics that actually affect the worker's position within the wage distribution, including not only individual-level abilities and skills but also contextual-level characteristics such as workplace conditions. Thus, we show how immigrant

workers who are proficient in Spanish within the various segments of the earning distribution are affected relative to their non-proficient counterparts. The major advantage of this approach is that it prevents us from comparing proficient individuals enjoying an advantageous earning capacity with non-proficient individuals subject to an unfavorable earning condition. This approach has proven fruitful in the economics of education literature to ascertain whether certain educational attributes are associated with unobserved earning ability (McGuinness and Bennett 2007; Bárcena et al. 2012).

An additional contribution of the paper is that it illustrates the role of host language proficiency in shaping wage inequality. Average estimates assume that the marginal impact of language proficiency on wages is constant over the wage distribution. If true, the effect of having language skills can be represented by a shift (to the right) of the conditional wage distribution. In contrast, the IVQR estimates represent the impact of host language proficiency on wages at different points of the distribution, thus describing changes not only in the location but also in the shape of the distribution. By combining average with quantile estimates, we can assess the impact of Spanish proficiency on wage inequality between and within groups: while average returns measure the average differential between proficient and non-proficient immigrants, differences in quantile returns represent the wage differential between proficient immigrants that are located at different quantiles of the distribution.

To provide a more detailed view, the paper conducts separate regressions for immigrants with different educational attainment. There are reasons to believe that language proficiency and schooling are complementary inputs of the earning-generating process. Language skills are more likely to represent a valuable asset in occupations that require higher levels of formal education. Moreover, since poor language skills may hamper life opportunities, social mobility, and job offers, we expect stronger effects from language proficiency among the highly educated. The evidence collected so far is scarce and suggestive of diverging degrees of complementarity between schooling and language skills (Chiswick and Miller 2003; Casale and Posel 2011). This paper documents not only differences among education groups in terms of the return to host language proficiency but also differences within groups. The aim is to examine whether the extent of heterogeneity surrounding the returns to host language proficiency depends on an individual's education. Such heterogeneous effects may have pronounced implications for the design of effective immigrant integration policies. An immigration policy priority in OECD countries is language training (OECD 2012). This has become extremely relevant as a result of the increasing amount of refugees (i.e., non-economic immigrants) that are migrating to high-income countries in Europe. Unfortunately, the scope attributed to such policies may be more modest than presumed if workers with low qualifications and in the lowest segments of the earning distribution fail to reap relevant returns from language training.

The paper is organized as follows. Section 2 provides a brief background of recent immigration to Spain and a review of the literature. Section 3 describes the dataset, the estimating sample, and the Spanish language proficiency question. Section 4 depicts the estimation strategy including the IVQR approach. Section 5 presents and examines the estimates for the impact of Spanish language proficiency on immigrants' earnings. Section 6 discusses issues of instrument quality and outlines theoretical implications. Section 7 contains the concluding remarks.

2 Background and previous literature

Immigration and its impact on the labor market in Spain are extremely relevant topics. There has been a rapid and intense transformation of the structure and composition of the population in Spain during the period 2000–2011. According to OECD estimates (2015), the stock of foreign-born population increased from 3.4% of the total population in 2000 to a peak of 12.4% in 2010–2011, representing roughly 5,751,000 immigrants. Correspondingly, Spain ranks fifth among OECD countries in stocks of foreign-born population. Moreover, the economic downturn initiated in the third quarter of 2008 has slowed down migration inflows significantly, increased migration outflows, and more than doubled the unemployment rate. As a result of the decline in new entries (OECD 2013) and the increase in return migration due to worsening labor market conditions, Spain has experienced negative net migration since 2010. In 2015, immigration exceeded emigration for the first time since 2010. In particular, 291,387 immigrants arrived in Spain and 253,069 people left the country according to the Spanish National Statistics Institute (INE). Interestingly, Colombians, Ecuadorians, Bolivians, and Peruvians accounted for almost half of the leavers for the period 2010–2014. Indeed, language may play a crucial role in return migration since five of the ten largest immigrant populations that arrived in Spain are from Spanish-speaking countries. Since Spain is the only Spanish-speaking country in the EU and Spaniards typically exhibit very poor foreign language skills as shown by the Eurobarometer 2012, many young Spaniards compete for jobs with immigrants. The economic recession has continued in Spain since late 2011, and the latest and most adverse consequence of the double-dip recession is the second highest unemployment rate in the European Union—19.8% (Eurostat 2016). According to new figures released by Eurostat in 2015, unemployment among foreigners in Spain is 29.8%, vastly exceeding the unemployment rate for nationals.

Moreover, international evidence shows that immigrants experience a negative wage gap with respect to native earnings. This gap is inversely related to years since migration, although the degree of earning assimilation is found to differ across studies (Hu 2000; Friedberg 2000; Adsera and Chiswick 2007; Beenstock et al. 2010). Furthermore, additional efforts have been conducted in the literature to test whether there are asymmetric effects in the immigrant-native wage gap across the wage distribution. In particular, Chiswick et al. (2008) measure the immigrant-native gap in the USA and Australia focusing on the partial impact of schooling and work experience at each decile on the earning distribution. Their results show that immigrants from non-English-speaking countries experience lower returns to human capital skills at each decile of the earning distribution than do immigrants from English-speaking countries. Specifically, the earning penalty for non-English-speaking immigrants increases beyond the third decile of the wage distribution. Similarly, Billger and Lamarche (2010) examine native-immigrant earning differentials across the wage distribution in the USA and the UK and find that immigrants from non-English-speaking countries receive substantially lower wages throughout the wage distribution. The wage penalty is stronger for male immigrants at the bottom of the wage distribution. This may highlight that these immigrants select into low-paying jobs and/or the presence of wage discrimination in the UK. Conversely, in the USA, the wage penalty is greater for non-English-speaking female workers at the top of the earning distribution. Le and Miller (2012) investigate

the variation in the earning gender gap across the distribution for immigrant and native women in the USA. Their main findings show that female immigrants from English-speaking countries experience a disadvantage exclusively as a result of their gender condition whereas female settlers from non-English-speaking countries experience a double disadvantage as women and immigrants. All in all, these results reveal that migrants cannot fully utilize their human capital attributes and that immigrants with high and low unobserved earning capacity are similarly affected. Lastly, using data from Spain, Anton et al. (2010) find that the immigrant-native wage gap increases to a maximum of 25% as we move up along the wage distribution. Their results suggest that there may be a glass ceiling for immigrant workers in Spain.

2.1 Host language proficiency and earnings

The great majority of studies in the field have been carried out in English-speaking countries including the USA, Canada, the UK, and Australia. In particular, OLS estimates of the impact of host language proficiency on earnings are typically moderate, ranging between 5 and 10% (Rivera-Batiz 1992; Dustmann and Fabbri 2003; Lui 2007; Chiswick and Miller 2010). Nonetheless, language knowledge may depend on unobservable individual characteristics that are potentially related to unmeasurable earning determinants. To address this issue, researchers have attempted to control for the potential endogeneity of language skills. Chiswick and Miller (1995) find that IV returns to English proficiency for immigrants in Canada and the USA represent between 40 and 57%, whereas Chiswick and Miller (2003) report estimates within 26 and 42% in Canada. Bleakley and Chin (2004) provide probably the most convincing IV strategy for the return to language proficiency up to the date. In their paper, the IV estimates of the return to English-speaking ability in the USA are around 30%. Moreover, by relaying on IV, researchers mitigate the extent of bias arising from measurement error in the (subjectively assessed) language proficiency variables. In non-English-speaking countries, Dustmann and van Soest (2002) for Germany and several studies for Israel (Chiswick and Repetto 2001; Berman et al. 2003) likewise report positive impacts of host language proficiency on immigrant earnings. Again, estimates tend to be considerably higher under IV. For instance, Chiswick (1998) reports a figure above 35% for Hebrew fluency among migrants in Israel. Gao and Smyth (2011) analyze the return to standard Mandarin among internal migrants in China and find a 40% return to language proficiency. Finally, in Spain, Budría and Swedberg (2014) show that the IV returns to host language knowledge for immigrants represent roughly 20%.

A limitation of the aforementioned studies is that returns to host language proficiency are assumed to be evenly distributed across the earning distribution. This is an unrealistic interpretation. All individuals including immigrants differ in unobserved abilities and skills. To the extent that these skills determine their earning capacity and the return from host language proficiency, one must expect some degree of heterogeneity across the earning distribution. However, the available evidence is still scarce. In particular, Boyd and Cao (2009) use quantile regression to examine the returns to English and French language proficiency among Canadian immigrants. Their results show that returns tend to be higher at the upper quantiles of the wage distribution. Namely, relative to the non-proficient, high-earning immigrants experience a greater

wage premium than low-earning immigrants. Nevertheless, their results can be hardly interpreted as causal impacts, as they do not control for the potential endogeneity of the language variable nor for measurement error in the language variable. Wang and Wang (2011) address this issue by adopting an IVQR approach based on Bleakley and Chin's instrument. Using a sample of immigrants that arrived in the USA as children for the 1990 and 2000 cohorts, the authors find a significant degree of heterogeneity across the earning distribution. In particular, while returns to language proficiency for the 1990 cohort are higher at the bottom quantiles of the distribution, their findings for the 2000 cohort show less heterogeneous returns and a less clear pattern of variation across quantiles. Ginsburgh and Prieto-Rodriguez (2011) use international comparable data from the 1994–2001 waves of the European Community Household Panel to examine the returns to foreign languages on native workers' earnings in European countries. Their IVQR estimates suggest that the impact of foreign language proficiency is stronger at the top of the wage distribution for half of the countries, even though there are significant differences between countries. More recently, using the NIS, Isphording (2013) estimates the returns to foreign (English, German, and French) language skills for immigrants in Spain. Their IVQR results for English proficiency do not show a further degree of heterogeneity beyond the IV results. Interestingly, when native speakers are excluded, the returns to German and French language proficiency decrease as you move up the earning distribution.

3 Data and definition of variables

The data is taken from the National Immigrant Survey of Spain (NISS), a large-scale immigration survey carried out by the National Statistics Institute of Spain. The data was collected between November 2006 and February 2007 and is based on the Municipal Census (MC). The original survey sample comprises approximately 15,500 individuals. Despite some years have passed since its publication and there is only one available wave, the NISS provides unique information for the study of immigration phenomena. It contains data on the socio-demographic characteristics of immigrants and their previous and current employment status. Immigrants are defined as individuals born abroad (regardless of their nationality) who at the time of being interviewed had reached at least 16 years of age and had resided in a Spanish home for at least a year or longer or had the intention to remain in Spain for at least 1 year. For further information regarding the NISS, see Reher and Requena (2009).

The estimating sample consists of private sector men who are between 18 and 65 years old and work regularly between 15 and 70 h a week. Self-employed individuals, as well as those whose main activity status is paid apprenticeship, training, and unpaid family workers, have been excluded from the sample. Since only a fraction of women participates in the labor market, and this group may be not representative of the total population of women, women are disregarded on behalf of the extra complications derived from potential selectivity bias. Immigrants at the top and bottom 2% of the hourly wage distribution are dropped to reduce the influence of outliers. Observations from Spain's two autonomous cities, Ceuta and Melilla, located in Northern Africa, are also dropped due to potential problems of representation (0.6% of the initial sample). Dropping observations, including item non-response, leave us with a final sample of 3592 individuals.[2]

3.1 Spanish language proficiency

The Spanish language proficiency question on the NISS is:

- *Thinking of what you need for communicating at work, at the bank, with the public authorities/administration. How well do you speak Spanish?*

Available answers range from 1 ("very well") to 4 ("need to improve"). The responses were used to define SP, a dummy variable that takes value 1 if the immigrant is proficient in Spanish (1—very well), zero otherwise.[3] According to this criterion, nearly 65.5% of the sample reports being proficient in Spanish.

The use of subjective evaluations is standard in the field, partly due to the high costs of test-based assessments of language ability. Admittedly, respondents may have different perceptions under identical circumstances of how well they speak a foreign language. These, notwithstanding, subjective questions are typically found to be highly correlated with scores from tests designed to accurately measure language ability as well as functional measures of language skills (Akbulut-Yuksel et al. 2011)

It must be noted that respondents that report Spanish as a foreign language are also asked to self-assess (yes/no) whether they possess a satisfactory skill level in different language areas, including comprehension, speaking, reading, and writing.[4] Nevertheless, this information is provided on a yes/no basis, and as many as 99.7% (comprehension), 100% (speaking), 89.7% (reading), and 81.6% (writing) of the sample answer "yes" to the corresponding question. These figures are far higher than the 65.5% of language proficient immigrants that emerges from the central question used for this paper. Therefore, relying on these indicators provides a far less stringent criterion for Spanish language ability. As a consequence, the present paper does not attempt to differentiate between different types of language skills.

Table 1 provides summary statistics by language proficiency level. As is apparent, relative to the non-proficient, proficient immigrants earn higher wages, have higher levels of educational attainment, have lived in Spain for a longer period, and are more likely to have a child in school, even though they have a similar number of children living at home. As expected, they are also less likely to come from a non-Spanish-speaking country (42.5%, against 99.0% among the non-proficient). Moreover, proficient immigrants are more likely to work in the technological and science sector and in the manufacturing and construction sectors. There are also some differences in terms of geographical origin, with non-proficient immigrants being more likely to come from Northern Africa and Eastern Europe.

4 Estimation strategy

The paper adopts an IVQR approach. The main advantage of this strategy is that, under instrument validity and classical measurement error, it allows us to obtain causal effects, not mere correlations. Moreover, these effects can be examined across the earning distribution.

4.1 Quantile regression

The τth conditional quantile model of this paper can be written as

$$Q_{ln(w_i)}(\tau|X_i,\ \mathrm{SP}_i) = X_i\beta(\tau) + \gamma(\tau)\mathrm{SP}_i \tag{1}$$

where $\tau \in (0, 1)$ is the quantile being analyzed, X_i is the vector of $k - 1$ socio-demographic variables, and SP measures Spanish language proficiency. It should be noted that the

Table 1 Summary statistics by Spanish proficiency

	Proficient	Non-proficient
Share	0.653	0.347
Hourly wage	7.682	6.339
	4.699	3.259
Years of schooling	11.140	8.614
	3.586	4.830
Age	38.004	36.340
	9.277	8.510
Age at arrival	24.520	28.310
	12.150	7.564
Not from Spanish-speaking country	0.425	0.990
	0.494	0.098
Single	0.357	0.333
	0.479	0.471
Divorced	0.066	0.076
	0.248	0.264
Married	0.577	0.591
	0.494	0.492
Children in school	0.375	0.271
	0.484	0.445
No. of children at home	0.663	0.633
	0.473	0.482
Region of origin		
Maghreb	0.087	0.320
	0.282	0.467
Sub-Saharan Africa	0.023	0.083
	0.148	0.275
Eastern Europe	0.076	0.355
	0.266	0.479
Latin America	0.603	0.035
	0.489	0.183
Asia	0.012	0.061
	0.107	0.239
Australia-North America	0.009	0.007
	0.096	0.082
Central and Western Europe	0.197	0.147
	0.398	0.354
Occupation sector		
Army	0.002	0.001
	0.045	0.028
Management	0.056	0.023
	0.229	0.151
Technology and sciences	0.181	0.093
	0.385	0.291

Table 1 Summary statistics by Spanish proficiency *(Continued)*

Services	0.142	0.097
	0.349	0.296
Administration	0.044	0.011
	0.206	0.102
Agriculture and fishery	0.221	0.329
	0.415	0.470
Manufacturing and construction	0.150	0.067
	0.358	0.251
Unqualified occupations	0.203	0.379
	0.402	0.485

Note: (a) Source: Spanish National Immigrant Survey; (b) standard deviations are in smaller type

return to Spanish proficiency, $\gamma(\tau)$, is allowed to vary with τ. $Q_{\ln(w_i)}(\tau|X_i, \mathrm{SP}_i)$ denotes the τth conditional quantile of ln w given X and SP. To estimate $\beta(\tau)$ and $\gamma(\tau)$, the τth regression quantile coefficients, the following minimization problem is solved (Koenker and Bassett 1978):

$$Q_{ln(w)}(\tau|X_i, \mathrm{SP}_i) = arg\ \ minE\left[\rho_\tau(ln(w_i) - X_i\beta(\tau) - \gamma(\tau)\mathrm{SP}_i)\right] \qquad \beta(\tau), \gamma(\tau) \tag{2}$$

where $\rho_\tau(u)$ is the loss function, defined as $\rho_\tau(u) = u(\tau - I[u < 0])$, being $I[\cdot]$ an indicator function. This optimization problem can be alternatively written in the more straightforward form of Eq. (3):

$$\mathrm{Min}_{\beta_{(\tau)}} \in R^k, \gamma_{(\tau)} \in R \left\{ \sum_{i:\ \ln w_i \geq X_i\beta(\tau)+\gamma(\tau)\mathrm{SP}_i} \tau|lnw_i - X_i\beta(\tau) - \gamma(\tau)\mathrm{SP}_i| + \sum_{i:lnw_i < X_i\beta_{(\tau)}+\gamma(\tau)\mathrm{SP}_i} (1-\tau)|lnw_i - X_i\beta(\tau) - \gamma(\tau)\mathrm{SP}_i| \right\} \tag{3}$$

Since the loss function is piecewise linear, Eq. (3) is actually a linear programming problem that can be solved using linear programming methods. Standard errors for the vector of coefficients are obtainable by using the bootstrap method described in Buchinsky (1998).

4.2 Instrumental variable quantile regression

Unbiasedness (and consistency) of quantile regression estimates relies firmly on the assumption of exogeneity of the regressors. However, language ability may depend on unobservable individual characteristics that are potentially related to unmeasurable earning determinants. That would be the case if, for example, more productive and capable individuals are more likely to be proficient in Spanish. In this case, the estimated coefficients would be biased and would not reflect the true benefits derived from language proficiency. At the same time, self-reported measures of speaking fluency typically suffer from measurement error, which leads to biased OLS and QR estimates.

Chernozhukov and Hansen (2008) developed an instrumental variable quantile regression procedure (IVQR) that relaxes the exogeneity assumption. The IVQR approach relies on the existence of instrumental variables Z that are related with the Spanish language proficiency variable (SP) but not with the error term. It is also worth noting that the use of an IV procedure is also intended to reduce the extent of

attenuation bias that may stem from errors in the measurement of the individual's self-assessed Spanish language proficiency SP.

The assumption of independence between the error terms, X and Z, implies an important moment restriction to obtain the IVQR estimator. From the definition of the τth conditional quintile shown in (4)

$$P\left[ln(w)\leq Q_{ln(w)}(\tau|X,\mathrm{SP})|X,Z\right]=\tau \tag{4}$$

Substituting (1) in (4), it becomes

$$P[ln(w)-X\beta(\tau)-\gamma(\tau)\mathrm{SP}\leq 0|X,Z]=\tau \tag{5}$$

This moment condition implies that the value of the τth conditional quantile of $\ln(w) - X\beta(\tau) - \gamma(\tau)\mathrm{SP}$ is zero, an equality that is the main equation for identification. The IVQR estimator for $\beta(\tau)$ and $\gamma(\tau)$ can be obtained by solving the following minimization problem:

$$\underset{\beta(\tau),\gamma(\tau),\lambda(\tau)}{arg\ min}\ E\left[\rho_\tau\left(ln(w_i)-X_i\beta(\tau)-\gamma(\tau)\mathrm{SP}_i-\hat{Z}_i\lambda(\tau)\right)\right] \tag{6}$$

In this equation, $\hat{Z}_i$ is the linear projection of SP_i on X_i and Z_i, i.e., the fitted values for SP obtained from auxiliary regression (7):

$$\mathrm{SP}_i=X_i\delta+Z_i\mu+\nu_i \tag{7}$$

For each τ, as $\hat{\beta}(\tau)$ and $\hat{\gamma}(\tau)$ converge in probability to $\beta(\tau)$ and $\gamma(\tau)$, respectively, $\hat{\lambda}(\tau)$ converges in probability to 0. Therefore, the estimator for the parameter $\gamma(\tau)$ can be obtained by choosing a value that drive the estimates of λ as close to zero as possible, as described in Wang and Wang (2011). The estimation algorithm in practice involves repeating j times the following two steps over a grid of potential values for $\gamma(\tau)$: (1) for a given value of $\gamma(\tau)^j$, run an ordinary quantile regression of $\ln(w_i) - \gamma(\tau)^j\mathrm{SP}_i$ on X_i and Z_i to obtain the estimates $\hat{\beta}(\gamma^j(\tau),\tau),\hat{\lambda}(\gamma^j(\tau),\tau)$, and (2) save the Wald statistic, W^j, to test whether $\lambda(\gamma^j(\tau),\tau)=0$. Finally, the value of $\hat{\gamma}^j(\tau)$ that minimizes W is the IVQR estimate of $\gamma(\tau)$ and the corresponding $\hat{\beta}(\gamma^j(\tau))$.

4.3 Model variables and instruments

Variable w is hourly earnings and vector X includes educational attainment, age and its square, age at arrival, years since migration, marital status (single, divorced or widowed, reference: married), number of children living at home, the region of residence (there are 17 autonomous communities), a dummy variable for being born in a non-Spanish-speaking country, and the immigrant's source geographical region (Eastern Europe, Northern Africa, Sub-Saharan Africa, Latin-America, Asia, Australia-North America, reference: Western Europe). The choice of these variables is duly motivated by the immigration literature. We also include occupational dummies (according to the one digit level National Classification of Occupations).

Finding valid instrumental variables is not trivial. Instruments must be *exogenous* (i.e., uncorrelated with earnings) and *relevant* (i.e., they must account for a significant variation in SP). In this paper, we build upon Bleakley and Chin (2004) and use the interaction term between age at arrival and a dummy variable showing non-Spanish-

speaking country of birth.[5] Age at arrival is negatively correlated with language knowledge, since younger children learn languages more easily than adolescents and adults. Cognitive scientists refer to this as the *critical period hypothesis* according to which there is a critical age range in which individuals learn languages more easily. According to the literature, the critical age range stands between 5 and 15 years (Chiswick et al. 2008). However, age at arrival itself cannot be an instrument, since early arrival fosters better knowledge of the host society and cultural convergence and, therefore, may lead to higher future wages. Therefore, conditioning on the interaction term between age at arrival and non-Spanish-speaking country of birth allows us to partial out the non-language effects of early arrival.[6] This occurs because upon arrival in Spain, immigrants from Spanish-speaking countries experience everything that immigrants from non-Spanish-speaking countries encounter, except for learning a new language. Instrument validity requires that non-language age-at-arrival effects on labor market performance are the same for the two types of immigrants. This validity is well grounded on previous research (Bleakley and Chin 2004; Wang and Wang 2011) and reinforced by the inclusion of region of origin in the earning equation, an information that partially factors out potential differences in the non-language effects met by immigrants from different countries. As we shall see, the selected instrument is highly significant in the first stage equation and, jointly with our second instrument, passes well the validity tests.

The second instrument captures whether the respondent has a child in school living at home. Arguably, the child's interaction with natives and his school attendance are crucial factors that contribute to develop his destination language communication abilities. Upon arrival, immigrant children in school are probably more likely to learn the destination language more quickly. At the same time, parents' exposure to communication with their children in the destination country's language and access to their children's superior pronunciation and grammar skills acts as a transmission mechanism. The validity of this instrument seems well grounded a priori, for there is no presumption that apart from linguistic effects, having children in school fosters parental earnings.

5 Results

Table 2 shows the ordinary OLS and QR estimates. According to our OLS results, being proficient in Spanish increases wages by 7.0%. Nevertheless, the QR estimates are suggestive of some differences across quantiles. While the impact of Spanish language proficiency fails to be statistically significant at the bottom two deciles of the distribution, it is well defined and above 8.5% at the upper half of the earning distribution. An F test for the equality of coefficients indicates that differences across quantiles are statistically significant (p value = 0.01). We will re-examine this pattern once we control for the endogeneity of the language proficiency variable SP in the following section.

It is noteworthy to unveil the role of the remaining covariates included in the equation. Since the OLS results reported on Table 2 are consistent with the bunch of the literature and have been discussed earlier, we mostly focus on the QR estimates. The returns to education and host language proficiency seem to follow a similar pattern. That is, the impact of an additional year of education becomes stronger as we move upwards along the earning distribution from 0.4% at the bottom decile to almost 1% in the upper segments of the distribution. Although the fluctuation is small, the increasing returns along the wage distribution are consistent with previous findings by

Table 2 OLS and QR estimates

	OLS	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
Spanish language proficiency	0.070***	0.014	0.021	0.045**	0.069***	0.081***	0.093***	0.098***	0.088***	0.098***
	0.014	0.032	0.024	0.021	0.020	0.019	0.017	0.017	0.021	0.033
Years of schooling	0.008***	0.004**	0.004**	0.002	0.005**	0.008***	0.010***	0.009***	0.009***	0.008***
	0.001	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.002	0.002
Age	0.005	0.024***	0.013**	0.005	0.004	0.001	0.001	0.000	0.003	0.001
	0.004	0.009	0.006	0.006	0.005	0.006	0.005	0.005	0.008	0.007
Age^2 (×100)	−0.041	−0.312***	−0.167**	−0.062	−0.026	0.025	0.020	0.033	−0.040	0.012
	0.052	0.121	0.080	0.075	0.066	0.082	0.071	0.069	0.113	0.089
Age at arrival	−0.003***	−0.002***	−0.003***	−0.003***	−0.003***	−0.003***	−0.003***	−0.003***	−0.004***	−0.006***
	0.001	−1.780	0.001	0.001	−3.670	0.001	−3.740	0.001	0.001	0.001
Single	−0.047***	−0.048**	−0.065***	−0.084***	−0.061***	−0.054***	−0.045***	−0.037**	−0.027	−0.051**
	0.012	0.022	0.018	0.015	0.013	0.014	0.014	0.015	0.020	0.022
Divorced	−0.034	0.001	0.033	0.004	−0.013	−0.046**	−0.059***	−0.073***	−0.017	−0.031
	0.022	0.056	0.036	0.027	0.022	0.019	0.014	0.027	0.043	0.051
No. of children	0.005	−0.002	0.007	0.000	0.001	−0.003	−0.006	−0.005	0.000	0.004
	0.006	0.011	0.011	0.007	0.006	0.006	0.006	0.007	0.008	0.012
Not from Spanish-speaking country	−0.031	0.054	0.096	0.049	−0.001	−0.055*	−0.088***	−0.155***	−0.145**	−0.069
	0.034	0.104	0.049	0.032	−0.040	0.032	0.030	0.043	0.058	0.060
Region of origin										
Maghreb	−0.145***	−0.107***	−0.165***	−0.182***	−0.175***	−0.176***	−0.163***	−0.154***	−0.143***	−0.104***
	0.019	0.032	0.036	0.027	0.025	0.021	0.017	0.022	0.036	0.038

Table 2 OLS and QR estimates *(Continued)*

Sub-Saharan Africa	−0.134***	−0.217***	−0.210***	−0.200***	−0.190***	−0.171***	−0.122***	−0.094***	−0.109***	−0.093*
	0.029	0.048	0.052	0.038	0.031	0.032	0.036	0.033	0.039	0.055
Eastern Europe	−0.095***	−0.146***	−0.143***	−0.135***	−0.110***	−0.095***	−0.073***	−0.064**	−0.066**	−0.049
	0.020	0.040	0.030	0.025	0.024	0.025	0.022	0.026	0.031	0.033
Asia	−0.079**	−0.199**	−0.124*	−0.103**	−0.082**	−0.080**	−0.048	−0.044	0.006	0.036
	0.035	0.069	0.072	0.041	0.036	0.038	0.039	0.038	0.061	0.064
Latin America	−0.144***	−0.164***	−0.180***	−0.174***	−0.179***	−0.189***	−0.176***	−0.159***	−0.132***	−0.095***
	0.016	0.029	0.025	0.022	0.017	0.020	0.016	0.021	0.028	0.030
Australia-North America	−0.036	−0.015	−0.098	−0.049	−0.095	−0.103	−0.053	−0.029	−0.036	−0.059
	0.058	0.159	0.108	0.087	0.084	0.091	0.088	0.082	0.130	0.120
Occupation sector										
Army	0.339***	0.307*	0.362**	0.228	0.316*	0.275	0.177	0.450	0.378	0.524***
	0.126	0.157	0.149	0.146	0.171	0.218	0.261	0.297	0.243	0.166
Management	0.295***	0.006	0.072	0.122**	0.248***	0.321***	0.400***	0.481***	0.567***	0.599***
	0.028	0.081	0.047	0.057	0.068	0.067	0.066	0.067	0.055	0.055
Technology and sciences	0.386***	0.350***	0.340***	0.340***	0.369***	0.369***	0.51***	0.355***	0.417***	0.537***
	0.019	0.044	0.034	0.028	0.027	0.023	0.028	0.032	0.044	0.039
Services	−0.022	−0.051*	−0.039	−0.024	−0.019	−0.037**	−0.044**	−0.018	−0.008	0.006
	0.018	0.030	0.003	0.023	0.020	0.017	0.020	0.020	0.025	0.029
Administration	0.140***	0.144***	0.099***	0.092**	0.110***	0.128***	0.141***	0.130***	0.177***	0.239***
	0.030	0.043	0.029	0.038	0.041	0.034	0.036	0.036	0.049	0.082
Agriculture and fishery	0.174***	0.166***	0.147***	0.168***	0.187***	0.183***	0.169***	0.143***	0.140***	0.169***
	0.014	0.023	0.020	0.018	0.016	0.013	0.016	0.017	0.023	0.028

Table 2 OLS and QR estimates *(Continued)*

Manufacturing and construction	0.187***	0.234***	0.207***	0.239***	0.252***	0.217***	0.176***	0.133***	0.076***	0.100
	0.019	0.039	0.030	0.024	0.018	0.015	0.016	0.018	0.023	0.027
Constraint	1.589***	0.985***	1.329***	1.556***	1.584***	1.671***	1.693***	1.841***	1.869***	1.974***
	0.092	0.183	0.118	0.121	0.114	0.126	0.110	0.129	0.180	0.142
R^2 no. of observations	0.300	0.126	0.135	0.169	0.190	0.203	0.227	0.223	0.237	0.257
	3592	3592	3592	3592	3592	3592	3592	3592	3592	3592

Note: (i) Source: Spanish National Immigrant Survey; (ii) heteroskedastic-robust standard errors are in smaller type; (iii) additional controls: 16 dummies for Spanish autonomous communities
*denotes significant atthe 10% level, **denotes significant at the 5% level, ***denotes significant at the 1% level

Chiswick et al. (2008) and with several regularities highlighted by the authors: (i) the fact that over-education is more prevalent among low-income workers, especially in the case of immigrants; (ii) high ability and motivation as omitted variables are more prevalent among highly educated workers; and (iii) school quality variations are positively associated with educational attainment.

Furthermore, age is not significantly related with wages at most segments of the distribution. However, due to assimilation effects, age at arrival is a significant determinant of wages. Late arrivers earn lower wages, this effect being stronger at the top quantile of the wage distribution. Marital status exhibits a diverging pattern across the earning distribution, with singlehood being more closely related with earnings in the bottom quantiles and divorce being significant only at the intermediate quantiles of the distribution. On average, having children is not significantly related with earnings, and this pattern holds for all segments of the distribution. There are conspicuous earning differentials between immigrants from different geographical regions. Relative to the reference individual (an immigrant from Central-Western Europe), workers from Maghreb, Sub-Saharan Africa, Eastern Europe, America, and Asia reap significantly lower earnings. The predicted wage penalty ranges between 7.9% for Asians and 14.4% for Latin-American immigrants. It is very interesting to note that in most cases, this penalty tends to be stronger at the bottom quintiles of the distribution. Thus, for example, the average 7.9% penalty for Asians masks a 19.9% wage decrease at the bottom quantile and a non-significant impact at the upper part of the earning distribution. This illustrates a general pattern: in most cases, geographical origin penalty is less significant and substantially weaker at the upper quantiles, relative to the bottom quantiles and the average estimates. In other words, unlike average and low-earning workers, workers at high-paying jobs are not affected much by their region of origin. A candidate explanation has to do with bargaining power. According to the latest data (Eurostat, 2015), the unemployment rate for immigrants is 9.5% above the unemployment rate for natives. Given that the unemployment rate is higher among individuals endowed with poor labor market credentials and low-earning potential, these non-EU workers may experience a greater wage penalty due to weaker bargaining power. In this respect, region of origin may act as a screening device among firms employing workers with low-earning potential. Moreover, Spain also experiences the highest level of overqualification among foreign-born workers in OECD countries (OECD, 2014). As result of the greater skill and qualification mismatch among non-EU workers and since this mismatch is more likely to happen among low-income individuals, these individuals are more likely to experience a wage penalty with respect to Western European workers.

Finally, the results suggest roughly 30–40% higher earnings among workers in the army, management, and technology and science sectors, relative to the reference category "Unqualified occupations." Administrative workers and those working in the agricultural and fishing and manufacturing and construction sectors carry a lower despite significant premium. Again, the QR estimates uncover a substantial amount of heterogeneity in many cases. Thus, for example, the wage increase associated with the managerial sector ranges from a non-significant 0.6% at the bottom quantile to a 59.9% increase at the top quantile, the average being 29.5%. A similar pattern applies to other sectors, including technology and Science. Manufacturing and construction also shows some differences across the distribution, although in this case, the coefficient is

decreasing as we move along the wage distribution. A candidate explanation is that workers at higher quantiles, with stronger unobserved skills, can reap a larger return from white-collar occupations, including management and technology and science. This is consistent with the idea that information technology favors the brain rather than brawn. In contrast, workers with weaker skills and a lower earning capacity obtain higher return from blue-collar occupations, relative to those workers with stronger unobserved skills.

5.1 IVQR estimates

The returns to language proficiency presented so far assume that SP is an exogenous variable and, thus, tell us little about causal effects. As a result, we will report IV and IVQR regression results in Table 3. We focus on the Spanish language proficiency coefficient SP, since the coefficients for the remaining covariates of the model present very little variation relative to the previous OLS estimates. We also report the ordinary OLS and QR estimates on Table 3 for the sake of comparison.

The IV estimates suggest that assuming exogenous SP yields a downward-biased prediction. Specifically, when we switch from OLS to IV, the impact of Spanish language proficiency on immigrant wages increases from 7.0 to 17.2%. This outcome suggests that detaching from language endogeneity may largely underestimate the true returns to host language proficiency. Therefore, the upward bias of the OLS estimate due to potential ability-bias story is more than offset by the severe downward bias associated with measurement error in the Spanish language variable. Next, we turn to the crux of our analysis. The estimates at different quantiles allow us to test whether the impact of destination language proficiency on earnings differs across segments of the earning distribution. The most remarkable finding is that the impact of language proficiency on wages cannot be regarded as constant across the earning distribution. Switching from the bottom to the top quantiles, the estimated effect rises from being statistically non-significant to a maximum of 30.6% in the top quintile. In other words, the average IV estimate reported for the previous section, 17.2%, masks a substantial amount of heterogeneity across the distribution.

For illustrative purposes, Fig. 1 depicts the quantile-return profile, along with its 95% confidence interval and the average estimate. For the sake of comparison, the results from the simple QR model are also depicted. The difference is striking. While QR

Table 3 IV and IVQR estimates

	OLS	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
		QR								
Exogenous SP	0.070***	0.014	0.021	0.045**	0.069***	0.081***	0.093***	0.098***	0.088***	0.098***
	0.015	0.032	0.024	0.021	0.020	0.019	0.017	0.017	0.021	0.033
	IV	IVQR								
Endogenous SP	0.172***	0.12	0.049	0.137***	0.216***	0.280***	0.229***	0.241***	0.252***	0.306***
	0.040	0.066	0.054	0.049	0.047	0.047	0.047	0.050	0.055	0.071

Note: (i) Source: Spanish National Immigrant Survey; (ii) heteroskedastic-robust standard errors are in smaller type; (iii) additional controls: educational attainment, age and its square, age at arrival, marital status, number of children at home, non-Spanish-speaking country of origin, occupational dummies, the immigrant's source geographical region, and dummies for region of residence in Spain; (iv) no. of obs. = 3592

**denotes significant at the 5% level, ***denotes significant at the 1% level

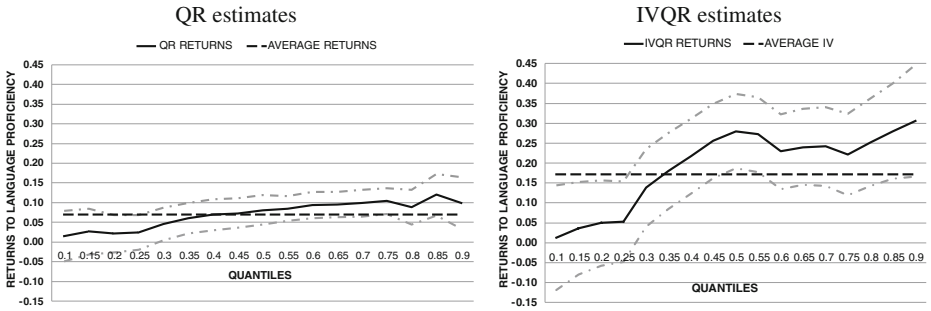

Fig. 1 QR and IVQR estimates—total sample. Note: (i) *dotted line* in QR returns denotes non-significant at those quintiles, (ii) *upper and lower curves* comprise the 95% confidence interval, and (iii) models with occupation controls

estimates are low and show a moderate level of dispersion, the IVQR model yields very heterogeneous returns across the distribution and a markedly increasing quantile-return profile. Workers at the bottom quantile fail to reap a reward from Spanish proficiency, while workers at the top quintile reap a return that almost doubles the return earned by the average worker. As a result, ordinary quantile regression not only underestimates the true returns to host language skills but also underestimates the extent of variation in the returns across the earning distribution.

In Table 4, we exclude occupation dummies from the estimating equation, as these variables can be regarded as potentially endogenous. Different occupations require different communication skills or, to put it differently, language proficiency may be a determinant of occupational selection. Consistent with this view, the occupation channel has been found to be important in explaining the earning effects of language skills (Wang and Wang 2011). We find that excluding occupation controls yields sensitively lower IV returns to Spanish proficiency. These are, on average, 12.4%, fail to be significant at the first four quintiles of the earning distribution, and reach a maximum of 25.4% at the top quintile. As a result, the upward profile of the return-quantile curve is again very apparent.

Immigrants with superior host language skills are expected to access better paid occupations. However, our findings seem to be at odd with this notion, since excluding occupation controls yields lower returns to Spanish proficiency. This finding suggests that some immigrants with superior Spanish language skills end up in low-paying

Table 4 IV and IVQR estimates—no occupation controls

	OLS	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
		QR								
Exogenous SP	0.076***	0.017	0.041**	0.031*	0.043***	0.073***	0.085***	0.096***	0.132***	0.128***
	0.016	0.027	0.018	0.018	0.021	0.021	0.019	0.021	0.026	0.036
	IV	IVQR								
Endogenous SP	0.124***	−0.017	0.050	0.014	−0.007	0.171***	0.161***	0.195***	0.190***	0.254***
	0.039	0.067	0.055	0.050	0.048	0.047	0.048	0.050	0.054	0.067

Note: (i) Source: Spanish National Immigrant Survey; (ii) heteroskedastic-robust standard errors are in smaller type; (iii) additional controls: educational attainment, age and its square, age at arrival, marital status, number of children at home, non-Spanish-speaking country of origin, occupational dummies, the immigrant's source geographical region, and dummies for region of residence in Spain; (iv) no. of obs. = 3592

*denotes significant atthe 10% level, **denotes significant at the 5% level, ***denotes significant at the 1% level

occupations. A candidate explanation for these results is that in Spain, occupational sorting is more importantly driven by education credentials than by host language skills. Highly educated individuals may manage to join high-paying occupations regardless of their knowledge of Spanish (e.g., tourism industry and/or international corporations where English is the main language). In contrast, low-educated individuals may be precluded from entering high-paying occupations, regardless of their knowledge of the host language.[7] This explanation is consistent with previous research by Di Paolo (2011), who finds that education and not Catalan language proficiency is the main channel for entering high-skilled occupations in the Spanish region of Catalonia. Moreover, in Spain, a country with historically above average unemployment rates, language knowledge may determine the sorting of immigrants between employment and unemployment, whereas education may act as a screening device to access high-skilled occupations. Although testing these hypotheses is beyond the scope of the present paper, we may lucubrate that, on average, immigrants with stronger Spanish language proficiency earn more, but the increase in earnings is greater if they work in occupations that require higher educational attainment.

5.2 IVQR estimates within education groups

To provide a more detailed view, and on the account for the potential complementarity between schooling and language skills (Chiswick and Miller 2003; Casale and Posel 2011; Budría and Swedberg 2014), in this section, we conduct separate regressions for immigrants with different educational attainment.

The results are reported in Table 5. To avoid problems derived from small cell size, two broad educational categories are considered. Panel 1 shows the result for immigrant workers that completed at least upper secondary education, whereas panel 2 is concerned with immigrants with less than upper secondary education. Splitting the sample as opposed to including a language proficiency-schooling interaction term is intended to allow for disparate endogeneity and earning-determination processes within the two groups.

Several conclusions emerge from the results. Firstly, returns to Spanish language proficiency for migrants with an upper secondary or higher education degree are remarkably high. More specifically, highly educated individuals that possess strong Spanish language skills are expected to earn, ceteris paribus, 32.1% more than highly educated individuals with a limited knowledge of the destination language. This complementarity between host language skills and schooling is not new and has been documented, although to varying degrees, in previous research. Secondly, we find that in both educational categories, standard OLS and QR estimates substantially underestimate the impact of language proficiency on wages and the extent of variation across quantiles. Thirdly, there is a large degree of heterogeneity with regard to IVQR returns to language proficiency within the highly educated category. In this group, workers at the two bottom quantiles fail to reap a sizeable return, while workers at the top quantile earn an impressive return of almost 50%. A very similar picture emerges from a model without occupation dummies (Table 6).

By contrast, immigrants with less than upper secondary education do not benefit, on average, from language skills. In this group, the average IV return is low and non-significant. However, discriminating among quantiles, we find an interesting pattern.

Table 5 Estimates by educational attainment

	OLS	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
		QR								
Upper sec. or more exogenous SP	0.110***	0.103**	0.105***	0.098***	0.113***	0.125***	0.134***	0.122***	0.133***	0.111***
	0.021	0.051	0.027	0.028	0.027	0.028	0.030	0.035	0.035	0.028
	IV	IVQR								
Endogenous SP	0.321***	−0.037	0.154	0.386***	0.449***	0.565***	0.468***	0.461***	0.35***	0.478***
	0.075	0.121	0.095	0.090	0.088	0.090	0.088	0.092	0.098	0.126
	OLS	QR								
Less than upper sec. exogenous SP	0.022	−0.039	−0.039	−0.027	0.029	0.049*	0.063**	0.076***	0.072***	0.104***
	0.021	0.045	0.039	0.038	0.037	0.028	0.026	0.022	0.022	0.038
	IV	IVQR								
Endogenous SP	0.065	0.012	0.077	0.079	0.0155***	0.133**	0.127**	0.122**	0.143**	0.182**
	0.047	0.078	0.064	0.057	0.055	0.053	0.054	0.057	0.065	0.079

Note: (i) Source: Spanish National Immigrant Survey; (ii) heteroskedastic-robust standard errors are in smaller type; (iii) additional controls: educational attainment, age and its square, age at arrival, marital status, number of children at home, non-Spanish-speaking country of origin, occupational dummies, the immigrant's source geographical region, and dummies for region of residence in Spain; (iv) no. of obs. = 2034 top panel, 1558 bottom panel

*denotes significant at the 10% level, **denotes significant at the 5% level, ***denotes significant at the 1% level

Table 6 Estimates by educational attainment, no occupational controls

	OLS	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$	$\tau = 0.9$
		QR								
Upper sec. or more exogenous SP	0.125***	0.086*	0.069**	0.080***	0.116***	0.115***	0.150***	0.168***	0.169***	0.166***
	0.022	0.046	0.033	0.032	0.030	0.028	0.031	0.033	0.029	0.035
	IV	IVQR								
Endogenous SP	0.301***	−0.08	0.007	0.111	0.172**	0.283***	0.461***	0.183***	0.381***	0.407***
	0.039	0.122	0.102	0.092	0.087	0.087	0.091	0.096	0.104	0.129
	OLS	QR								
Less than upper sec. exogenous SP	0.025	−0.051	−0.050	−0.023	−0.015	0.020	0.042	0.071***	0.086***	0.064
	0.022	0.041	0.041	0.035	0.039	0.037	0.029	0.026	0.027	0.043
	IV	IVQR								
Endogenous SP	0.072	−0.032	−0.028	0.086	0.121**	0.209***	0.118**	0.155***	0.128**	0.022
	0.045	0.076	0.060	0.055	0.052	0.052	0.052	0.055	0.061	0.072

Note: (i) Source: Spanish National Immigrant Survey; (ii) heteroskedastic-robust standard errors are in smaller type; (iii) additional controls: educational attainment, age and its square, age at arrival, marital status, number of children at home, non-Spanish-speaking country of origin, occupational dummies, the immigrant's source geographical region, and dummies for region of residence in Spain; (iv) no. of obs. = 2034 top panel, 1558 bottom panel
*denotes significant at the 10% level, **denotes significant at the 5% level, ***denotes significant at the 1% level

While immigrants at the bottom segments of the distribution fail to reap a sizable return, those at the intermediate and upper quintiles earn a significant 10% or more due to Spanish proficiency. These patterns are illustrated in Figs. 2 and 3 and remain once occupation dummies are dropped from the estimating equation.

All in all, we find that the high returns earned by educated workers with strong unobserved earning skills differ from the null return earned by workers with lower qualifications and/or weak unobserved earning skills. This finding suggests a strong degree of heterogeneity around the impact of host language proficiency on immigrants' earnings in Spain.

6 Discussion

6.1 Instrument quality

An important concern with IV is the validity of the instruments. Validity is not assured if the excluded instruments have direct effects on earnings beyond those flowing

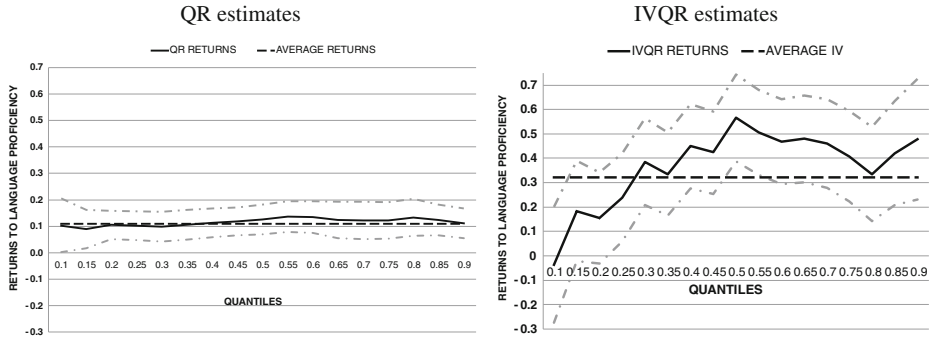

Fig. 2 QR and IVQR estimates—workers with at least upper secondary education. Note: (i) *dotted line* in QR returns denotes non-significant at those quintiles, (ii) *upper and lower curves* comprise the 95% confidence interval, and (iii) models with occupation controls

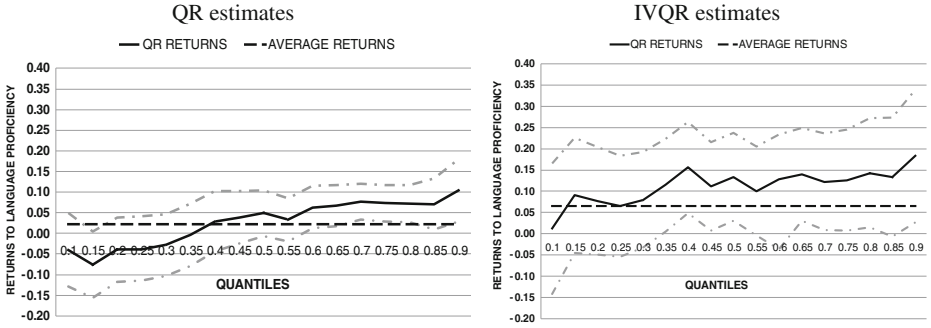

Fig. 3 QR and IVQR estimates—workers with less than upper secondary education. Note: (i) *dotted line* in QR returns denotes non-significant at those quintiles, (ii) *upper and lower curves* comprise the 95% confidence interval, and (iii) models with occupation controls

indirectly through Spanish proficiency. This problem may yield biased estimates and will be exacerbated by a weak correlation between the endogenous variable and the instruments (Bound et al. 1995). Even though the validity of the interaction between age at arrival and non-Spanish-speaking country of origin is well grounded on previous research, we performed two quality checks. Firstly, we tested econometrically whether the selected instruments are uncorrelated with the unexplained part of the earning equation. This was performed using the Sargan-Hansen test of orthogonality 2.35 ($p = 0.130$) suggesting that the two instruments used in the paper are not significantly related to the immigrant's earnings.

To examine the weak IV problem, two diagnosis tests were performed: the F test for the joint significance of the selected instruments (253.05, p value = 0.000) and their relative contribution to R^2 (0.162) in the second stage equation. These figures are well above the lower range of values considered admissible in the literature (about $F = 10$, see Stock et al. 2002) and suggest that the selected instruments account for a significant variation in the SP variable. Having such relevant instruments is very important to attenuate any potential bias arising from potential instrument non-validity. Finally, we tested for the exogeneity of the instrumented variable by means of the Durbin-Wu-Hausman statistic (7.29, p value = 0.007). The rejection of the null hypothesis suggests that SP cannot be regarded as exogenous and, therefore, the necessity of resorting to IV.

Finally, we also explored a just-identified model with the most significant variable as the only exclusion restriction. This was Bleakley and Chin's instrument (z-statistic = −21.91 in the first stage equation, against children in school age, z-statistic = 2.99). Just-identified models are median-unbiased, even with weak instruments (Angrist and Pischke, 2009, p. 209) and also deserve some consideration. Moreover, differences in compliant populations may explain variability in treatment effects from one instrument to another. The results (details available upon request) for the just-identified model and for the benchmark model used in this paper were very similar. Specifically, we found that moving from the over-indentified model to the just-identified model, the impact of Spanish proficiency on earnings decreased slightly from 17.2 to 16.4%. Similarly, this return was 31.2% for the high-educated category and a non-significant 5.9% for the low-educated category, against 32.1% and a non-significant 6.5%, respectively, in our benchmark model. The dispersion of the estimates across the earning distribution

presented also small variations, with non-significant returns at the lowest segments of the distributions and relatively high returns at the upper quantiles. These were generally above 20% in the total sample (26.6% in the top quantile), 40% in the high-educated group (despite a somewhat lower return in the top of the distribution, 27.2%), and 14% in the low-educated group (16.1% in the top quantile).

6.2 Theoretical implications

The results reported in the paper suggest that controlling for the endogeneity of host language proficiency is crucial. That is, standard estimates are likely to under-predict the impact of host language skills as well as the extent of variation across segments of the earning distribution. Previous research confirms these findings. In particular, Boyd and Cao (2009) find that in Canada, returns to host language proficiency tend to be higher for immigrants located at the upper quintiles of the wage distribution. The problem is that their results are likely to underestimate the extent of the dispersion, for they assume that language ability is exogenous. In order to adjust for endogeneity, Wang and Wang (2011) adopt an IVQR approach for the USA and find a large degree of heterogeneity across the 1990 cohort earning distribution of immigrants in the USA. However, their results using their most recent data, 2000 cohort, show very little variation across quantiles. In particular, the difference between the highest coefficient (37% at the 80th percentile) and the lowest coefficient (31% at the 60th percentile) is only 6%. Moreover, workers at the upper quantiles do not reap significantly larger returns from host language proficiency than workers at the bottom part of the distribution. On the contrary, our results show that the difference between the return for workers at the top and bottom of the earning distribution is as high as 42.3% for the full sample and 54.7% for workers that have completed at least upper secondary education. This suggests that the results are not consistent across countries. This notion finds support in Ginsburgh and Prieto-Rodriguez (2011), who use international comparable data from the 1994–2001 waves of the European Community Household Panel. Their IVQR estimates suggest that the impact of foreign language proficiency is stronger at the top section of the wage distribution for half of the countries analyzed, even though there are significant differences between countries. However, our results are in contrast with Isphording's (2013) findings, which are also based on the NISS. His IVQR estimates for English proficiency do not show a further degree of heterogeneity beyond the IV results. He also finds that when German and French native speakers are excluded, the returns to German and French language proficiency decrease as you move up the earning distribution. We interpret these observations as evidence that in Spain, returns to Spanish proficiency are much more heterogeneous than the returns to foreign languages.

The results presented in this paper suggest that there is a large degree of heterogeneity in the expected returns to host language proficiency in Spain, with returns actually rising considerably as we move up along the wage distribution. A candidate explanation for these results is unobserved ability. In the quantile regression framework, the estimates at different quantiles represent the effects of a given covariate for individuals that have the same observable characteristics but, due to unobservable earning capacity, are located at different quantiles of the conditional distribution. Breaking the labor market down by ability quantiles allows us to obtain estimates at different quantiles that provide snapshots of the impact of Spanish language knowledge for proficient individuals

within different ability groups. Our results show that high-ability individuals reap a substantially larger return from host language investments.

This finding provides a novel view on the interplay between language skills and ability. Earlier work points to relevant complementarities between schooling and language skills (Chiswick and Miller 2003; Casale and Posel 2011). In our research, we find that after controlling for educational attainment, unobserved ability still accounts for a significant degree of variation in the return that immigrants derive from host language knowledge. We use a broader definition of ability that includes all those unmeasured/unobserved characteristics that affect the worker's position along the earning distribution. In fact, technological change has increased the demand for high-skilled workers. Because of the complementarities between language skills and other forms of human capital mentioned earlier, the stronger demand for highly skilled workers suggests a rise in the demand for high-skilled workers with stronger destination language skills. As a result, the return to host language proficiency at the upper tail of the distribution is relatively high.

A complementary explanation is that immigrant enclaves where destination language proficiency is not as necessary may provide better opportunities for low-skilled immigrant with limited host language knowledge to find jobs within those neighborhoods (Wang and Wang 2011). Ethnic enclaves provide means for costless communication, labor market opportunities, and transportation. Within ethnic neighborhoods, immigrants can communicate in their native language and receive revenue from trading without the need and cost of learning a new language. Although we control for occupation sector, our categories are too broad to capture enclave-specific occupations.

7 Conclusions

This paper has shown that the impact of Spanish language proficiency on earnings cannot be well described in an average sense. This impact is remarkably stronger at the upper segments of the earning distribution, especially among the more highly educated immigrants. The estimates are based on an instrumental variable quantile regression (IVQR) approach that controls for the endogeneity of host language proficiency. This is achieved by means of two instruments that pass several validity tests well. As a consequence, the results obtained can be interpreted as causal effects. Ordinary quantile regression estimates tend to underestimate the true returns to language skills, and the level of underestimation differs greatly across the earning distribution.

We break down the labor market by ability quantiles, with individual ability indexed by the immigrant's position at the conditional distribution, and find that high-ability individuals reap a substantially larger return from investment in host language proficiency than low-ability individuals. This outcome highlights the substantial individual heterogeneity that surrounds the benefits that immigrants derive from destination language skills. The heterogeneity not only stems from differences in educational attainment (returns are substantially larger among more educated individuals) but also holds within education categories as well.

From a policy perspective, labor market integration and the strengthening of educational aspects, including destination language training, represent a priority for OECD countries (OECD 2012). In line with this view, the Spanish Strategic Plan for Citizenship and Integration acknowledges the fact that immigration poses specific challenges that

must be tackled, such as "the promotion of improvements in immigrants' knowledge of the official languages and social norms in Spain, prerequisites for a cohesive society and for the social integration of immigrants" (Ministry of Labour and Social Affairs 2007). The social and economic integration of immigrants has become an even more relevant issue that governments must deal with as a result of the current refugee crisis. The results obtained in this paper have important implications for policy makers, enabling them to target the immigrant population based on their needs and their future return to host language proficiency. Firstly, the scope attributed to such language policies may be more modest than presumed if less educated workers located at the lower segments of the wage distribution fail to reap relevant returns from destination language training. Researchers and policy makers should take this heterogeneity into consideration when attempting to determine the impact of language policies for different population categories. To that purpose, focusing on averages may be seriously misleading.

Secondly, the results also bear important implications on income inequality. More specifically, the outcome shows that the conditional wage distribution for proficient immigrants is more dispersed than the conditional wage distribution for non-proficient immigrants. As a consequence, if we provide language skills to workers who have the same observable characteristics but are located at different quantiles of the wage distribution, their wages will become more dispersed. To put it differently, host language proficiency is a source of earning inequality among immigrants. Therefore, offering indiscriminate language training programs in Spain may be counter-productive. Under the light of our results, destination language training directed towards immigrants at the lower segments of the earning distribution should be combined with courses that promote the development of marketable job skills intended to improve their labor market status.

Thirdly, language proficiency can be seen as a risky investment in human capital. In our setting, the risk is represented by the differences in the returns to host language ability across quantiles; as such, differences represent residual wage inequalities after controlling for the effect of skill differences by regression results. Our results show that to the extent that immigrants are not fully aware of the personal characteristics that place them at some point of the wage distribution, the returns to their investment in language skills are to some extent unpredictable. This uncertainty may have differential effects on the immigrant's willingness to learn the host language, depending on their attitude towards risk, subjective perceptions, and the cost of acquiring destination language skills.

Endnotes

[1]For instance, in Bleakley and Chin (2004), OLS estimates double after controlling for classical measurement error, while IV results present little variations after controlling for non-classical errors. In Dustmann and van Soest (2002), the effects of language on earnings rise by a factor of more than two in a model that controls for classical measurement error. However, Dustmann and van Soest (2004) show that in specific contexts, non-classical errors can lead to sensitive biases.

[2]The incidence of earnings non-response is relatively low, 4.6% compared to other microdatasets available. A particularity of the NISS is that some 19% of the respondents prefer to report the income information using intervals (10 alternatives are provided). We decided to keep them by recoding the income classes using the interval center. Excluding this subsample from the estimations produced similar results.

[3]The distribution of responses was the following: (1) "very well," 65.6%; (2) "well," 21.3%; (3) "sufficient," 7.4%; and (4) "need to improve," 5.8%. The paper follows a stringent criterion by considering only individuals who claim to be able to speak Spanish "very well." Results under the alternative classifications 1–2 against 3–4 displayed slightly lower returns and are available upon request.

[4]The precise wording of the questions is: "Comprehension? 1. Yes. 2. No; can you speak this language? 1. Yes. 2. No; can you read in this language? 1. Yes. 2. No; can you write in this language? 1. Yes. 2. No." Immigrants whose mother tongue is Spanish are not required to provide such information, the underlying assumption being that they are fully proficient in these four areas.

[5]We thank an anonymous referee for very insightful comments regarding instrument validity in a previous version of the paper and for drawing the attention on the benefits of Bleakley and Chin's approach.

[6]There are several functional forms to operationalize the interaction between age at arrival and the dummy for non-Spanish-speaking country of origin. These are described in Table A2 in Bleakley and Chin's paper. Specifically, the dummy for non-Spanish-speaking country can be interacted with (i) age at arrival, (ii) age at arrival—critical period threshold, (iii) age at arrival dummy variables, and (iv) a dummy for age at arrival before critical threshold. We have chosen (i) for reasons of simplicity. We found that alternative parameterizations yielded very similar results.

[7]Encompassing calculations support this view. We have modeled the probability of working in a specific occupation as a function of the socio-economic characteristics of the worker, including Spanish proficiency and education. While in most occupations, the sign of these two variables is the same (negative in low-skill and positive in high-skill occupations), there is a notable exception. The probability of working in technology and science, the best-paid occupation (see Table 2), depends crucially on education and very *negatively* on the Spanish-level proficiency.

Acknowledgements
We would like to thank the editor and an anonymous referee for the helpful comments. Santiago Budría acknowledges the financial support provided by the Spanish Ministry of Education through grants ECO2012-33993 and ECO2012-36480 and Aristos Campus Mundus Program through grant ACM2016_22. This paper uses the Spanish National Immigrant Survey (NISS), a large-scale immigration survey released by the Spanish National Statistics Institute. We thank the NISS data providers.
Responsible editor: Denis Fougère

Competing interests
The IZA Journal of Development and Migration is committed to the IZA Guiding Principles of Research Integrity. The authors declare that they have observed these principles.

Author details
[1]Department of Quantitative Methods, Universidad Pontificia Comillas, C/Alberto Aguilera 23, s/n, 28015 Madrid, Spain. [2]Department of Business & Economics, Saint Louis University, Avenida del Valle 34, Padre Arrupe Hall, 28003 Madrid, Spain.

Received: 25 October 2016 Accepted: 11 April 2017
Published online: 29 August 2017

References

Adsera A, Chiswick B. Are there gender and country of origin differences in immigrant labor market outcomes across European destinations? J Popul Econ. 2007;20(3):495–526.
Akbulut-Yuksel M, Bleakley H, Chin A. The effects of English proficiency among childhood immigrants: are Hispanics different? In: Leal D, Trejo SJ, editors. Latinos and the economy: integration and impact in schools, labor markets, and beyond. 2011.
Angrist J, Pischke S. Mostly harmless econometrics: an empiricist's companion. New Jersey: Princeton University Press; 2009.

Anton JI, Muñoz de Bustillo R, Carrera M. From guests to hosts: immigrant-natives wage differentials in Spain. Int J Manpow. 2010;31(6):645–59.
Bárcena E, Budría S, Moro-Egido AI. Skill mismatches and wages among European university graduates. Appl Econ Lett. 2012;19(15):1471–5.
Beenstock M, Chiswick BR, Paltiel A. Testing the immigrant assimilation hypothesis with longitudinal data. Rev Econ Househ. 2010;8(1):7–27.
Berman E, Lang K, Siniver E. Language-skill complementarity: returns to immigrant language acquisition. Labour Econ. 2003;10(3):265–90.
Billger S, Lamarche C. Immigrant heterogeneity and the earnings distribution in the United Kingdom and United States: new evidence from a panel data quantile regression analysis. Bonn: IZA DP 5260, IZA; 2010.
Bleakley H, Chin A. Language skills and earnings: evidence from childhood immigrants. Rev Econ Stat. 2004;86:481–96.
Bound J, Jaeger D, Baker R. Problems with instrumental variable estimation when the correlation between instruments and the endogenous explanatory variables is weak. J Am Stat Assoc. 1995;90(430):443–50.
Boyd M, Cao X. Immigrant language proficiency, earnings, and language policies. Can Stud Popul. 2009;36(1-2):63–86.
Buchinsky M. Recent advances in quantile regression models: a practical guideline for empirical research. J Hum Resour. 1998;33(1):88–126.
Budría S, Swedberg P. The impact of language proficiency on immigrants' earnings in Spain. Rev Econ Apl. 2014;23(67):63–91.
Casale D, Posel D. English language proficiency and earnings in a developing country: the case of South Africa. J Socio-Econ. 2011;40:385–93.
Chernozhukov V, Hansen C. Instrumental variable quantile regression: a robust inference approach. J Econ. 2008;142(1):379–98.
Chiswick B. Hebrew language usage: determinants and effects on earnings among immigrants in Israel. J Popul Econ. 1998;11:253–71.
Chiswick B Miller P. The Endogeneity between language and earnings: international analyses. J Labor Econ. 1995;13(2):246–288.
Chiswick B, Miller P. Language skills and earnings among legalized aliens. J Popul Econ. 1999;12:63–89.
Chiswick B, Miller P. The complementarity of language and other human capital: immigrant earnings in Canada. Econ Educ Rev. 2003;22(5):469–80.
Chiswick B, Miller P. Occupational language requirements and the value of English in the US labor market. J Popul Econ. 2010;23:353–72.
Chiswick B, Repetto G. Immigrants adjustment in Israel: the determinants of literacy and fluency in Hebrew and the effects on earnings' international migration: trends, policy and economic impact. New York: Routledge; 2001. p. 204–28.
Chiswick B, Le A, Miller P. How immigrants fare across the earnings distribution in Australia and the U.S. Ind Labor Relat Rev. 2008;61(3):353–73.
Di Paolo A. Knowledge of Catalan, public/private sector choice and earnings: evidence from a double sample selection model. Hacienda Pública Española. 2011;197(2):9–35.
Di Paolo A, Raymond JL. Language knowledge and earnings in Catalonia. J Appl Econ. 2012;15(1):89–118.
Dustmann C, Fabbri F. Language proficiency and labor market performance of immigrants in the UK. Econ J. 2003;113:695–717.
Dustmann C, van Soest A. Language and the earnings of immigrants. Ind Labor Relat Rev. 2002;55(3):473–92.
Dustmann C, Van Soest A. An analysis of speaking fluency of immigrants using ordered response models with classification errors. J Bus Econ Stat. 2004;22:312–21.
Eurostat (2015) Migrant integration statistics – labour market indicators. Available at: http://ec.europa.eu/eurostat/statisticsexplained/index.php/Migrant_integration_statistics_%E2%80%93_labour_market_indicators.
Eurostat (2016) Unemployment statistics. Available at: http://ec.europa.eu/eurostat/statisticsexplained/index.php/Unemployment_statistics.
Friedberg R. You can't take it with you? Immigrant assimilation and the portability of human capital. J Labor Econ. 2000;18(2):221–51.
Gao W. Smyth R. Economic returns to speaking "standard Mandarin" among migrants in China's urban labor market. Econ Educ Rev. 2011;30(2):342–352.
Ginsburgh VA, Prieto-Rodriguez J. Returns to foreign languages of native workers in the European Union. Ind Labor Relat Rev. 2011;64(3):599–618.
Hayfron J. Language training, language proficiency and earnings of immigrants in Norway. Appl Econ. 2001;33:1971–9.
Hu W-Y. Immigrant earnings assimilation: estimates from longitudinal data. Am Econ Rev Pap Proc. 2000;90:368–72.
Isphording I. Returns to foreign language skills of immigrants in Spain. Labour. 2013;27(4):443–61.
Koenker R, Bassett G. Regression quantiles. Econometrica. 1978;46:33–50.
Le A, Miller P. Glass ceiling and double disadvantage effects: women in the U.S. labour market. Appl Econ. 2012;42:603–13.
Lui HK. The returns to language ability in Hong Kong: before and after the handover. Appl Econ Lett. 2007;14(2):121–5.
McGuinness S, Bennett J. Overqualification and the graduate labour market: a quantile regression approach. Econ Educ Rev. 2007;26(5):521–31.
Ministry of Labour and Social Affairs (Ministerio de Trabajo y Asuntos Sociales) (2007): 'Strategic plan for citizenship and integration', Subdirección General de Información Administración y Publicaciones. Available at: http://extranjeros.empleo.gob.es/es/IntegracionRetorno/Plan_estrategico/pdf/PEClingles.pdf.
OECD. Migration policy developments, International Migration Outlook 2012. OECD: Publishing; 2012. Available at: http://dx.doi.org/10.1787/migr_outlook-2012-6-en. Accessed 16 June 2015.
OECD. Migration policy developments, International Migration Outlook 2013. OECD: Publishing; 2013. Available at: http://www.oecd.org/els/mig/imo2013.htm. Accessed 14 Nov 2015.

OECD. International Migration Outlook 2014. OECD: Publishing; 2014. Available at: http://www.oecd-ilibrary.org/social-issues-migration-health/international-migration-outlook-2014/overqualification-rates-among-the-highly-educated-in-employment-15-to-64-year-olds-by-migration-status-2013_migr_outlook-2014-graph26-en.
Reher D, Requena M. The National Immigrant Survey of Spain: a new data source for migration studies in Europe. Demogr Res. 2009;20:253–78.
Rendón S. The Catalan premium: language and employment in Catalonia. J Popul Econ. 2007;20:669–86.
Rivera-Batiz F. English proficiency and the earnings of young immigrants in the U.S. labor markets. Rev Policy Res. 1992;11(2):165–75.
Stock JH, Wright J, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat. 2002;20(4):518–29.
Wang C, Wang L. Language skills and the earnings distribution among child immigrants. Ind Relat. 2011;50(2):297–322.
Zhen Y. The Effects of English Proficiency on Earnings of U.S. Foreig-Born Immigrants: Does Gender Matter? J of Financ Econ. 2013;1(1):27–41.