Language Skill Acquisition in Immigrant Social Networks: Evidence from Australia

Jean-William P. Laliberté*

First version: December 2014; This version: March 2018

Abstract

This paper estimates the effect of linguistic enclaves on one of the most important determinants of both the economic and social integration of immigrants: language skills. Cross-sectional estimates of the relationship between linguistic concentration and English proficiency cannot distinguish learning from sorting on the basis of pre-existing language skills. Using rich longitudinal data, I find that enclave size significantly impedes language acquisition, albeit the effect is smaller than cross-sectional models suggest. An unusually rich set of variables is used to generate bounds on the effect of enclaves and a complementary instrumental variable approach confirms the robustness of the results. Enclaves are unrelated to formal language course take-up rates, indicating that they affect language learning via social interactions among friends and colleagues rather than through formal education.

1 Introduction

In countries with large foreign-born populations, immigrants tend to be spatially concentrated in a few metropolitan areas. Policy-makers and researchers posit that residential segregation may produce externalities affecting the economic and social outcomes of the cities in which ethnic enclaves are formed. The implementation of refugee placement and desegregation policies in many immigrant-receiving countries reflects these concerns. Enclaves might also affect immigrants' long-term economic assimilation: (Borjas, 2015)

*I am very grateful to Michael Baker, Dwayne Benjamin, Arthur Blouin, Julien Gagnon, Nicolas Gendron-Carrier, Michael Gilraine, Maripier Isabelle, Kory Kroft, Mathieu Marcoux, Robert McMillan, Romuald Méango, Juan Morales, Ismael Mourifié, Philip Oreopoulos, Scott Orr, Marc-Antoine Schmidt, Michel Serafinelli, and Aloysius Siow for helpful discussions and to seminar participants at the University of Toronto, CIREQ, and the Canadian Economics Association meeting for comments. I would also like to thank the Australian Department of Immigration and Citizenship for help with data access. All mistakes are my own. Email: jeanwilliam.lalibert@ucalgary.ca

demonstrates that about 20% of the decline in earnings convergence experienced by most recent cohorts of immigrants to the U.S. can be attributed to changes in the size of ethnic groups. These trends notably correlate strongly with slower rates of language acquisition, suggesting an important role for English proficiency as a mechanism.

Living in a linguistic enclave may diminish the incentives to invest in the acquisition of host-country language skills (Lazear, 1999). Similarly, highly segregated areas may provide fewer interactions with natives, and therefore make learning the dominant language more difficult. Any resulting lower language proficiency can hinder the economic assimilation of migrants, given that proficiency in the host-country language has a sizable economic return estimated to approximately 15% higher earnings in most English-speaking countries.[1] Moreover, qualitative evidence suggests that employers are strongly concerned about applicants' language skills when reviewing resumes (Oreopoulos, 2011). The consequences of lower proficiency rates may not be exclusively borne by the individuals who lack these language skills. Low fluency rates may have external effects on economic and social outcomes by giving rise to communication barriers between groups.[2] Linguistic isolation can also generate intergenerational externalities by reducing the incentives to acquire language skills for future cohorts of immigrants.

This paper estimates the effect of linguistic enclaves on one of the most important determinants of both the economic and social integration of immigrants: language skills. I focus on Australia, a nation with one of the highest shares of foreign-born population among developed countries, and with substantial segregation of immigrants in urban areas.[3] Using the Longitudinal Survey of Immigrants to Australia (LSIA), I directly track changes in language skills over time to explicitly measure language *acquisition*. Existing estimates rely on cross-sectional variation in immigrants' language skills, making it difficult to separate out learning from sorting: if immigrants make their location decisions on the basis of preexisting language skills, a strong correlation between linguistic concentration and proficiency in the host-country language may be found among newly arrived immigrants, that is even before any learning has possibly occurred in the host country.[4]

[1] Chiswick and Miller (2014) provide an excellent review of existing empirical estimates of the effect of dominant language proficiency on earnings. Also see Lewis (2013); Bleakley and Chin (2004); Berman, Lang and Siniver (2003); Dustmann and van Soest (2002, 2001).

[2] English proficiency may generate positive human capital externalities as in Borjas (1995) and Moretti (2004). Also, the equilibrium fluency rate may not be socially optimal, in that it might produce too little productive social interactions (Konya, 2007; Lazear, 1999; Church and King, 1993). Other possible externalities relate to the costs associated with the use of interpreters, which are not necessarily internalized by non-fluent individuals. More importantly, when professional interpretation services are not available, relying on ad hoc interpreters (e.g. family members) can have disastrous consequences, notably on the quality of health care services provided (Flores, 2006).

[3] In 2011, more than one out of every four Australians was foreign-born, the third highest ratio among OECD countries (OECD, 2013). That same year, two thirds of individuals speaking a language other than English at home resided in either Greater Sydney or Greater Melbourne, while less than 40% of the total Australian population lived in these two cities.

[4] In the United States, Lazear (1999) and Chiswick and Miller (2005) find a negative association between the probability that an immigrant speaks English well and the size of his linguistic group in his area. A negative relationship between ethnic/linguistic concentration and English language proficiency is also found in Australia (Chiswick and Miller, 1996), in Canada (Warman, 2007), and in the United Kingdom (Dustmann and Fabbri, 2003). Danzer and Yaman (2016) exploits West Germany's Guest-

This is particularly important in English-speaking countries, given that most immigrants possess English language skills to varying degrees before they migrate. To assess the robustness of my estimates, I use two complementary empirical strategies. The first one leverages an unusually rich set of observable characteristics associated with immigrants' abilities and their location decisions to generate bounds on enclave effects. The second approach focuses on sponsored immigrants, for whom information on their sponsor's location can be used as the basis for a instrumental variable.

I further contribute to the literature on enclave effects by examining different channels through which spatial concentration of language groups can affect the acquisition of language skills. The level of geography at which segregation matters the most for language skills is notably treated as an empirical question. Also, generally unavailable information on English language course take-up is used to investigate whether enclaves slow down language acquisition by reducing investments in language skills through formal education. Additional survey questions related to language used at work and job search strategies are used to investigate the link between enclaves, language acquisition, and economic incentives.

As a starting point, the paper first replicates the negative association between linguistic concentration and language skills documented in earlier studies. I then show that about a third of standard cross-sectional estimates actually reflects differences in pre-immigration proficiency rather than an effect on learning. My estimates suggest that a (within language-group) standard deviation increase in city-wide linguistic concentration reduces the probability of becoming fluent in English by 2.6 percentage points, a magnitude equivalent to roughly 15% of the total increase in proficiency observed over the survey's duration.[5] This result is very robust across specifications and empirical strategies. I also find that conditional on city-level enclave size, there is no incremental negative effect of residing in a relatively high-concentration neighborhood. Finally, I show that enclave size is unrelated to English course take-up, suggesting that social interactions outside the classroom – at work, in particular – are important in the learning process.

The remainder of the paper is organized as follows. Section 2 describes the survey data, and the empirical strategy is outlined in Section 3. The results are then presented in Section 4. Possible mechanisms are discussed in Section 5, and Section 6 concludes.

Worker Programme as an exogenous source of spatial variation, which allows them to isolate the causal effect of enclaves on learning. The immigrant population they study is substantially different than the one considered here, however. Danzer and Yaman (2016) focus on low-skilled immigrants who all had no prior knowledge of German.

[5]Note that low proficiency is not merely transitory phenomenon: in 2011, among the 1.3 million immigrants who had been living in Australia for at least 15 years and whose best spoken language is not English, one in five still spoke English "not well" or "not at all".

2 Data

2.1 The Longitudinal Survey of Immigrants to Australia (LSIA)

The primary dataset used in this paper is the Longitudinal Survey of Immigrants to Australia (LSIA1), a survey of a representative sample of recent immigrants undertaken by the Department of Immigration and Multicultural and Indigenous Affairs of Australia.[6] The LSIA1 sample consists of 5,192 Primary Applicants (PAs), which respresents approximately 7% of all PAs aged 15 or above who arrived in Australia between September 1993 and August 1995 and were offshore visaed.[7] In-depth personal interviews were conducted approximately five or six months (wave 1), 18 months (wave 2), and 42 months (wave 3) after the date of arrival in Australia. The questionnaire covers a comprehensive set of topics, notably including information about the reasons for immigrating to Australia, work status and educational attainment prior to immigration, place of residence in Australia, enrollment in English language courses in Australia, language best spoken, and English language skills.[8]

The longitudinal structure of the survey allows me to examine *changes* in language skills that occur with time spent in Australia. In each wave, respondents whose best spoken language is not English were asked to self-assess their ability to speak English on a four-point scale ranging from “Not at all” to “Very well”. In line with previous studies (Danzer and Yaman, 2016; Dustmann and Fabbri, 2003; Lazear, 1999), I collapse self-assessed language skills into a binary variable in most analyses. In particular, an individual is considered fluent in English if she reports speaking it “Well” or “Very well”. I complement this information with an arguably more objective measure of language skills – an indicator of whether the interview was conducted in English or not.[9]

This paper’s main analyses are based on a balanced panel of PAs who were interviewed in all three waves, whose best spoken language is not English, and who were of working age (18-64 years old) at the time of the first interview.[10] The analytical sample is further restricted to individuals for which I am able to produce measures of enclave size, which excludes smaller language groups. The final sample consists of 2,053 Primary

[6]Other work that also use the LSIA to examine immigrants’ English proficiency include Chiswick, Lee and Miller (2006) and Chiswick, Lee and Miller (2004).

[7]Immigrants with special eligibility visas, whose country of birth could not be identified, or who did not settle in a major urban area, as well as New Zealand citizens, were excluded from the survey. These subgroups only make up about four or five percent of all otherwise eligible PAs.

[8]Two other rounds of the LSIA were conducted on more recent cohorts of immigrants. The LSIA1 cohort arrived in Australia prior to the implementation, in 1999, of a major reform of the points test system which strengthened the English language ability requirements for two of the five main visa categories. English proficiency upon arrival was relatively weaker for LSIA1 participants than for LSIA2 and LSIA3 cohorts, leaving more room for language skill acquisition in the host-country (Chiswick, Lee and Miller, 2004; Cobb-Clark, 2004). LSIA1 also has a longer follow-up period than the later rounds.

[9]Interviews could be conducted with the assistance of friends or members of the respondent’s family, or with the help of an accredited interpreter or through a bilingual interviewer.

[10]The attrition rate between waves 1 and 3 is of 30%. Sampling weights are adjusted accordingly.

Applicants.[11]

2.2 Measurement: Linguistic concentration

The key independent variable of interest is linguistic concentration, which I measure using Bertrand, Luttmer and Mullainathan's (2000) *contact availability* ratio:

$$CA_{jk} = \left[\frac{\left(\frac{\text{Number of people of language-group } k \text{ in area } j}{\text{Number of people in area } j}\right)}{\left(\frac{\text{Total number of people of language-group } k \text{ in the country}}{\text{Total population of the country}}\right)} \right].$$

Deflating the proportion of individuals belonging to language-group k in area j by the share of that language-group in the entire country has the advantage of avoiding underweighting smaller language groups. It also has a clear interpretation as a segregation index: if language-group k makes up the same proportion of the population in area j as it does at the country-level, then the index will be equal to one. Consequentially, a contact availability measure above one implies that the language-group of interest is over-represented in the area, whereas an index below one means that the language-group is under-represented.[12]

The population counts of the Australian Census Community Profiles are used to compute measures of CA_{jk} for 31 non-English linguistic communities defined by language spoken at home in 1996.[13] The smaller geographic unit for which this is feasible is a Statistical Subdivision (SSD), a "socially and economically homogeneous region characterized by identifiable links between inhabitants" (Australian Bureau of Statistics, 1996). Linguistic concentration is also calculated for Statistical Divisions (SD), which are made up of one or more SSDs, and more or less correspond to MSAs in the United States and CMAs in Canada. For example, the SD of Sydney (population of 3,741,290 in 1996) consists of 14 SSDs (with an average population of 267,235). There is substantial variation in the size of SDs, with Sydney being the most populous, and some having a population as small as 44,798 (Pilbara, Western Australia). The individuals in my sample live in 16 Statistical Divisions (64 SSDs) in wave 3, and form 223 unique language-SD cells (671 language-SSD cells).

The choice of level of aggregation is not insignificant. One may argue that the social interactions that matter the most in the present context mainly take place at the neighborhood level. If this is the case, then, linguistic concentration measured at the SSD level should be a more accurate measure of immigrant social networks

[11] I ignore possible general equilibrium effects associated with the location decisions of the LSIA1 cohort, which only represent 2% of Australia's total non-English speaking population in 1996. In the language of Angrist (2014), there is a separation between the subjects of peer effects (the LSIA1 cohort) and the peers who provide a mechanism for causal effects on those subjects (neighbors, notably individuals from previous waves of immigration).

[12] As is customary in the literature on ethnic concentration, I use $\ln CA_{jk}$ in most econometric specifications. Because language-group fixed effects are included in the model, results are numerically equivalent to using the log of the *exposure index*: $EI_{jk} = 100 * \left(\frac{\text{Number of people of language-group } k \text{ in area } j}{\text{Number of people in area } j}\right)$.

[13] Results are qualitatively unchanged if country of birth is used instead of language spoken at home (available upon request).

than at the broader SD level. However, there are a few shortcomings to this approach that can be addressed by using a more aggregate measure. Firstly, measures of concentration for linguistic groups who make up a very small proportion of the population in certain areas may be subject to measurement error; using larger geographic units should attenuate this problem (Danzer and Yaman, 2013). Secondly, smaller geographic units do not allow for cross-neighborhood interactions (Warman, 2007). If labour markets stretch beyond the borders of SSDs, then concentration measured at the SD level may capture the economic benefits of English proficiency more accurately. Finally, because it is arguably easier to move within cities than across metropolitan areas, using a larger aggregation level may reduce the selection bias.

A priori, it is unclear what level of aggregation better captures the reach of social interactions likely to affect language acquisition. I treat this issue as an empirical question that I investigate by contrasting results at the SD and SSD levels. Bertrand, Luttmer and Mullainathan (2000) and Cutler, Glaeser and Vigdor (2008) also simultaneously consider two levels of geography. Other studies of enclave effects cover a broad range of geographical units: wards in the UK (Dustmann and Fabbri, 2003), municipalities in Sweden (Edin, Fredriksson and Åslund, 2003) and in Denmark (Damm, 2009), CMAs in Canada (Warman, 2007), and regions (*Anpassungsschichten*) in West Germany (Danzer and Yaman, 2016).

2.3 Descriptive Statistics

To illustrate how people who settle in enclaves might differ from those who do not, I split the sample between immigrants above (enclave) and below (non-enclave) their language-specific median linguistic concentration measured at the SD level. Descriptive statistics for observable individual characteristics are reported in Table 1. The enclave status indicators are based on the respondents' location at the time of the third interview.[14]

In terms of demographics (panel A), immigrants living in high-concentration cities (column (2)) are significantly older, less educated, live in larger households but are less likely to be married, and are disproportionately female. In panel B, I present differences on a set of variables that are plausibly associated with one's ability and motivation for learning English. The patterns are suggestive of negative selection into enclaves. For instance, immigrants living in enclaves were less likely to be working prior to immigrating to Australia and to have chosen to move to Australia primarily for reasons related to employment opportunities. The fraction of immigrants expecting to receive some help finding work is higher among non-enclave residents. There is, however, no difference between immigrants living in high- and low-concentration areas in terms

[14]Sorting is most likely a slow process. Firstly, upon arrival, immigrants may not possess sufficient information on the entire set of possible locations, and so their initial location may be transitory. Secondly, if there are frictions in the housing market, it may take time for someone to find an affordable unit in their preferred area. Wave 3 locations are therefore used in order to fully uncover sorting patterns.

of whether they expect to receive help learning English. There is some suggestive evidence that enclave residents were more likely to immigrate for family reunion reasons, given their higher propensity to have visited Australia before and the likelihood they choose their State of residence because they have more family there.[15] All these wave one variables are retrospective and are considered as pre-determined characteristics in the empirical analyses.

Figure 1 shows the average English fluency rates in each of the three waves for enclave and non-enclave immigrants separately. There is clear evidence of sorting on the basis of pre-existing language skills: upon arrival in Australia (wave one), immigrants located in enclaves are 6 percentage point less likely to be fluent in English, and 7 percentage points less likely to have done the interview in English. The longitudinal nature of the survey allows to directly measure the amount of learning while living in Australia. Approximately 4 years after immigration, the initial gaps in language skills widen to 11 (15) percentage points for English fluency (interview in English), suggesting that enclaves do impede language acquisition. Yet, Figure 1 demonstrates that cross-sectional differences in language skills between immigrants living in high- and low-concentration area don't provide a causal estimate of enclave effects. About half of the unadjusted gap observed in wave three already existed upon arrival.

3 Empirical Framework

As a starting point, consider the cross-sectional econometric specification used in most previous studies (Danzer and Yaman, 2016; Warman, 2007; Chiswick and Miller, 2005):

$$L_{ijk} \quad = \quad \tilde{\beta}\,(\ln CA_{jk}) + \tilde{\gamma}' X_{ijk} + \tilde{\delta}_j + \tilde{\delta}_k + \tilde{\varepsilon}_{ijk} \tag{1}$$

where the dependent variable L_{ijk} is a measure of language skills and i indexes individuals, j indexes areas, k indexes language-groups. The coefficient of interest, $\tilde{\beta}$, represents the relationship between linguistic concentration and English language skills. The model includes pre-determined individual characteristics X_{ijk} as well as language-group ($\tilde{\delta}_k$) and location fixed effects ($\tilde{\delta}_j$).

The effect of enclaves is identified by comparing immigrants who belong to the same language-group but live in different cities, and by comparing immigrants who live in the same location but belong to different language-groups. Any difference across locations in terms of local labor market conditions and opportunities for learning English (e.g. differences in supply of English courses) that do not vary across language-groups

[15] For brievity, Table 1 only shows the most common reasons for moving to Australia/choosing State living in. In regression analyses, I use sets of dummies for all possible values as control variables.

are absorbed by the area fixed effects. Similarly, estimates of enclave effects are unaffected by unobserved heterogeneity between language-groups that is constant across locations. For example, economic incentives to learn English might be weaker for language-groups who are subject to more discrimination in the labour market, and groups may differ in terms of social norms regarding social integration, as well as in terms of values and attitudes towards labour market participation (Bertrand, Luttmer and Mullainathan, 2000). Language-groups fixed effects also account for the relationship between cost of language skill acquisition (degree of difficulty) and linguistic distance from English (Isphording and Otten, 2014).

One way to interpret the coefficients in equation (1) is through the lens of a language production function. The workhorse model of host-country language fluency among immigrants conceptualizes the human capital accumulation process as a function of three broad class of factors: economic incentives, degree of exposure to the language, and efficiency in language acquisition (Chiswick and Miller, 1995, 1996). This approach formalizes the idea that language acquisition reflects a response to the economic benefits to fluency (incentives) as well as to the costs, which depend on one's ability to learn new languages and the number opportunities for using the language (Van Tubergen and Kalmijn, 2009). In this sense, enclaves may affect language acquisition either by reducing the incentives to invest in language skills, or by making learning more difficult.[16]

In cross-sectional analyses, one only observes immigrants' degree of proficiency in the host-country language – the *stock* of English language skills – at one point in time, making it impossible to distinguish preexisting language skills at the time of migration from those learned subsequently. Unless immigrants are randomly allocated to locations, estimates from equation (1) do not identify the effect of enclaves on learning, a *flow* measure.

Language skills constitute a form of human capital which can be accumulated. It is therefore useful to decompose the stock of L_{ijkt} at time t into two components: the stock at time $t-1$ and the amount of learning experienced between the two periods ($Learn_{ijkt}$):

$$L_{ijkt} = (1-\zeta)L_{ijk,t-1} + Learn_{ijkt}$$

where $\zeta \in (0,1)$ is a depreciation parameter, allowing for one's proficiency in English to decreases if the skill is never or rarely used. Provided that $L_{ijk,t-1}$ is observed, an estimate of the relationship between enclave size and the acquisition of new language skills can be obtained by estimating a version of the cross-sectional model in which past skills $L_{ijk,t-1}$ are included as a covariate. More precisely, with $t=0$ corresponding to

[16] The Online Appendix lays out a simple conceptual model that clarify these distinct interpretations.

the first wave of LSIA, and $t = 1$ to the third wave:

$$
\begin{aligned}
Learn_{ijk1} &= \beta \left(\ln CA_{jk}\right) + \gamma' X_{ijk} + \delta_j + \delta_k + \varepsilon_{ijk} \\
L_{ijk1} &= \beta \left(\ln CA_{jk}\right) + (1-\zeta) L_{ijk0} + \gamma' X_{ijk} + \delta_j + \delta_k + \varepsilon_{ijk}.
\end{aligned} \tag{2}
$$

Equation (2) is effectively a flow specification, and the coefficient on β now only reflects the effect of enclaves on learning.[17] By including lagged proficiency as a covariate, this empirical strategy shares many similarities with the value-added literature in education, in which the endogeneity of some educational inputs (e.g. assignment to a better teacher) is accounted for by controlling for lagged values of the outcome of interest, generally test scores (Chetty, Friedman and Rockoff, 2014).

Including past stock of language skills as a control variable accounts for selection into enclaves on the basis of pre-immigration proficiency and unobserved characteristics correlated with it, notably prior exposure to English. Likewise, within language-group differences in ambition and ability to learn new languages are likely to be echoed by past language skills. For instance, for most migrants, the decision to move to Australia was made at least some time before they actually migrated.[18] In this context, the most ambitious individuals plausibly invested relatively more resources into developing their ability to speak English before migrating, anticipating this skill would be valuable in the host-country.

Yet, immigrants with the same linguistic background and the same degree of proficiency upon arrival may still differ in unobserved ways associated with both the propensity to locate in enclaves and the ability to learn English. I use two complementary strategies to address this concern. First, I exploit the scope of the LSIA1 questionnaire to examine how the estimated coefficient varies with the inclusion of a rich set of individual characteristics after conditioning on past proficiency. I then construct bounds on the causal effect of enclaves in section 4.2. Second, in section 4.3, I use the fact that immigrants sponsored by a family member (e.g. a spouse, a child, a close relative) tend to move close to their sponsor as the basis for an instrumental variable derived from the sponsor's location.

[17] This approach is equivalent to using a "gain" measure ($\Delta L_{ijk1} = L_{ijk1} - L_{ijk0}$) on the left hand-side. Subtracting L_{ijk0} from both sides, $\Delta L_{ijk1} = \alpha + \beta \left(\ln CA_{jk}\right) - \zeta L_{ijk0} + \gamma' X_{ijk} + \delta_j + \delta_k + \varepsilon_{ijk}$.

[18] The LSIA sample includes only offshore visaed immigrants. Moreover, under the "cap and queue" system, immigrants applying for the Independent or Skilled-Australian Sponsored visa categories might have been put on a waiting list if the yearly limit of immigrants allowed under these visa categories had been reached at the time of their application (Chiswick and Miller, 2004).

4 Results

4.1 The Effect of Linguistic Concentration on Language Acquisition

Baseline estimates of the effect of linguistic concentration on language skills are shown in Table 2 for two outcome measures: self-reported binary fluency and an indicator of whether the interview was conducted in English or not.[19] I use linear probability models, in conformity with previous work (Danzer and Yaman, 2016; Dustmann and Fabbri, 2003; Chiswick and Miller, 1995). In all cases, the standard errors are clustered within language-group by area of residence cells, that is at the level of aggregation at which CA_{jk} varies.[20] Panel A reports results for concentration measured at the broader SD level, and panel B presents the corresponding estimates for enclave size measured at the SSD level.

In columns (1) and (4), the set of individual characteristics is restricted to standard demographic variables (age, gender, marital status, household size, presence of children in household) and lagged proficiency is omitted. In line with previous findings, the relationship between linguistic concentration and language skills is negative and significant at the city-level (Statistical Division, panel A). A large set of observable characteristics plausibly related to ability and motivation to learn English (listed in Table 1) is further added in columns (2) and (4). The change in R^2 confirms that these variables have sizable predictive power for language skills.[21] The coefficient on enclave size decreases slightly, from -0.096 to -0.084 for self-reported speaking ability, and from -0.102 to -0.092 for language of interview. For comparison, Chiswick and Miller (1996) find that a one percentage point increase in own language-group's share of the population at the metro level is associated with a 5 percentage points reduction in English proficiency in Australia. The point estimate reported in column (2) implies that a similar increase of one percentage point in concentration translates into a 5.6 percentage points decline in fluency.[22]

Lagged language skills is added as a regressor in columns (3) and (6), and the corresponding coefficient on enclave size represents the effect on learning. The coefficient of interest is strongly statistically significant, but shrinks by a third relative to models that do not control for pre-immigration proficiency. In other words, linguistic concentration does appear to have a negative impact on language skill acquisition, but a considerable fraction of the cross-sectional estimate is attributable to sorting on the basis of pre-immigration proficiency. To better grasp the economic significance of these results, I first calculate the within-language

[19] Appendix Table A.1 presents results for three alternative measures of spoken English: an indicator for both conducting the interview in English and self-reporting being fluent, the principal component of these two variables, and an indicator for whether the respondent indicated (in wave three) that their ability to speak English had improved at least moderately since the last interview.

[20] Twoway clustering at the area and language group levels produce almost identical results (unreported).

[21] The adjusted R^2 increases by 38%, from 0.33 to 0.455, for self-reported proficiency, for example.

[22] The average SD-level language-group share in my sample is 0.015. A one percentage point increase therefore corresponds to a 66% increase from the mean, and a $(0.0844 * 0.66 =)$ 5.6 percentage point decrease in fluency.

group standard deviation in concentration relative to the sample mean, which is 45% of the mean at the SD-level. The impact of a one standard deviation increase in enclave size is a (0.058∗0.45 =) 2.6 percentage points reduction of in the probability of being fluent in English. To put this magnitude in perspective, consider the total change in proficiency between the first and third waves – a 17.6 percentage points (43%) increase, from 41% to 59%. The effect of a (within-group) standard deviation increase in linguistic concentration on language acquisition is equivalent to about (2.6/17.6 =) 15% of the sample's total improvement between the two waves.[23]

In Panel B, the effect of SSD-level concentration (CA_{jk}^{SSD}) is small and generally statistically insignificant. This might be surprising given that these estimates rely on variation across SSDs both within and between SDs.[24] In Online Appendix B, I show that SSD-level estimates can be interpreted as a weighted average of (a) the partial effect of SD-level concentration (CA_{jk}^{SD}), and (b) the partial effect of within-SD *relative* SSD-level concentration. Perhaps surprisingly, I find no evidence of an impact of CA_{jk}^{SSD} on language acquisition working through the second channel. A decomposition of the estimates reported in panel B between these two channels suggests that any effect of CA_{jk}^{SSD} works entirely through its relationship with SD-level concentration. The resulting coefficients are particularly small because considerably more weight is put on the (null) effect of relative SSD-level concentration conditional on SD-level enclave size. In other words, conditional on residing in a city where many people share one's mother tongue, there is no incremental negative effect of residing in a relatively high-concentration neighborhood. Similarly, Cutler, Glaeser and Vigdor (2008) find that MSA-level group share is consistently associated with lower English ability in the U.S., but that within-MSA segregation, if anything, is positively associated with English ability conditional on MSA-level group share. In light of these results, the remaining of the paper will mainly focus on SD-level enclave size.

To validate that the effect of enclaves on learning is not driven by groups of outliers or by functional form assumptions, I perform a number of robustness and specification checks. First, I verify that the negative effect of linguistic concentration does not hinge on age restrictions (Table A.2, rows (b)-(d)) and confirms that the effect is not driven by a few large language groups (rows (e)-(h)). Excluding Chinese (21% of the sample) and Arabic (10%) speakers notably leads to coefficients larger in magnitude.[25] Excluding the two

[23]My estimates are strikingly close the those obtained by Danzer and Yaman (2016), who find that a one standard deviation increase in ethnic concentration at the regional level reduces the probability of being fluent in German by 3.8 percentage points. A full one standard deviation increase in concentration in my sample decreases proficiency by ($0.058 * 0.66 =$) 3.8 percentage points. My results are, however, markedly lower than most cross-sectional estimates. For example, Lazear (1999) finds that a one standard deviation increase in ethnic concentration lowers English proficiency by about 23 percentage points, that is by about a third of the overall fluency rate prevailing in the U.S. in 1990.

[24]The magnitudes of the coefficients at the SD and SSD level are not directly comparable because people do not locate randomly within cities. For instance, a 1% increase in linguistic concentration at the SSD level will conceivably translate into a less-than-1% increase in SD-level concentration.

[25]Note that it may be the case that linguistic concentration is an inaccurate measure of social networks for these broad

metropolitan areas where the vast majority of immigrants settle also leaves the main conclusion unchanged (rows (i) and (j)). Second, I take into account that individuals might have moved non-randomly across cities between the two interviews. I would over-estimate the negative effect of enclave size on language skills if, for example, individuals who were able to improve their ability to speak English markedly while living in an enclave then moved to a low-concentration area just before the third interview. Yet, this form of sorting is not cause for concern in the case of analyses conducted at the SD level, since less than 5% of respondents in my sample moved across SDs (37 percent moved between SSDs). The last four rows of Table A.2 display the estimated effect of linguistic concentration on language acquisition for different subgroups of "stayers". Finally, Tables A.3 and A.4 respectively show that the results remain qualitatively unchanged if one uses probit models or measures concentration in levels rather than in log.

4.2 Sensitivity to controls

In this section, I examine the robustness of the estimated enclave effects to the inclusion of a comprehensive set of individual characteristics. Intuitively, if controlling for lagged language skills accounts for most of the relevant unobserved heterogeneity, then the coefficient β should be stable across specifications that do and do not account for observable characteristics once we have conditioned on past proficiency.

In Table 3, I re-estimate flow equation (2) at the SD-level and gradually add more covariates. In columns (1) and (4), the set of individual characteristics is empty, while columns (2) and (5) include baseline demographics, and numerous proxies for ability and motivation to learn English are further added in columns (3) and (6). These variables notably include a full set of education dummies, indicators for the main reason for choosing the State of residence, visa category dummies, the number of languages spoken well, and whether the PA expected to receive help learning English on arrival. For both self-reported proficiency and language of interview, the coefficients are very similar across columns and difference are not statistically significant.

To provide a more complete picture, I then implement a test of how strong selection on unobservables would have to be, relative to selection on observables, to explain away the entire estimated effect. Building upon the work of Altonji, Elder and Taber (2005), Oster (2017) shows that under proportional selection on observables and unobservables, one can calculate bounds on the coefficient of interest using the magnitude of the movements of the coefficient and of the R^2 as controls are added. Bias-adjusted effects are approximately given by

groups. The definition of Chinese language used in the Australian census encompasses more than six dialects (Cantonese, Hokkien, Kan/Hakka, Mandarin, Min, Teo-Chiew and Wu). The same is true of Arabic, which is a catchall category for all varieties of Arabic languages.

$$\beta^* \approx \beta^L - \delta\left(\beta^S - \beta^L\right)\left[\frac{R_{max} - R^L}{R^L - R^S}\right]$$

where β^L and β^S are the estimated effects in the long (with covariates) and short (no covariate) models, respectively. R^L and R^S are the associated R^2s. The parameter R_{max} is the R^2 we would observe if all relevant unobservables were also included in the full regression, and δ is the ratio of selection on unobservables over selection on observables. A fair approximation of R_{max} can be obtained by running a regression that includes fixed effects for all possible combinations of SSDs and language-groups. Doing so drives up the R^2 to 0.70 for self-assessed English language skills. For completeness, I also compute more conservative bounds using $R_{max} = 1$.

Identified sets $[\beta^L, \beta^*]$ for $\delta = 1$ are shown in Table 3.[26] In all cases, the bounds are relatively tight, and the identified sets never include zero. In addition, I report values of δ that would be necessary to bring β down to zero. The results indicate that if the variation in language skills could be fully explained by observables and unobservables ($R_{max} = 1$), selection on unobservables would have to be between 2 and 5 times larger than selection on observables for the estimated effect to vanish completely (columns (3) and (6)). If the relevant unobservables were only to raise the R^2 to 0.7, then selection on unobservables would have to be 8 to 19 times larger than selection on observables to completely explain away enclave effects.

4.3 Sponsored immigrants

The second approach I use to verify that my estimates of enclave effects are not driven by sorting is to focus on a subgroup of immigrants for whom the location decision is transparent: PAs sponsored by a relative. For these individuals, where they locate depends largely on where their sponsor happens to residence. For instance, when asked whether their sponsor's location had an important influence on where they chose to live, 94% of sponsored said that it did. By the time of the third interview, the vast majority (90%) of sponsored PAs indicated still living in the same city or town as their sponsor.

Table 4 report estimates of enclave effects on language skills for sponsored PAs.[27] Column (1) reproduces the baseline specification for the relevant subsample, and I gradually incorporate sponsor-related variables in columns (2) through (5). First, I control for sponsor observable characteristics that are plausibly correlated

[26] Oster (2017) suggests that equal selection ($\delta = 1$) is an appropriate upper bound.
[27] For completenesse, Table A.5 present results for the alternative measures of language skills also used in Table A.1.

with the sponsor's language abilities and degree of social integration (column (2)).[28] Including these covariates does not materially affect the magnitude of the coefficient of interest. This result is robust to adding fixed effects for the sponsor's State of residence (column (3)), restricting the sample to PAs whose sponsor has been living in Australia for at least 5 years (column (4)), or adding a set of dummies for the sponsor's relationship with the PA (column (5)) – e.g. spouse, uncle, daughter.

In columns (6) through (10), enclave size is instrumented with linguistic concentration in the sponsor's State of residence.[29] The exclusion restriction under this approach postulates that unobserved sponsor characteristics that directly impede or facilitate the immigrant's acquisition of language skills are not systematically related to the sponsor's propensity to locate in enclaves. One concern, for instance, is that sponsors located in enclaves are systematically less likely to directly help PAs becoming fluent, in which case any relationship between enclave size in the sponsor's location and learning might operate via the sponsor's behavior rather than the enclave environment itself. Note that in theory, however, the sponsor's own English language skills could have either a positive or a negative effect on the PAs' learning, since proficient sponsors could act as either teachers or interpreters.[30] Examining the relationship between linguistic concentration in the sponsor's location and the likelihood that the PA received help from their relatives learning English, I find no evidence that either a positive or negative link exists. Linguistic concentration in the sponsor's location is also unrelated to PAs' language skills upon arrival.[31]

In most specifications, the 2SLS estimates are moderately smaller than the corresponding OLS estimates, and are considerably less precise. Given that I can only measure enclave size in the sponsor's location at a very high level of aggregation, these results must be interpreted with caution. Yet, the magnitude of the point estimates tend to be line with my baseline estimates of enclave effects, supporting the hypothesis that enclaves causally slow down language acquisition.

[28]The characteristics are the number of year lived in Australia, whether the sponsor is an Australian citizen, and whether English is the language spoken at home. Some immigrants who were sponsored by their fiancé/spouse declared that English was spoken at home at the time of the first interview, suggesting that their spouse do not share their mother tongue.

[29]Sponsor location is only recorded at a fairly aggregate level that cannot be accurately matched with Statistical Divisions. Yet, the first-stage remains strong (F-stat>75). This instrument shares some similarities with the approach taken by Edin, Fredriksson and Åslund (2003), who instrument ethnic concentration in the current area of residence with enclave size in the municipality to which refugees were assigned at the time of migration under a refugee placement program in Sweden. It also incorporates insight from Bertrand, Luttmer and Mullainathan (2000), who instrument PUMA-level concentration with concentration at a more aggregate level of geography (MSA).

[30]Chiswick and Miller (1995) make a similar point regarding the influence of children born in the host-country on their immigrant parents' proficiency.

[31]In the first interview, LSIA respondents were asked if their relatives helped them learning English, as well as if they helped them getting a job. Regressing binary variables indicating whether help was received on linguistic concentration in the sponsor's location, conditioning on sponsor characteristics and the baseline set of fixed effects (language-group and SSD), yields coefficients statistically indistinguishable from zero, with F-stats of 1.17 and 0.5 for learning English and finding work, respectively. In a similar regression with $L_{ijk,t-1}$ as the dependent variable, the F-stat on the coefficient for enclave size in the sponsor's location is 0.6.

5 Heterogeneous effects and mechanisms

To shed light on the possible mechanisms through which linguistic concentration impedes language acquisition, I first examine which groups of immigrants are most affected by enclaves. Table 5 displays the results of an analysis of the heterogeneity of effects. Linguistic concentration appears to strongly influence the language acquisition of women, but not of men.[32] If women's attachment to the labor market is weaker, then their labour market participation, and therefore indirectly their decision to invest in host-country human capital, are plausibly more responsive to environmental circumstances. Secondly, English language acquisition seems unrelated to linguistic concentration for immigrants older than 35 at the time of immigration. Presumably, the lifetime benefits of learning English are lower for older individuals who have fewer years left on the job market. These two pieces of evidence suggest that language acquisition indeed responds to economic incentives. For the most and least educated immigrants, learning appears relatively insensitive to the size of their linguistic group. Arguably, it might be the case that high (low) ability individuals find it particularly easy (difficult) to learn a new language independently of the social context.

Table 6 further unpacks the possible ways in which enclaves might affect language acquisition via economic incentives. My objective, here, is to uncover some general patterns rather than isolate causal effects of enclaves on intermediate outcomes. The results are therefore suggestive at best.

In column (1), the relationship between enclave size and the propensity to take up English courses is reported.[33] The small and statistically insignificant coefficient implies that enclave residents do not have a lower propensity to enroll in formal language courses than immigrant living outside of enclaves. The effect of enclaves on proficiency therefore probably works through social interactions, not differences in formal training. Not only does English proficiency provide strong economic value by increasing the number of people with whom one can interact, but learning itself depends on the number of opportunities to practice speaking the language. The absence of a correlation between enclave size and English course take-up is consistent with the fact that linguistic concentration has little or no effect on reading and writing skills, the learning of which plausibly depend much less on social interactions outside the classroom than speaking abilities (Table A.6). From a policy perspective, these patterns suggest that additional public subsidies towards formal English courses would not induce enclave residents to invest more in language skills, considering that such courses were already provided for free in Australia.

[32] Danzer and Yaman (2016) find that enclaves affect men and women equally in Germany, whereas Warman (2007) report larger effects of enclaves on language skills for women than men in Canada.

[33] In Australia, immigrants are provided with up to 510 hours of free English language tuition through the Adult Migrant English Program (AMEP). Notably, the program also offers correspondence or distance learning courses and Home Tutor services, in addition to formal classroom tuition (Martin, 1998).

In columns (2) through (7) of Table 6, I examine indirect pieces of evidence regarding the role of social interactions. In column (2), the dependent variable is the amount of contact between people from different countries and cultures, as perceived by the PA. Concentration is associated with a decrease in the amount of such contacts, but the relationship is not statistically significant. Columns (3) and (4) indicate that enclave size is not associated with lower income or lower employment rates, in line with Danzer and Yaman (2016); Damm (2009); Edin, Fredriksson and Åslund (2003).[34] Columns (5) and (6) summarize the type of jobs occupied by workers and their job search strategies, conditional on being employed at the time of the third wave. Again, the relationships are not statistically significant, but the point estimates suggest that more enclave residents hold jobs at which they speak languages other than English and that they found their jobs through friends and famil.[35] Finally, in column (7), I directly test the hypothesis that the return to proficiency is lower in enclaves. To do so, for each SD by language-group cell, I calculate the difference in employment rates between fluent and non-fluent individuals. I then regress this measure of the return to proficiency on enclave size, and find that they are indeed negatively related: a 10% increase in linguistic concentration is associated with a 1.5 percentage points reduction in the English fluency premium (expressed in terms of employment probabilities).

While the estimates presented in Table 6 are all very imprecise, a general picture emerges. The penalty for reduced English proficiency, in terms of probability of finding work, is smaller in enclaves, plausibly because jobs with weaker language requirements are more available. As a result, enclave residents are just as likely to find work, despite having lower language skills. One may then speculate that the work environment itself provides fewer opportunities to practice English in enclaves, thereby reducing language skill acquisition.

6 Concluding Remarks

This paper estimates the relationship between linguistic concentration and language skill acquisition in an English-speaking country. While a negative effect of enclaves on learning is confirmed, I demonstrate that non-random sorting of immigrants on the basis of pre-immigration proficiency is prevalent, and considerably inflates cross-sectional estimates. The stability of the estimated relationship across several empirical strategies attests to the robustness of this result. Overall, I find that a (within language-group) standard deviation increase in concentration translates into a 2.6 percentage points reduction in learning rates. Considering the large economic returns to proficiency, this negative effect of enclaves possibly contributes to slowing down

[34] Any reduction in income due to weaker language skills in enclaves relative to non-enclaves is possibly counterbalanced by the benefits of language group-based job information networks.

[35] The results are fairly similar at the SSD level, but the negative coefficients for "contact between cultures" and "English only at work" are statistically significant at that level of aggregation.

the economic assimilation of immigrants (Borjas, 2015).

The evidence presented in this paper suggests that policies aimed at reducing the cost of formal English courses are unlikely to attenuate the negative effect of linguistic concentration.[36] The results are perhaps more consistent with social interactions being the main mechanism through which enclaves affect the aquisition of host country-specific human capital. If English skills are practiced and improved at work, then some immigrants might be caught in a circular relationship between lack of English job opportunities and low English proficiency. Speculatively, poor initial language skills might prevent individuals from finding certain types of jobs, and the resulting scarcity of interactions in English with colleagues further hinders language skill acquisition. The question of which type of intervention will be more successful at reducing the effects of enclaves on language skill acquisition – e.g. anti-discrimination and labor market integration policies, or social integration activities – warrants further research.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–184.

Angrist, Joshua D. 2014. "The perils of peer effects." *Labour Economics.*

Australian Bureau of Statistics. 1996. *Statistical Geography: Volume 1. Australian Standard Geographical Classification (ASGC), 1996 Edition.* Catalogue no.1216.0.

Bauer, Thomas, Gil Epstein, and Ira Gang. 2005. "Enclaves, language, and the location choice of migrants." *Journal of Population Economics*, 18(4): 649–662.

Bayer, Patrick, Robert McMillan, and Kim S. Rueben. 2004. "What drives racial segregation? New evidence using Census microdata." *Journal of Urban Economics*, 56(3): 514–535.

Beaman, Lori A. 2012. "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S." *Review of Economic Studies*, 79(1): 128–161.

Berman, Eli, Kevin Lang, and Erez Siniver. 2003. "Language-skill complementarity: returns to immigrant language acquisition." *Labour Economics*, 10(3): 265–290.

Bertrand, Marianne, Erzo F. P. Luttmer, and Sendhil Mullainathan. 2000. "Network Effects And Welfare Cultures." *The Quarterly Journal of Economics*, 115(3): 1019–1055.

Bleakley, Hoyt, and Aimee Chin. 2004. "Language Skills and Earnings: Evidence from Childhood Immigrants." *The Review of Economics and Statistics*, 86(2): 481–496.

Borjas, George J. 1995. "Ethnicity, Neighborhoods, and Human-Capital Externalities." *American Economic Review*, 85(3): 365–90.

Borjas, George J. 2015. "The slowdown in the economic assimilation of immigrants: Aging and cohort effects revisited again." *Journal of Human Capital*, 9(4): 483–517.

[36] Such policies may very well help increase proficiency rates in general, however. The analyses presented here only indicate that the effect of enclaves on proficiency is not mediated by English course take-up.

Calvó-Armengol, Antoni, and Matthew O. Jackson. 2004. "The Effects of Social Networks on Employment and Inequality." *American Economic Review*, 94(3): 426–454.

Calvó-Armengol, Antoni, and Yves Zenou. 2005. "Job matching, social network and word-of-mouth communication." *Journal of Urban Economics*, 57(3): 500–522.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *American Economic Review*, 104(9): 2593–2632.

Chiswick, Barry R., and Paul W. Miller. 1995. "The Endogeneity between Language and Earnings: International Analyses." *Journal of Labor Economics*, 13(2): 246–88.

Chiswick, Barry R., and Paul W. Miller. 1996. "Ethnic Networks and Language Proficiency among Immigrants." *Journal of Population Economics*, 9(1): 19–35.

Chiswick, Barry R., and Paul W. Miller. 2004. "Language Skills and Immigrant Adjustment: What Immigration Policy Can Do!" Institute for the Study of Labor (IZA) IZA Discussion Papers 1419.

Chiswick, Barry R., and Paul W. Miller. 2005. "Do enclaves matter in immigrant adjustment?" *City & Community*, 4(1): 5–35.

Chiswick, Barry R., and Paul W. Miller. 2014. "International Migration and the Economics of Language." Institute for the Study of Labor (IZA) IZA Discussion Papers 7880.

Chiswick, Barry R., Yew Liang Lee, and Paul W. Miller. 2004. "Immigrants' language skills: The Australian experience in a longitudinal survey." *International Migration Review*, 38(2): 611–654.

Chiswick, Barry R., Yew Liang Lee, and Paul W. Miller. 2006. "Immigrants' Language Skills and Visa Category." *International Migration Review*, 419–450.

Church, Jeffrey, and Ian King. 1993. "Bilingualism and Network Externalities." *Canadian Journal of Economics*, 26(2): 337–45.

Cobb-Clark, Deborah A. 2004. "Selection Policy and the Labour Market Outcomes of New Immigrants." Institute for the Study of Labor (IZA) IZA Discussion Papers 1380.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 2008. "When are ghettos bad? Lessons from immigrant segregation in the United States." *Journal of Urban Economics*, 63(3): 759–774.

Damm, Anna Piil. 2009. "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence." *Journal of Labor Economics*, 27(2): 281–314.

Danzer, Alexander M., and Firat Yaman. 2013. "Do Ethnic Enclaves Impede Immigrants' Integration? Evidence from a Quasi-experimental Social-interaction Approach." *Review of International Economics*, 21(2): 311–325.

Danzer, Alexander M, and Firat Yaman. 2016. "Ethnic concentration and language fluency of immigrants: Evidence from the guest-worker placement in Germany." *Journal of Economic Behavior & Organization*, 131: 151–165.

Dustmann, Christian, and Arthur van Soest. 2001. "Language Fluency And Earnings: Estimation With Misclassified Language Indicators." *The Review of Economics and Statistics*, 83(4): 663–674.

Dustmann, Christian, and Arthur van Soest. 2002. "Language and the earnings of immigrants." *Industrial and Labor Relations Review*, 55(3): 473–492.

Dustmann, Christian, and Francesca Fabbri. 2003. "Language proficiency and labour market performance of immigrants in the UK." *Economic Journal*, 113(489): 695–717.

Edin, Per-Anders, Peter Fredriksson, and Olof Åslund. 2003. "Ethnic Enclaves And The Economic Success Of Immigrants - Evidence From A Natural Experiment." *The Quarterly Journal of Economics*, 118(1): 329–357.

Evans, William N, Wallace E Oates, and Robert M Schwab. 1992. "Measuring peer group effects: A study of teenage behavior." *Journal of Political Economy*, 100(5): 966–991.

Flores, Glenn. 2006. "Language barriers to health care in the United States." *New England Journal of Medicine*, 355(3): 229–231.

Goel, Deepti, and Kevin Lang. 2009. "Social Ties and the Job Search of Recent Immigrants." National Bureau of Economic Research, Inc NBER Working Papers 15186.

Ioannides, Yannis M., and Linda Datcher Loury. 2004. "Job Information Networks, Neighborhood Effects, and Inequality." *Journal of Economic Literature*, 42(4): 1056–1093.

Isphording, Ingo E, and Sebastian Otten. 2014. "Linguistic barriers in the destination language acquisition of immigrants." *Journal of Economic Behavior & Organization*, 105: 30–50.

Konya, Istvan. 2007. "Optimal Immigration and Cultural Assimilation." *Journal of Labor Economics*, 25: 367–391.

Lazear, Edward P. 1999. "Culture and Language." *Journal of Political Economy*, 107(S6): S95–S126.

Lewis, Ethan. 2013. "Immigrant-Native Substitutability and the Role of Language." In *Immigration, Poverty, and Socioeconomic Inequality.* , ed. David Card and Steven Raphael. Russell Sage Foundation.

Martin, Shirley. 1998. "New life, new language: The history of the Adult Migrant English Program."

Moretti, Enrico. 2004. "Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data." *Journal of econometrics*, 121(1): 175–212.

Munshi, Kaivan. 2003. "Networks In The Modern Economy: Mexican Migrants In The U.S. Labor Market." *The Quarterly Journal of Economics*, 118(2): 549–599.

OECD. 2013. *International Migration Outlook 2013.* OECD Publishing.

Oreopoulos, Philip. 2011. "Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes." *American Economic Journal: Economic Policy*, 3(4): 148–71.

Oster, Emily. 2017. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, 0(0): 1–18.

Patacchini, Eleonora, and Yves Zenou. 2012. "Ethnic networks and employment outcomes." *Regional Science and Urban Economics*, 42(6): 938–949.

Stevens, Gillian. 1992. "The social and demographic context of language use in the United States." *American Sociological Review*, 171–185.

Van Tubergen, Frank, and Matthijs Kalmijn. 2009. "Language proficiency and usage among immigrants in the Netherlands: Incentives or opportunities?" *European Sociological Review*, 25(2): 169–182.

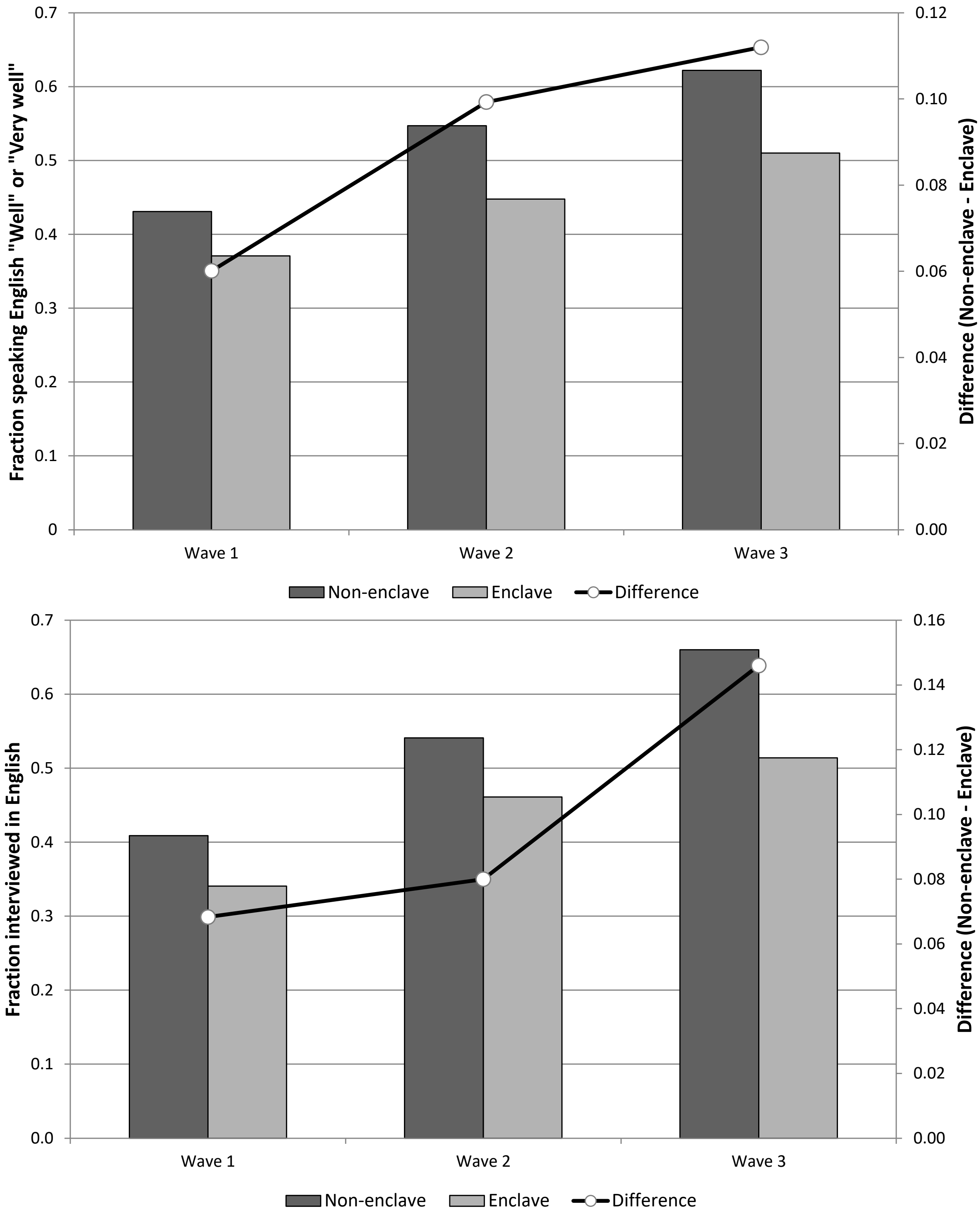
Vervoort, Miranda, Jaco Dagevos, and Henk Flap. 2012. "Ethnic concentration in the neighbourhood and majority and minority language: A study of first and second-generation immigrants." *Social science research*, 41(3): 555–569.

Wahba, Jackline, and Yves Zenou. 2005. "Density, social networks and job search methods: Theory and application to Egypt." *Journal of Development Economics*, 78(2): 443–473.

Warman, Casey. 2007. "Ethnic enclaves and immigrant earnings growth." *Canadian Journal of Economics*, 40(2): 401–422.

Tables and Figures

Figure 1: Language Acquisition



Notes: Data from the Longitudinal Survey of Immigrants to Australia. Enclaves defined as areas with linguistic concentration above the language-specific median, at the SD level. Statistics are based on 2,053 observations and are calculated using the provided survey weights.

Table 1: Descriptive statistics

Individual characteristics	Full sample	Enclave	Non-enclave	Difference (2) - (3)	Difference (s.d.)
	(1)	(2)	(3)	(4)	(5)
A. Demographics					
Women	0.508	0.541	0.493	0.0475**	(0.0242)
Age	32.56	33.963	31.97	1.993***	(0.487)
Married	0.770	0.725	0.789	-0.0641***	(0.0203)
Children present	0.345	0.385	0.329	0.0563**	(0.0230)
Household size	4.012	4.192	3.937	0.255**	(0.0998)
B. Ability and motivation proxies					
Education:					
Bachelor/Professional degree or more	0.273	0.231	0.290	-0.0590***	(0.0215)
Trade/Highschool	0.454	0.417	0.470	-0.0531**	(0.0241)
Less than High school	0.273	0.352	0.240	0.112***	(0.0214)
Visa:					
Preferential Family	0.586	0.553	0.601	-0.0480**	(0.0238)
Concessional Family	0.079	0.088	0.076	0.0120	(0.0131)
Business Skills/ENS	0.025	0.031	0.023	0.0085	(0.0076)
Independent	0.122	0.127	0.121	0.0058	(0.0159)
Humanitarian	0.187	0.202	0.180	0.0218	(0.0189)
Reason choosed State living in:					
Spouse/partner lived here	0.466	0.408	0.490	-0.0825***	(0.0241)
Employer is located here	0.018	0.019	0.017	0.0014	(0.0064)
Job opportunities	0.069	0.067	0.070	-0.0026	(0.0122)
More family in this state	0.293	0.365	0.264	0.101***	(0.0219)
Friends living here	0.073	0.060	0.078	-0.0188	(0.0126)
Reason moved to Australia:					
Better employment opportunities	0.198	0.159	0.215	-0.0563***	(0.0193)
To join family/relatives	0.381	0.407	0.371	0.0362	(0.0235)
To get married	0.167	0.156	0.171	-0.0147	(0.0180)
Better future for family	0.147	0.173	0.137	0.0357**	(0.0171)
Was working in Prior Country	0.697	0.661	0.713	-0.0521**	(0.0222)
Visited Australia before	0.273	0.306	0.259	0.0470**	(0.0215)
Requested job-related information	0.461	0.448	0.466	-0.0176	(0.0241)
Expected to receive help finding work	0.515	0.463	0.536	-0.0730***	(0.0241)
Expected to receive help learning English	0.525	0.534	0.521	0.0129	(0.0242)
Number of languages spoken well	1.473	1.507	1.459	0.0481	(0.0356)

Notes: Enclaves defined as areas with linguistic concentration above the language-specific median, at the SD level. Statistics are based on 2,053 observations and are calculated using the provided survey weights. All individual characteristics are derived from wave one survey questions. For variables pertaining to the main reason for choosing to move to Australia/for choosing State living in, only the most popular answers are shown. "Requested job-related information" and the "Expected to receive help..." are retrospective question, referring to the period just before moving to Australia.

Table 2: Main results

Dependent variable:	Spoken English			Interview in English		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Statistical Division level (SD)						
ln(Enclave size)	-0.0958***	-0.0844***	-0.0584***	-0.102***	-0.0918***	-0.0613***
	(0.0237)	(0.0226)	(0.0206)	(0.0212)	(0.0213)	(0.0205)
R^2	0.363	0.491	0.568	0.386	0.492	0.553
B. Statistical Subdivision level (SSD)						
ln(Enclave size)	-0.0234	-0.0345**	-0.0212	-0.0236	-0.0326**	-0.0218
	(0.0162)	(0.0147)	(0.0131)	(0.0153)	(0.0149)	(0.0137)
R^2	0.359	0.490	0.567	0.382	0.489	0.552
N	2053	2053	2053	2053	2053	2053
Language group fixed effects	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X
Motivation/Ability proxies		X	X		X	X
Wave 1 dependent variable			X			X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level in panel A, and the SSD by language-group level in panel B, are reported in parentheses. Data are weighted using the provided sample weights. All specifications are linear probability models. Demographic characteristics are age, gender, household size, marital status, and presence of children. Motivation/ability proxies are education dummies, main reason for moving to Australia, main reason for choosing State of residence, number of languages spoken well, having visited Australia prior to immigration, visa category, work status in former country, whether the PA requested job-related information before moving, whether the PA expected to receive help finding work, expected to receive help learning English. In columns (1) and (2), the dependent variable is equal to one if English is either spoken "Well" or "Very well".
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table 3: Main results - Sensitivity to observable characteristics

Dependent variable:	Spoken English			Interview in English		
Specification:	Short	Long 1	Long 2	Short	Long 1	Long 2
	(1)	(2)	(3)	(4)	(5)	(6)
ln(Enclave size)	-0.0610***	-0.0569***	-0.0584***	-0.0658***	-0.0597***	-0.0613***
	(0.0228)	(0.0196)	(0.0206)	(0.0222)	(0.0202)	(0.0205)
R^2	0.457	0.555	0.568	0.461	0.539	0.553
Identified set if						
R_{max} = 0.7		[-0.0569,-0.0508]	[-0.0584,-0.0553]		[-0.0597,-0.0471]	[-0.0613,-0.0541]
R_{max} = 1		[-0.0569,-0.0383]	[-0.0584,-0.0483]		[-0.0597,-0.0236]	[-0.0613,-0.0394]
Selection necessary for β=0 if						
R_{max} = 0.7		9.4	18.9		4.7	8.5
R_{max} = 1		3.1	5.8		1.7	2.8
Language group fixed effects	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X
Demographic characteristics		X	X		X	X
Motivation/Ability proxies			X			X
Wave 1 dependent variable	X	X	X	X	X	X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. The method used to calculate the identified set as well as the amount of selection on unobservable necessary for β to shrink to zero is from Oster (2017) and is discussed in the main text.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table 4: Sponsored Immigrants

Estimation:	OLS					2SLS				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Dependent variable: Spoken English										
ln(Enclave size)	-0.0774***	-0.0705***	-0.0709***	-0.0721**	-0.0721***	-0.0498	-0.0457	-0.0559*	-0.0598	-0.0546
	(0.0240)	(0.0234)	(0.0235)	(0.0281)	(0.0253)	(0.0395)	(0.0383)	(0.0313)	(0.0389)	(0.0335)
B. Dependent variable: Interview in English										
ln(Enclave size)	-0.0555**	-0.0478**	-0.0471**	-0.0413*	-0.0581**	-0.0480	-0.0456	-0.0238	-0.0310	-0.0359
	(0.0240)	(0.0228)	(0.0233)	(0.0245)	(0.0241)	(0.0595)	(0.0579)	(0.0291)	(0.0311)	(0.0327)
N	1422	1422	1422	1127	1231	1422	1422	1422	1127	1231
Language group fixed effects	X	X	X	X	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X	X	X	X	X
Motivation/Ability proxies	X	X	X	X	X	X	X	X	X	X
Wave 1 dependent variable	X	X	X	X	X	X	X	X	X	X
Sponsor's characteristics		X	X	X	X		X	X	X	X
Sponsor's location fixed effects			X	X	X			X	X	X
Sample restriction: Sponsor in Australia for ≥ 5 years				X					X	
Sponsor's relationship with PA					X					X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. The sample is restricted to PAs sponsored by a relative. The sponsor's location is at the State level. Sponsor characteristics are number of year lived in Australia, Australian citizen status, and whether English is the language spoken at home. In columns (5) and (10), a full set of dummies for categories of the sponsor's relationship with the PA are included. In columns (6) through (10), the instrument is ln(*Enclave size*) in the sponsors' State of residence. ***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table 5: Heterogeneous effects

	Dependent variable		Sample size
	Spoken English	Interview in English	
Subsample	(1)	(2)	(3)
Baseline	-0.0584***	-0.0613***	2053
	(0.0206)	(0.0205)	
Men	-0.0367	-0.0205	1137
	(0.0323)	(0.0312)	
Women	-0.0901***	-0.0945***	916
	(0.0300)	(0.0324)	
Education: Low	-0.0333	-0.0478*	729
	(0.0299)	(0.0278)	
Education: Middle	-0.0782***	-0.103***	917
	(0.0229)	(0.0251)	
Education: High	-0.0745	0.0240	407
	(0.0830)	(0.0560)	
Age >= 35	-0.0122	-0.0060	769
	(0.0398)	(0.0377)	
Age < 35	-0.0697***	-0.0744***	1284
	(0.0230)	(0.0236)	

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. High education denotes a bachelor degree or more, medium education is a professional or trade degree, or 12 years of schooling or more, and low education is less than 12 years of schooling.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table 6: Possible Mechanisms

Dependent variable:	English course take-up	Contact between cultures	ln(Income)	Working	English only language at work	Found job through friends or family	Return to proficiency
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(Enclave size)	0.0187	-0.0267	0.0260	-0.0118	-0.0355	0.0508	-0.149*
	(0.0278)	(0.0242)	(0.0854)	(0.0329)	(0.0383)	(0.0340)	(0.0847)
N	2053	2053	2031	2053	1226	1226	1782
Language group fixed effects	X	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X	X
Motivation/Ability proxies	X	X	X	X	X	X	X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. English course take-up takes value one if the PA has ever been enrolled in a course after immigration by the time of the third interview. Contact between cultures takes values one if the PA believes there at a lot of some contact between cultures in Australia. Income is measured by the mid-range value of 13 possible weekly income brackets. In column (4), the dependent variable indicates whether the PA is working in wave three. For working individuals, columns (5) and (6) use indicators for only speaking English at work, and having found the job through friends or family. The dependent variables in column (7) is a SD by language-group specic measure of the return to proficiency, in terms of employment probabilities. This outcome is missing for cells in which either everyone is fluent or no one is fluent.

***, **, * indicate significance at the 1%, 5%. and 10% levels.

A Appendix Tables and Figures

Table A.1: Baseline results - Alternative measures of spoken English

Dependent variable:	Both self-declares speaking English and interviews in English			Composite of self-declared spoken English and interview in English			How English has improved since last interview		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A. Statistical Division level (SD)									
ln(Enclave size)	-0.121***	-0.111***	-0.0746***	-0.220***	-0.196***	-0.122***	-0.0549*	-0.0478*	-0.0455
	(0.0245)	(0.0233)	(0.0227)	(0.0452)	(0.0429)	(0.0395)	(0.0286)	(0.0288)	(0.0283)
R^2	0.374	0.499	0.579	0.424	0.557	0.648	0.123	0.162	0.163
B. Statistical Subdivision level (SSD)									
ln(Enclave size)	-0.0357**	-0.0467***	-0.0350***	-0.0521*	-0.0745**	-0.0432*	-0.0402**	-0.0388**	-0.0377**
	(0.0162)	(0.0148)	(0.0132)	(0.0313)	(0.0289)	(0.0253)	(0.0174)	(0.0177)	(0.0177)
R^2	0.369	0.497	0.578	0.420	0.555	0.648	0.124	0.164	0.164
Language group fixed effects	X	X	X	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X	X	X	X
Motivation/Ability proxies		X	X		X	X		X	X
Wave 1 dependent variable			X			X			X[a]

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level in panel A, and the SSD by language-group level in panel B, are reported in parentheses. Data are weighted using the provided sample weights. Columns (1)-(3): the outcome is an indicator for both doing the interview in English and speaking English "Well" or "Very well". Columns (4)-(6): the outcome is the principal component of doing the interview in English and speaking English "Well" or "Very well". Column (7)-(9): The outcome is a binary variables that takes value one if the PA declares that her English has improved at least moderately since the last interview (the question is asked in wave three). In column (9), the lagged dependent variable is speaking English "Well" or "Very well" in wave one. Note that the outcome in columns (7)-(9) is a flow measure.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table A.2: Baseline results - Robustness to sample restrictions

	Dependent variable		
	Spoken English	Interview in English	Sample size
Subsample	(1)	(2)	(3)
(a) Baseline	-0.0584***	-0.0613***	2053
	(0.0206)	(0.0205)	
(b) Age: 25-64	-0.0408*	-0.0464**	1741
	(0.0226)	(0.0201)	
(c) Age: 18-54	-0.0604***	-0.0787***	1952
	(0.0210)	(0.0196)	
(d) Age: 25-54	-0.0418*	-0.0690***	1640
	(0.0228)	(0.0193)	
(e) Chinese speakers excluded	-0.0663***	-0.0634***	1753
	(0.0207)	(0.0206)	
(f) Arabic speakers excluded	-0.0797***	-0.0781***	1806
	(0.0204)	(0.0202)	
(g) Spanish speakers excluded	-0.0561***	-0.0639***	1877
	(0.0211)	(0.0206)	
(h) Vietnamese speakers excluded	-0.0545**	-0.0548***	1934
	(0.0241)	(0.0207)	
(i) Sydney excluded	-0.0991***	-0.0833***	1149
	(0.0260)	(0.0247)	
(j) Melbourne excluded	-0.0728***	-0.0792***	1461
	(0.0235)	(0.0209)	
(k) Stayers: Always same SD	-0.0502**	-0.0585***	1933
	(0.0226)	(0.0221)	
(l) Stayers: Always same SSD	-0.0671**	-0.0982***	1119
	(0.0288)	(0.0257)	
(m) Same dwelling for at least 24 months	-0.0297	-0.0257	1092
	(0.0363)	(0.0309)	
(n) Same dwelling for at least 18 months	-0.0588*	-0.0381	1274
	(0.0312)	(0.0307)	

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. In rows (m)-(n), the subsample of stayers is based on a questions asked in wave three.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table A.3: Baseline results - Probit specification

Dependent variable:	Spoken English (1-4 scale)			Spoken English (binary)			Interview in English		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A. Statistical Division level (SD)									
ln(Enclave size)	-0.213***	-0.229***	-0.177**	-0.325***	-0.320***	-0.240**	-0.300***	-0.326***	-0.238*
	(0.0750)	(0.0714)	(0.0714)	(0.0996)	(0.102)	(0.119)	(0.114)	(0.119)	(0.142)
B. Statistical Subdivision level (SSD)									
ln(Enclave size)	-0.0598	-0.0722	-0.00684	-0.0842	-0.0962	-0.0469	-0.0704	-0.0768	-0.0627
	(0.0466)	(0.0471)	(0.0469)	(0.0609)	(0.0631)	(0.0660)	(0.0632)	(0.0667)	(0.0738)
N	2053	2053	2053	2053	2053	2053	2053	2053	2053
Language group fixed effects	X	X	X	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X	X	X	X
Motivation/Ability proxies		X	X		X	X		X	X
Wave 1 dependent variable			X			X			X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD-level, are reported in parentheses. Data are weighted using the provided sample weights. Columns (1)-(3) show estimates of an ordered probit model. Columns (4) to (9) are binary probit models.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table A.4: Baseline results - Alternative functional forms

Dependent variable:	Spoken English			Interview in English		
	(1)	(2)	(3)	(4)	(5)	(6)
Statistical Division level (SD)						
Contact availability	-0.0245	-0.253***		-0.0249	-0.219***	
	(0.0154)	(0.0605)		(0.0159)	(0.0589)	
Contact availability^2		0.0575***			0.0489***	
		(0.0144)			(0.0143)	
Contact availability: 2nd quintile			-0.115***			-0.0672**
			(0.0355)			(0.0277)
Contact availability: 3rd quintile			-0.0452			-0.0322
			(0.0411)			(0.0408)
Contact availability: 4th quintile			-0.114***			-0.157***
			(0.0369)			(0.0366)
Contact availability: 5th quintile			-0.0940**			-0.0320
			(0.0365)			(0.0330)
N	2053	2053	2053	2053	2053	2053
Language group fixed effects	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X
Motivation/Ability proxies	X	X	X	X	X	X
Wave 1 dependent variable	X	X	X	X	X	X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. In columns (3) and (6), the effect of enclave size is estimated using dummies for each quintile of the distribution of CA^{SD}_{jk}. The omitted category is the first quintile.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table A.5: Sponsored Immigrants - Alternative measures of spoken English

Estimation:	OLS					2SLS				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Dependent variable: Both self-declares speaking English and interviews in English										
ln(Enclave size)	-0.0882***	-0.0788***	-0.0789***	-0.0787***	-0.0864***	-0.0583	-0.0510	-0.0590*	-0.0656*	-0.0646*
	(0.0253)	(0.0247)	(0.0249)	(0.0255)	(0.0265)	(0.0388)	(0.0376)	(0.0336)	(0.0352)	(0.0363)
B. Dependent variable: Composite of self-declared spoken English and interview in English										
ln(Enclave size)	-0.136***	-0.125***	-0.125***	-0.123**	-0.137***	-0.0993	-0.0953	-0.0842	-0.0994	-0.0943
	(0.0452)	(0.0442)	(0.0447)	(0.0477)	(0.0478)	(0.0887)	(0.0859)	(0.0578)	(0.0654)	(0.0648)
C. How English has improved since last interview										
ln(Enclave size)	-0.0647**	-0.0601**	-0.0596**	-0.0131	-0.0934***	-0.0915*	-0.1000*	-0.0971**	-0.0644	-0.122***
	(0.0307)	(0.0299)	(0.0301)	(0.0351)	(0.0318)	(0.0516)	(0.0521)	(0.0399)	(0.0512)	(0.0422)
N	1422	1422	1422	1127	1231	1422	1422	1422	1127	1231
Language group fixed effects	X	X	X	X	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X	X	X	X	X
Motivation/Ability proxies	X	X	X	X	X	X	X	X	X	X
Wave 1 dependent variable	X	X	X	X	X	X	X	X	X	X
Sponsor's characteristics		X	X	X	X		X	X	X	X
Sponsor's location fixed effects			X	X	X			X	X	X
Sample restriction: Sponsor in Australia for ≥ 5 years				X					X	
Sponsor's relationship with PA					X					X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. The sample is restricted to PAs sponsored by a relative. The sponsor's location is at the State level. Sponsor characteristics are number of year lived in Australia, Australian citizen status, and whether English is the language spoken at home. In columns (5) and (10), a full set of dummies for categories of the sponsor's relationship with the PA are included. In columns (6) through (10), the instrument is ln(*Enclave size*) in the sponsors' State of residence. In panel C, the lagged dependent variable is speaking English "Well" or "Very well" in wave one.
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table A.6: Reading and Writing skills

Dependent variable:	Reading skills			Writing skills		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Statistical Division level (SD)						
ln(Enclave size)	-0.0283	-0.0397**	-0.0290*	-0.0089	-0.0204	-0.0012
	(0.0204)	(0.0197)	(0.0174)	(0.0236)	(0.0238)	(0.0214)
R^2	0.422	0.453	0.508	0.404	0.442	0.511
B. Statistical Subdivision level (SSD)						
ln(Enclave size)	-0.0156	-0.0210	-0.0091	-0.0138	-0.0187	-0.0075
	(0.0147)	(0.0149)	(0.0134)	(0.0142)	(0.0132)	(0.0120)
R^2	0.422	0.453	0.508	0.428	0.443	0.511
N	2053	2053	2053	2053	2053	2053
Language group fixed effects	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X
Motivation/Ability proxies		X	X		X	X
Wave 1 dependent variable			X			X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. Both measures of language skills are binary, taking value one if the PA reads/writes English "Well" or "Very well".
***, **, * indicate significance at the 1%, 5%. and 10% levels.

Table A.7: Decomposition of SSD-level estimates

Measure of language skills:	Spoken English			Interview in English		
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	0.0037	-0.0160	-0.0075	0.0059	-0.0106	-0.0073
	(0.0209)	(0.0189)	(0.0171)	(0.0185)	(0.0179)	(0.0164)
β_2	-0.0959***	-0.0838***	-0.0581***	-0.103***	-0.0913***	-0.0610***
	(0.0244)	(0.0226)	(0.0204)	(0.0217)	(0.0225)	(0.0211)
$\rho_{SD,SSD}$	0.272***	0.273***	0.272***	0.272***	0.273***	0.271***
	(0.0292)	(0.0282)	(0.0280)	(0.0292)	(0.0282)	(0.0279)
Language group fixed effects	X	X	X	X	X	X
SSD fixed effects	X	X	X	X	X	X
Demographic characteristics	X	X	X	X	X	X
Motivation/Ability proxies		X	X		X	X
Wave 1 dependent variable			X			X

Notes: Heteroskedasticity-consistent standard errors, clustered at the SD by language-group level, are reported in parentheses. Data are weighted using the provided sample weights. The coefficients β_1 and β_2 are obtained by estimating equation (3). The coefficient $\rho_{SD,SSD}$ is from a regression of $\ln CA_{jk}^{SD}$ on $\ln CA_{jk}^{SSD}$ (and all other conditioning variables).
***, **, * indicate significance at the 1%, 5%. and 10% levels.

B Level of aggregation

To reconcile the relatively small effect of SSD-level linguistic concentration with the large effect of SD-level linguistic concentration, consider an econometric model in which contact availability has two separate effects on language acquisitions: (a) an effect of SD-level contact availability that is common to all immigrants of particular language group in a given city, irrespective of which neighborhood they live in, and (b) an additional effect of SSD-level contact availability *given* the SD-level concentration. In particular, let $CA_{jk}^{within} = CA_{jk}^{SSD}/CA_{jk}^{SD}$ denote SSD-level concentration relative to the SD-level contact availability and write

$$\begin{aligned} L_{ijk} =& \beta_1 \ln CA_{jk}^{within} + \beta_2 \ln CA_{jk}^{SD} + \dot{\gamma}' X_{ijk} + \dot{\delta}_j + \dot{\delta}_k + \dot{\varepsilon}_{ijk} \\ =& \beta_1 \ln CA_{jk}^{SSD} + (\beta_2 - \beta_1) \ln CA_{jk}^{SD} + \dot{\gamma}' X_{ijk} + \dot{\delta}_j + \dot{\delta}_k + \dot{\varepsilon}_{ijk}. \end{aligned} \tag{3}$$

Intuitively, conditional on living in a city where same-language contacts are plentiful (i.e. SD-level concentration is high), there might be little additional effect of living in a neighborhood with relatively more same-language contacts than the city average, in which case β_1 would be small. Similarly, there might be little impact of living in a high-concentration SSD (relative to the city average) if there are very few same-language contacts in the city to begin with (i.e. very low SD-level concentration).

Note that the baseline estimates of equation (2) when concentration is measured at the SSD-level (β_{SSD}) are a weighted average of β_1 and β_2:

$$\beta_{SSD} = \beta_1 (1 - \rho_{SSD}) + \beta_2 \rho_{SSD}$$

where ρ_{SSD} is the coefficient from a regression of $\ln CA_{jk}^{SD}$ on $\ln CA_{jk}^{SSD}$ (and all other conditioning variables).[37] Table A.7 reports estimates of β_1, β_2, and ρ_{SSD}.

An empirical strategy often employed in studies of contextual effects is to instrument the independent variable of interest with a corresponding variable measured at a higher level of aggregation (Evans, Oates and Schwab, 1992; Bertrand, Luttmer and Mullainathan, 2000; Cutler, Glaeser and Vigdor, 2008). For example, if we

[37] i.e. $\ln CA_{jk}^{SD} = \rho_{SSD} \ln CA_{jk}^{SSD} + \dot{\gamma}' X_{ijk} + \dot{\delta}_j + \dot{\delta}_k + \dot{\varepsilon}_{ijk}$.

instrument $\ln CA_{jk}^{SSD}$ with $\ln CA_{jk}^{SD}$, the IV estimates are in fact composite coefficients that relate directly to β_1, β_2, and π_{SD}, the coefficient from a regression of $\ln CA_{jk}^{SSD}$ on $\ln CA_{jk}^{SD}$ (i.e. the first-stage):

$$\hat{\beta}_{SSD}^{IV} = \frac{\beta_{SD}}{\pi_{SD}} = \frac{\beta_2 - \beta_1 \left(1 - \pi_{SD}\right)}{\pi_{SD}} = \beta_1 + \frac{\left(\beta_2 - \beta_1\right)}{\pi_{SD}}. \tag{4}$$

The 2SLS estimates reported in section 4.3 are a special case of equation (4).

C Data appendix

The 1996 Community Profiles report population counts for each SSD/SD by language spoken at home for 22 languages: English, Arabic (including Lebanese), Australian Indigenous Languages, Chinese languages, Croatian, French, German, Greek, Hungarian, Indonesian, Italian, Macedonian, Malay, Maltese, Netherlandic, Polish, Portuguese, Russian, Serbian, Spanish, Tagalog (Filipino), Turkish, and Vietnamese. To maximize the number of language groups for which I can compute measures of concentration, I instead use the 2001 Time Series Community Profiles, in which 1996 population counts are reported for all languages listed above, as well as for Hindi, Japanese, Khmer, Korean, Persian, Samoan, Sinhalese, Tamil, and Ukrainian.[38] Similarly, 1996 population counts for Thai and Assyrian speakers are collected from the 2006 Time Series Community Profiles. I drop Malay, Maltese and Samoan speakers because there are too few observations for these language groups in the LSIA sample. In the calculations, overseas visitors and people whose language spoken at home is not stated are excluded from the total population of each area. This may generate some measurement error, leading to an attenuation bias in the empirical analysis.

Note that between each Census, the boundaries of a few SSD were modified. Because the location reported in the LSIA corresponds to the 1996 definitions, I must adjust the population counts from the 2001 and 2006 Time Series Community Profiles so that they match the 1996 boundaries. For example, the SSD of Blacktown-Baulkham Hills was renamed Blacktown in 2001, and its size was reduced by the transfer of the Statistical Local Area of Baulkham Hills (A) to the SSD of Central Northern Sydney. Hence, to account for these changes, the population counts of Baulkham Hills (A) for the year 1996 are added to those of Blacktown (and subtracted from those of Central Northern Sydney) in the 2001 Time Series Community Profiles to reflect the 1996 boundaries. In most cases, the mapping from the 2001 to the 1996 boundaries is very straightforward, and so the 1996 population counts are calculated without error. In a few cases, only

[38] The category "South slavic, not further defined" can't be matched to one specific language in the LSIA data, hence this group is dropped. Australian Indigenous Languages are also ignored, as they are not spoken by immigrant populations.

part of a given geographic unit was transferred from one SSD to another, generating minor measurement error.

Finally, the Census Community Profiles also report population counts by country of birth for a subset of countries. If groups are assigned on that basis rather than by best language spoken, the sample size is substantially reduced (e.g. many Arabic and Spanish speakers in the LSIA1 data set were born in a country for which the population counts are not reported in the Community Profiles). The main results reported in this paper are qualitatively unchanged if country of birth is used to measure concentration or if population counts for year 2001 instead of 1996 are used (available upon request).

D Theoretical Model

To understand the mechanisms through which linguistic concentration may affect language skill acquisition, and how non-random location patterns may introduce bias in empirical estimates, it is convenient to consider the language acquisition process and the location decision in two steps. For simplicity, suppose there are only two languages in the host-country: a majority language (English), and a minority language. The proportion of minority speakers N_j varies across J cities. Consider a cohort of minority language-speaking immigrants who arrive in Australia and must decide in which city to reside and whether to learn English or not.[39] The entering cohort is sufficiently small so that its size has virtually no impact on the distribution of minority and majority language speakers in each location (i.e. values of N_j are exogenous to the model).

D.1 The learning process

This subsection focuses exclusively on immigrants who are not fluent upon arrival. Let $L_{ij} \in \{0, 1\}$ indicate the English proficiency status of individual i in area j. Immigrants who learn English incur a cost $c_{ij} = c(\mathbf{X}_i, v_i, N_j, \mathbf{G}_j)$ where $\mathbf{X}_i$ is a vector of observable individual characteristics (e.g. education, age) that affect the ease with which the individual is able to learn new languages, v_i is an unobserved idiosyncratic component (e.g. ability), and $\mathbf{G}_j$ is a vector of city characteristics such as availability of English courses. Note that the cost is allowed to depend directly on the concentration of minority speakers. The literature suggests that c_{ij} is an increasing function of N_j.[40]

[39] Immigrants already fluent in English only need to decide where to locate.

[40] In Lazear's (1999) random-encounter model, as N_j grows, the opportunity cost of learning English (i.e. the probability of "trade" occurring for a non-fluent immigrant) increases. Alternatively, a larger N_j implies that minority language-speakers rarely meet native speakers, and so opportunities to practice English are less common (Stevens, 1992). For instance, Vervoort, Dagevos and Flap (2012) find that ethnic concentration is negatively (positively) associated with the number of contacts with natives (co-ethnics) in the Netherlands, which in turn is related to greater (lower) proficiency in Dutch. Danzer and Yaman (2013) document a similar pattern in Germany.

Because individuals tend to interact disproportionately with people who speak their mother tongue, residential concentration may also reflect an underlying social network in which valuable information is disseminated. These networks can serve the function of a system of job referrals, potentially improving the labor market outcomes of its members.[41] For instance, it has been shown that once selection is accounted for, the generally uncovered negative relationship between linguistic concentration and earnings disappears, or even turns positive.[42] For the most part, studies tackling issues of selection into enclaves do not explore the different mechanisms through which concentration affect employment and earnings. In this paper, I shed light on the strength of the language acquisition channel, treating language skills as the main outcome of interest.

Individuals do not value fluency directly, but their utility is a function of earnings, which in turn depend on language skills. There are two types of jobs: those that require English, paying a wage w_j^H, and those that require limited language abilities and pay a wage $w_j^L < w_j^H$.[43] A worker i receives a type $q \in \{H, L\}$ job offer with probability $P_{ij}^q = P^q(L_{ij}, \mathbf{X}_i, v_i, N_j, \mathbf{G}_j)$. The job finding rates are functions of $\mathbf{X}_i$ and v_i, implicitly allowing for ability in language acquisition to be correlated with labour market abilities. The probability of finding type-q work is allowed to vary with linguistic concentration to reflect substantial evidence that informal job search and job referrals from family and friends are widespread (Ioannides and Loury, 2004), notably in immigrant social networks (Beaman, 2012). Under basic assumptions, some network models predict that the individual probability of finding work is increasing in the number of contacts (ties) (Patacchini and Zenou, 2012).[44]

Markets are segmented in the sense that workers fluent in English only look for type-H jobs and non-fluent workers are only employed in type-L jobs.[45] Immigrant i's expected utility in area j is separable in income,

[41]Goel and Lang (2009) and Patacchini and Zenou (2012) validate the use of ethnic concentration as a measure of network size by showing that it is related to the probability of finding a job through social networks.

[42]Munshi (2003) studies the employment probability of Mexican migrants in the United States and finds that Mexican workers are more likely to be employed, and to have a nonagricultural job (conditional on employment) if their social network is exogenously large. These results are in line with Cutler, Glaeser and Vigdor (2008), who instrument for segregation using variation in the distributions of occupations across areas and groups of immigrants. Exploiting the institutional features of refugee placement policies, Edin, Fredriksson and Åslund (2003) and Damm (2009) find strong evidence of low-ability refugees sorting into enclaves in Sweden and Denmark, respectively. Once the selection bias is removed, they find that the impact of living in an enclave on earnings is positive.

[43]Note that because the entering cohort is small, the learning decisions made by this group of immigrant has no general equilibrium effect on the return to proficiency, hence on market wages.

[44]This is not necessarily true in the short-run, because of within network competition (Calvó-Armengol and Jackson, 2004). Also, beyond a critical value of network size, the number of job matches can actually decrease because of congestion (Calvó-Armengol and Zenou, 2005; Wahba and Zenou, 2005). Also, if concentration affects labour demand (through job creation, for example), then the job finding rates captures both the dissemination of information channel and the effect on labour demand in the two types of jobs.

[45]That is, $P^H(0, \mathbf{X}_i, v_i, N_j, \mathbf{G}_j) = 0$ and $P^L(1, \mathbf{X}_i, v_i, N_j, \mathbf{G}_j) = 0$. Suppose workers are endowed with one indivisible unit of time which can either be allocated to searching English or non-English jobs. Then, as long as expected earnings of fluent workers are higher in type-H jobs ($P_j^H(1, X_i, v_i, N_j) w_j^H \geq P_j^L(1, X_i, v_i, N_j) w_j^L$), it is always optimal for them to search for English rather than non-English jobs.

the cost of learning English, and idiosyncratic preferences for location j (ϵ_{ij}):

$$U_{ij}(L_{ij}) = (P^H_{ij} w^H_j) L_{ij} + (P^L_{ij} w^L_j)(1 - L_{ij}) - L_{ij} c_{ij} + \epsilon_{ij}. \tag{5}$$

The optimal choice of language skill acquisition in location j is given by:

$$L^*_{ij}(\mathbf{X}_i, v_i, N_j, \mathbf{G}_j, w^H_j, w^L_j) = \begin{cases} 1 & \text{if} \quad \underbrace{P^H(1, \mathbf{X}_i, v_i, N_j, \mathbf{G}_j) w^H_j - P^L(0, \mathbf{X}_i, v_i, N_j, \mathbf{G}_j) w^L_j}_{\text{Benefits}(B_{ij})} \geq \underbrace{c(\mathbf{X}_i, v_i, N_j, \mathbf{G}_j)}_{\text{Costs}(c_{ij})} \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

Equation (6) can be seen as a formalization of Chiswick and Miller's (1995; 1996) language model, which defines immigrants' propensity to acquire host-country language skills as a function of three broad categories of factors: economic incentives, exposure, and efficiency. In my model, economic incentives are accounted for by the labour market benefits of English proficiency (B_{ij}). Individual characteristics $\mathbf{X}_i$ and v_i are efficiency variables, and N_j is an exposure factor.

A relatively weak sufficient condition for concentration to have a unambiguously negative effect on language acquisition is $\partial B_{ij}/\partial N_j < 0$. For instance, if language-based social networks only provide job referrals for jobs with weak English requirements (i.e. $\frac{\partial P^L_{ij}}{\partial N_j} \geq 0$ and $\frac{\partial P^H_{ij}}{\partial N_j} = 0$), the result follows. The condition is also satisfied if there are fewer type-H job opportunities in enclaves (P^H_{ij} decreases with N_j). If individuals were randomly assigned to different locations, regressing L^*_{ij} on N_j (as well as on all the other variables included in equation (6)) would provide an unbiased estimate of the average net effect of enclave size. Yet, location is not random, hence this uncomplicated empirical approach likely yields biased estimates.

D.2 The location decision and selection bias

The indirect utility of living in location j for individual i is given by $U^*_{ij} = U_{ij}(L^*_{ij})$. Individual i will therefore choose to reside in location s if and only if

$$U^*_{is} > U^*_{ij} \quad \forall j \neq s. \tag{7}$$

In this model, there is no uncertainty; immigrants know whether they will learn English or not when deciding where to live.[46] Bias in OLS specifications may arise for at least three reasons.

[46]I assume that individuals have perfect information to characterize optimal sorting patterns. The selection bias would likely

Bias 1: The classic ability bias. To illustrate this case, consider immigrants whose learning decision is perfectly inelastic with respect to N_j, that is high-ability individuals who find it optimal to learn English in every J locations (always-learners, with v_i exceeding some arbitrary $\bar{v}$), and low-ability individual who never "pick up" English in any location (never-learners). The former's indirect utility is given by $U^*_{ij} = P^H_{ij} w^H_j - c_{ij} + \epsilon_{ij}$, and the latter's by $U^*_{ij} = P^L_{ij} w^L_j + \epsilon_{ij}$. Always-learners will tend to locate out of enclaves disproportionately if (a) choosing a location with a relatively high N_j increases the cost of language acquisition more than it increases the probability of finding a type-H job, and (b) they dislike living in enclaves (e.g. $Cov(\epsilon_{ij}, N_j | v_i \geq \bar{v}) < 0$). Similarly, never-learners may select into high-concentration areas if the probability of finding type-L work is greater in enclaves, or if they have a preference for enclaves. In the extreme case of a sample made up exclusively of these two types of migrants, a negative correlation between concentration and proficiency might be found even in the absence of any effect of N_j on learning.

Bias 2: The comparative advantage bias. The effect of living in low-concentration areas on language skill acquisition may vary across individuals, and immigrants may sort on that basis. Because migrants do not value proficiency in itself, we cannot unambiguously predict whether immigrants whose learning decision is the most sensitive to linguistic concentration are systematically more or less likely to reside out of enclaves than individuals who are unaffected by concentration.

Bias 3: Selection on pre-immigration proficiency. Immigrants already fluent in English upon arrival do not have to make a learning decision. In every location, their indirect utility is given by $U^*_{ij} = P^H_{ij} w^H_j + \epsilon_{ij}$ and any sorting of these individuals is independent of learning costs. If English-job opportunities are more scarce in enclaves, or if this group of migrants simply prefer to live in low-concentration areas, then any regression of the "stock" of language skills at one point in time on linguistic concentration might be plagued by a problem of reverse causality: we may conclude that some immigrants became proficiency because they were living in low-concentration areas, whereas in reality they chose to live in these locations because they were already fluent upon arrival.[47]

be less severe if people did not locate optimally. For example, one may expect most sponsored immigrants to take their initial location as given (i.e. they move wherever their sponsor lives). Then, if moving costs are substantial, an immigrant may stay in the city where his sponsor is located even if he could achieve a higher utility living somewhere else. Similarly, uncertainty or imperfect information with respect to city characteristics may lead to sub-optimal location decision, hence to a smaller selection bias.

[47] There is considerable empirical evidence of this type of behavior (Bauer, Epstein and Gang, 2005; Bayer, McMillan and Rueben, 2004).