Aoki, Yu; Santiago, Lualhati

Working Paper

Deprivation, Segregation, and Socioeconomic Class of UK Immigrants: Does English Proficiency Matter?

IZA Discussion Papers, No. 11368

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Aoki, Yu; Santiago, Lualhati (2018) : Deprivation, Segregation, and Socioeconomic Class of UK Immigrants: Does English Proficiency Matter?, IZA Discussion Papers, No. 11368, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/177172

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11368

Deprivation, Segregation, and Socioeconomic Class of UK Immigrants: Does English Proficiency Matter?

Yu Aoki
Lualhati Santiago

FEBRUARY 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11368

Deprivation, Segregation, and Socioeconomic Class of UK Immigrants: Does English Proficiency Matter?

Yu Aoki
University of Aberdeen, HERU and IZA

Lualhati Santiago
Office for National Statistics, UK

FEBRUARY 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Deprivation, Segregation, and Socioeconomic Class of UK Immigrants: Does English Proficiency Matter?*

This paper studies the causal effect of English proficiency on residential location outcomes and the socioeconomic class of immigrants in England and Wales, exploiting a natural experiment. Based on the phenomenon that young children learn a new language more easily than older children, we construct an instrument for English proficiency using age at arrival in the United Kingdom. Taking advantage of a unique dataset, we measure the extent of residential segregation along different dimensions, and find that poor English skills lead immigrants to live in areas with a high concentration of people who speak their same native language, but not necessarily in areas with a high concentration of people of their same ethnicity or country of birth. This finding could suggest that, for immigrants with poor English proficiency, what matters for their residential location decision is language spoken by residents, as opposed to ethnicity or country of birth. We also find that language skills have an impact on the occupation-based socioeconomic class of immigrants: Poor English skills reduce the likelihood of being in the occupation-based class 'higher managerial and professional' and increase that of being in the class 'self-employment'.

JEL Classification: J15, J61, R23, Z13

Keywords: language skills, deprivation, residential segregation, socioeconomic class

Corresponding author:
Yu Aoki
Department of Economics
University of Aberdeen
Dunbar Street
Old Aberdeen, AB24 3QY
United Kingdom
E-mail: y.aoki@abdn.ac.uk

* We gratefully acknowledge the permission of the Office for National Statistics (ONS) to use the Longitudinal Study, and the help provided by staff of the Centre for Longitudinal Study Information and User Support, which is supported by the ESRC Census of Population Programme (Award Ref: ES/K000365/1). Financial support from the Scottish Institute for Research in Economics and Carnegie Trust for the Universities of Scotland is also gratefully acknowledged. This work contains statistical data from the ONS which is Crown Copyright and all statistical results remain Crown Copyright. The use of the ONS Statistics statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. The authors alone are responsible for the interpretation of the data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

1. Introduction

Millions of international migrants live across the world. Globally, 244 million international migrants were recorded in 2015, where Oceania recorded the highest proportion of immigrant population, 21%, followed by North America, 15%, and Europe, 10% (United Nations, 2015).[1] We live in an increasingly diverse society and the social integration of immigrants is a high priority for governments in many developed countries. It is becoming increasingly important to understand the key factors that influence immigrant integration, and we focus on one of these possible factors: Language skills. Language facilitates communication with native residents and colleagues which in turn can affect immigrants' living environments in a number of ways. To date, although the impact of language skills on the economic outcomes of immigrants, in particular earnings, has been intensively studied, there is limited knowledge on its impact on their residential outcomes and socioeconomic class. This paper contributes to this knowledge by studying the causal impact of English proficiency on the socioeconomic class of immigrants and a variety of residential outcomes that measure the extent of segregation and deprivation in the neighbourhoods that they reside in.

Our paper makes three contributions to the literature on immigrant outcomes in a host country. First, we construct measures of the extent of residential segregation of immigrants aimed at capturing the concept of enclave along three dimensions: Main language spoken by residents (language enclave), ethnicity (ethnic enclave), and country of birth (country-of-birth enclave). We analyse which of these three dimensions of enclave is affected by the language skills of immigrants when they are making a residential location decision. Second, we study the extent of deprivation in the neighbourhoods immigrants live in, where the extent of deprivation is measured at a small geographical area of an average of 1,500 individuals. We can conduct this analysis by exploiting a unique dataset from the Office for National Statistics (ONS) Longitudinal Study, which links individual-level data from the England and Wales 2011 Census to the Indices of Deprivation in England. The various measures of neighbourhood deprivation that we exploit capture different dimensions of deprivation (i.e., income, employment, and health deprivation of residents) allowing us to analyse residential environments immigrants with different levels of English proficiency live in. Third, we analyse how language proficiency affects the socioeconomic class of immigrants. This is an important but difficult question to investigate due to the difficulty in measuring socioeconomic classes. To address this challenge, we rely on the National Statistics Socioeconomic Classification (NS-SEC), which is a measure based on occupation aimed at capturing the socioeconomic position of an individual in the United Kingdom (UK). The NS-SEC is widely used in UK statistics, and although it is occupation-based, it has rules to provide coverage for the entire adult population. We are not aware of any other studies that have provided arguably causal evidence on the impact of language proficiency on residency in a language enclave, the extent of deprivation of the area where immigrants reside, and their socioeconomic class.

A major challenge to identify the causal effect of language is the endogeneity of language skills. First, there may be reverse causality. For example, poorer English skills may lead an individual to live in an area

[1] International migrants are defined as people living in a country different from the country in which they were born.

with a higher concentration of individuals speaking their native language, while at the same time living close to individuals who speak their native language may make it more difficult to improve their English skills. Second, there may be unobserved heterogeneity across individuals that is correlated with both English skills and immigrant outcomes (e.g., ability). Third, the self-reported measure of English skills used in our analysis may contain measurement error. To address these possible endogeneity concerns, we use an instrumental variable (IV) strategy where age at arrival in the UK is exploited to construct an instrument for English language skills. The idea of using age at arrival is proposed by Bleakley & Chin (2004), and is based on the "critical period hypothesis of language acquisition" by Lenneberg (1967), suggesting that individuals exposed to a new language within the critical period of language acquisition (i.e., childhood) can learn it more easily than those exposed to it outside of this critical period. This hypothesis implies that non-Anglophone immigrants who arrived in the UK when they were young would on average have better English skills than non-Anglophone immigrants who arrived at an older age.

However, age at arrival is unlikely to be a valid instrument for English skills on its own because it may influence an immigrant's socioeconomic outcomes through other channels than language acquisition; for example, through cultural assimilation. To overcome this problem, we incorporate immigrants born in Anglophone countries in our analysis to partial-out all age-at-arrival effects except language acquisition. After arriving in the UK, immigrants born in non-Anglophone countries would be exposed to a new language in addition to the new UK environment, while those born in Anglophone countries would be exposed to the same new UK environment but not to a new language. This implies that, conditional on individual characteristics, any difference observed in the outcomes of early- and late-arrivers born in Anglophone countries reflects age-at-arrival effects only, while this same difference observed in immigrants born in non-Anglophone countries reflects those same age-at-arrival effects and an additional effect, the language effect. Thus, the difference in an outcome of early- and late-arrivers for those born in non-Anglophone countries in excess of the equivalent difference for those born in Anglophone countries can arguably be attributed to the effect of language. Furthermore, among non-Anglophone countries, there is variation in how close their native languages are to English. For example, Dutch is linguistically more similar to English than Vietnamese. To account for this heterogeneity in similarity of immigrants' native languages to English, we construct our instrument by interacting age at arrival with linguistic distance between the origin-country language and English.[2]

Our IV estimates indicate that language skills have an impact on a residential segregation outcome: Poorer English skills significantly lead immigrants to live in areas with a higher concentration of individuals who speak their native language, but not necessarily in areas with higher concentrations of people with the same ethnicity or country of birth. This finding could suggest that, for immigrants with poor English language skills, what matters for their residential location decisions is language spoken by residents, as opposed to ethnicity or country of birth. We also find a sizeable impact on socioeconomic class. For example, poor English skills increase the likelihood of being in the occupation-based class 'self-employed' and reduce that of being in the class 'higher managerial and professional'. Supplementary regressions support the possibility that education is a mediator for the effects of language on socioeconomic class but not on residential segregation. Turning to the extent of deprivation in the areas where immigrants live, our

[2]Clarke & Isphording (2017) is the first to use this instrument for English proficiency in their study of the causal impact of English proficiency on immigrant health.

results indicate that poorer English skills lead immigrants to live in more deprived neighbourhoods, but these effects are imprecisely estimated.

The rest of the paper is structured as follows: Section 2 reviews the literature on the effects of language skills on residential and economic outcomes of immigrants, while Section 3 presents the identification strategy. Section 4 describes datasets and sample specifications, and explains how we construct our main variables such as various measures of residential segregation. Section 5 presents our empirical results and Section 6 conducts a series of robustness checks. Finally, Section 7 discusses policy implications and conclusions.

2. Literature Review

To the best of our knowledge, we are not aware of any research that has analysed the relation between host-country language proficiency and the extent of deprivation in the neighbourhoods immigrants live in. In contrast, the relation between language proficiency and residential segregation has been extensively studied by researchers in economics and other disciplines. In a seminal paper, Lazear (1999) proposes a model of cultural and language assimilation of immigrants that inversely relates an immigrant's language proficiency to the proportion of local population who speak their same native language. This model predicts that an immigrant residing in an area with a large proportion of people who speak their native language has less incentive to learn a new language. On the other hand, the model of spatial assimilation developed by Massey (1985) suggests that ethnic enclaves are a natural first stage for immigrants when entering a country, but they leave the enclaves once they have integrated to the new country's culture.

Empirically, a large number of papers have investigated the correlation between host-country language proficiency and ethnic residential segregation (e.g., Logan et al. 2002; Dustmann & Fabbri 2003; Bauer et al. 2005; Iceland & Scopilliti 2008; Beckhusen et al. 2013). Broadly, they find that having lower English language skills is positively correlated with ethnic-enclave residency. For example, Dustmann & Fabbri (2003), in their analysis of the determinants of language skills, find strong negative correlations between ethnic minority concentrations and English language skills of ethnic minority immigrants in the UK. There is also research on residency in language enclaves (e.g., Chiswick & Miller, 1995, 2005). For example, Chiswick & Miller (2005) study the relation between living in a language enclave and English proficiency of immigrants in the United States (US), and find that English proficiency is negatively associated with a higher extent of minority language concentration.

A limitation of these studies is that it is not clear which direction causality runs: Namely, whether poor language skills cause immigrants to live in enclaves, or whether they have poor language skills because they live in an enclave. Furthermore, there may be unobserved heterogeneity across individuals that affects residency in enclaves, such as variation in cultural attitude, which may be correlated with their language proficiency. Bleakley & Chin (2010) is the first paper to address this potential endogeneity issue using an IV for English proficiency, which is an interaction between age at arrival in the US and an indicator for being born in a non-Anglophone country. They find weak evidence of the effects of English proficiency on ethnic and country-of-birth enclave residency, unlike previous studies that found strong correlations between host-country language proficiency and enclave residency. This could suggest that the findings of the previous studies are biased due to the endogeneity of language proficiency, although it is important

to note that, to measure ethnic residential segregation, Bleakley & Chin (2010) use a relatively large unit called public-use microdata area (PUMA). The PUMA contains a minimum of 100,000 residents which may not be small enough to capture the concept of an enclave. We aim to overcome this issue by measuring residential segregation at a smaller geographical level which is arguably more suitable for the analysis of the effect of language proficiency on enclave residency. In addition, we aim to provide a more in-depth analysis of immigrants' segregation than previous studies, by comparing the effects of language skills on three different measures on segregation based on language, ethnicity and country of birth.

Turning to studies on the socioeconomic class of immigrants, we are not aware of any research that has analysed the relation between language proficiency and the socioeconomic class. Previous research most closely related to this topic are the studies on the relation between language proficiency and various labour market outcomes. In particular, the effects of language proficiency on earnings of immigrants have been extensively studied (e.g., Dustmann 1994; Chiswick & Miller 1995; Shields & Price 2002; Bleakley & Chin 2004; Miranda & Zhu 2013). Likewise, the impact of language proficiency on employment has also been intensively studied (e.g., Miller & Neo 1997; Dustmann & Fabbri 2003; Gonzalez 2005; Clausen et al. 2009). For example, Dustmann & Fabbri (2003) investigate the causal effect of English skills on employment probabilities among immigrants in the UK. Using propensity score matching and an instrumental variable estimation strategy to address the endogeneity issue of English skills, they find that better English skills raise employment probabilities. Although not extensive, there is some research on the effects of language on occupational choices in the context of studying mediating factors for the effects of language on earnings (e.g., Kossoudji, 1988; Aldashev et al., 2009; Chiswick & Miller, 2009). For example, Chiswick & Miller (2009) find that, in their earnings equations, effect sizes of English proficiency greatly diminish once occupations are controlled for, and suggest that some of the earnings disadvantages of immigrants in the US with limited English skills are likely due to this deficiency placing them in lower earning occupations.

3. Identification Strategy

We estimate the causal effect of English language proficiency on residential outcomes and the socioeconomic class of childhood immigrants in England and Wales by regressing these outcomes on a measure of English proficiency, controlling for various individual characteristics. The following model is specified:[3]

$$outcome_{ica} = \alpha_0 + \alpha_1 proficiency_{ica} + X'_{ica}\xi + \gamma_c + \delta_a + u_{ica} \quad (1)$$

where $outcome_{ica}$ represents the outcome of individual i born in country c who arrived in the UK at age

[3]Some of the outcomes that we analyse are dummy variables. Although we could specify non-linear models such as a logit model to analyse these outcomes, we use linear models for all outcomes for the following reasons. First, this allows us to be consistent in our model specification across regressions. Second, linear models have a more straightforward interpretation than non-linear models when working with instrumental variables. Angrist & Pischke (2009) argue that, although a non-linear model may fit the conditional expectation function for limited dependent variables more closely than a linear model, marginal effects computed from these two types of models are very similar.

a, and $proficiency_{ica}$ is a measure of English language skills. The individual characteristics, X_{ica}, and the parameter ξ are $K \times 1$ vectors, where K is the number of variables capturing individual characteristics such as age and gender. γ_c and δ_a are country-of-birth and age-at-arrival fixed effects, respectively, and u_{ica} is the disturbance term.

The main coefficient of interest is α_1, which measures the effect of English skills on the socioeconomic outcomes of immigrants. An econometric challenge to estimate equation (1) is the endogeneity of English skills. First, the socioeconomic outcomes of immigrants may affect their English skills (reverse causality). Second, unobserved individual characteristics, such as ability, are likely to be correlated with both English skills and our outcome variables. Third, our self-reported measure of language proficiency may contain measurement error. Thus, using OLS to estimate α_1 is unlikely to produce a causal estimate of the effect of English proficiency.

To identify the causal effect, we estimate equation (1) using the IV estimator which requires an IV that gives exogenous variation in English skills. To construct an IV for language skills, following the idea of Bleakley & Chin (2004), we exploit age at arrival in the UK. Their idea of using age at arrival in the host country is based on the "critical period of language acquisition" hypothesis (Lenneberg, 1967), which states that an individual exposed to a new language during the critical period of language acquisition (childhood) can learn the language relatively easily.[4] This hypothesis implies that, among immigrants from a non-Anglophone country, those who arrive in the UK at a young age can learn English relatively easily, while those who arrive at an older age find it harder to learn English.

For a variable to serve as an IV for English skills, the following assumptions are required: (i) it does not appear in equation (1) and (ii) it is uncorrelated with any other determinants of the socioeconomic outcomes of immigrants apart from proficiency in English. However, age at arrival per se is unlikely to satisfy these assumptions because it affects the extent of assimilation apart from language acquisition. For example, age at arrival may affect immigrant socioeconomic classes through knowledge about employment practice or a better social network in the UK. To overcome this problem, we incorporate immigrants born in Anglophone countries in our analysis to partial out all age-at-arrival effects except for language acquisition. On arrival in the UK, all immigrants are exposed to a new environment, but only those born in non-Anglophone countries encounter a new language. Thus, conditional on individual characteristics, differences in outcomes of early- and late-arrivers from Anglophone countries would only reflect age-at-arrival effects, while differences in outcomes of immigrants from non-Anglophone countries would reflect those same age-at-arrival effects and an additional effect, the language effect. Therefore, a difference in the outcomes between early- and late-arrivers born in non-Anglophone countries in excess of the corresponding difference for immigrants born in Anglophone countries can arguably be attributed to the effect of language.

[4]Lenneberg (1967) observes that, until early teens, individuals have an innate flexibility for the organisation of brain functions necessary for the acquisition of a language. If basic language skills have not been acquired by puberty, they tend to remain deficient for the rest of their life because the ability to adjust to physiological demands for verbal acquisition declines sharply after puberty due to physiological changes in the brain.

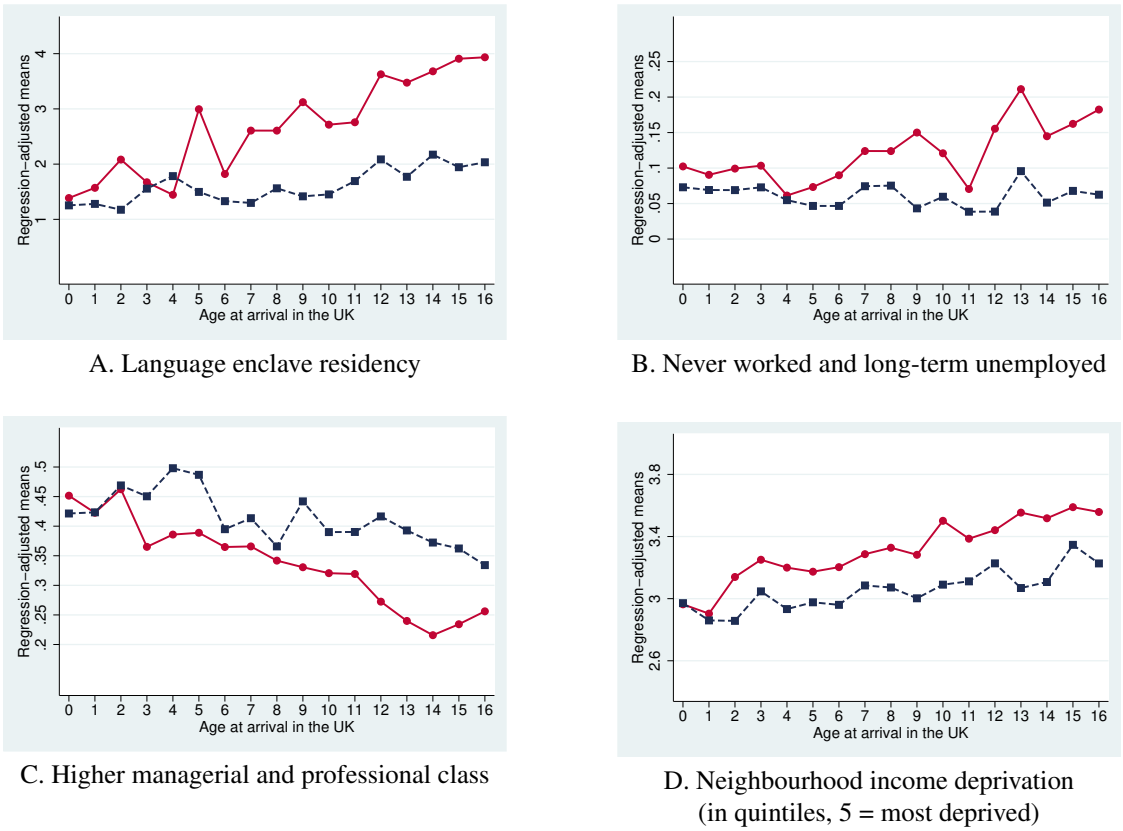

Figure 1: Immigrant outcomes by age at arrival

Notes: Immigrant outcomes are plotted by age at arrival where the outcomes are extent of residential segregation in terms of the main language spoken by residents (panel A), likelihood of being in the classes 'never worked and long-term unemployed' (panel B) and 'higher managerial and professional' (panel C), and extent of income deprivation in the neighbourhood immigrants live in (panel D). Each outcome is regression adjusted for age, sex and race. The sample corresponds to childhood immigrants aged 20 or over at the time of the 2011 Census.
Source: ONS Longitudinal Study.

Figure 1 plots immigrant outcomes by age at arrival, where these outcomes are extent of language residential segregation (panel A), likelihood of being in the occupation-based classes 'never worked and long-term unemployed' (panel B) and 'higher managerial and professional' (panel C), and extent of income deprivation in the neighbourhood immigrants live in (panel D).[5] The solid and dashed lines correspond to immigrants from non-Anglophone and Anglophone countries, respectively. Figure 1 indicates that, for late arrivers, the two series appears to diverge, although the pattern is not clear in panel D. Later

[5] As there are numerous outcome variables, we do not report graphs for every outcome to save space. Instead, we report the relation between age at arrival and each immigrant outcome (i.e., reduced-form estimates) in Table 2.

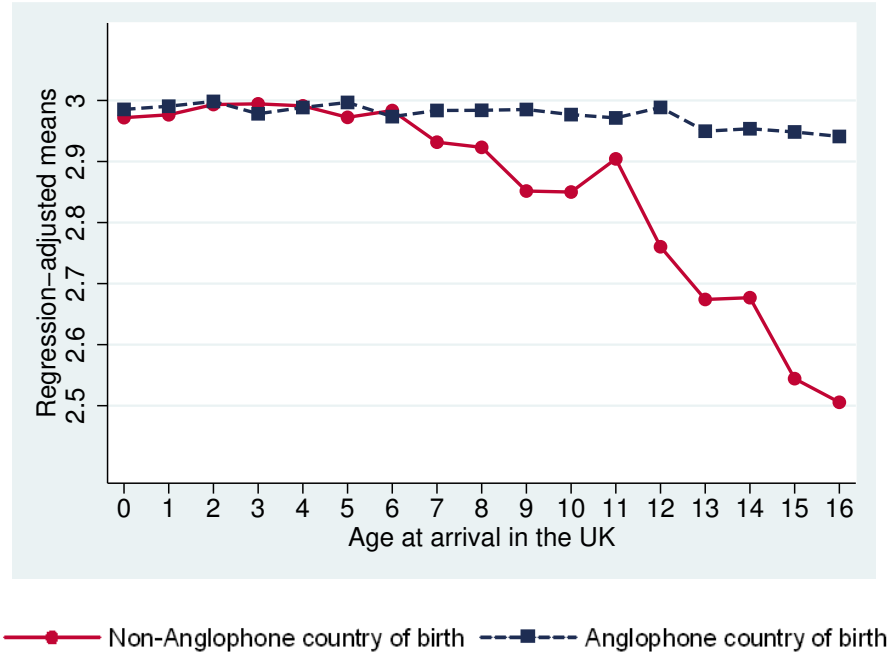

Figure 2: Age at arrival and English proficiency

Notes: Figure plots the average ordinal measure of English proficiency, where 3, 2, 1, and 0 correspond to speak English "very well", "well", "not well", and "not at all", respectively. English proficiency is regression adjusted for age, sex and race. The sample corresponds to childhood immigrants aged 20 or over at the time of the 2011 Census.
Source: ONS Longitudinal Study.

arrivers from non-Anglophone countries tend to cluster in areas with a higher concentration of residents who speak their native language (panel A), are more likely to be in the occupation-based class 'never worked and long-term unemployed' (panel B), are less likely to be in the class 'higher managerial and professional' (panel C), and tend to live in the neighbourhoods where residents are relatively more deprived (panel D). An interesting observation from Figure 1 is that immigrants from Anglophone countries also exhibit age-at-arrival effects. This observation implies that, apart from the effect of language, age at arrival is likely to have direct effects on immigrants' socioeconomic outcomes, confirming that age at arrival per se is not a valid instrument and it is important to control for age-at-arrival fixed effects in equation (1).

Figure 2 shows the relation between English language proficiency and age at arrival of immigrants who arrived in the UK during their childhood. Figure 2 shows that immigrants born in Anglophone countries score between 2.9 and 3 in the ordinal measure of English proficiency, where 3 corresponds to "speaks very well", and are generally proficient in English irrespective of their age at arrival. In contrast, immigrants born in non-Anglophone countries who arrived in the UK after age eight report having poorer English language skills than those who arrived at an earlier age. The two series start diverging at around age nine and, for those born in non-Anglophone countries, the later they arrived, the poorer their English

is on average, which is consistent with the critical period hypothesis.

In addition, there is heterogeneity in how similar immigrants' native languages are to English, and this may affect their capacity to become proficient in English. For example, an immigrant with a language that is more similar to English (e.g., Dutch) will find it easier to learn English than an immigrant with a native language that is very different to English (e.g., Vietnamese). To account for this heterogeneity in the similarity of the native language to English, we follow Clarke & Isphording (2017) and exploit linguistic distance between English and home-country language, $ldist_c$, to construct our instrument, θ_{ica}:

$$\theta_{ica} = max(0, a_i - 8) \times ldist_c \tag{2}$$

where a_i is age at arrival for individual i and the function $max(0, a_i - 8)$ corresponds to the additional years after age eight for those who arrived in the UK after age eight, and zero otherwise. An assumption underlying equation (2) is that, for those who arrived at age eight or before, there is no difference in English proficiency across immigrants born in different countries, but English proficiency and age at arrival are linearly related after age eight for immigrants born in non-Anglophone countries. We choose age eight as the cut-off value because Figure 2 indicates that the two series start diverging at around age nine.[6] Using the instrument in equation (2), the relation between English proficiency and age at arrival, which corresponds to our first-stage equation, can be specified as follows:

$$proficiency_{ica} = \beta_0 + \beta_1 \theta_{ica} + X'_{ica}\zeta + \gamma_c + \delta_a + u_{ica} \tag{3}$$

where the individual characteristics, X_{ica}, and the parameter ζ are $K \times 1$ vectors, where K is the number of variables capturing individual characteristics. γ_c and δ_a are country-of-birth and age-at-arrival fixed effects, respectively, and u_{ica} is the disturbance term.

4. Data and Sample

4.1. *Data*

To analyse the impact of English language skills on immigrant outcomes, we use an individual-level dataset from the ONS Longitudinal Study of England and Wales, which contains a sample of approximately 1% of the population of England and Wales. All our individual characteristics are collected from the 2011 Census sample within the ONS Longitudinal Study, including our measure for English proficiency, which is a self-reported ordinal measure that takes values 3, 2, 1, and 0, corresponding to speak English "very well", "well", "not well", or "not at all", respectively. We also extract our measure of parental education from the ONS Longitudinal Study by tracking the individuals in our dataset through all censuses contained in the Longitudinal Study. Once we have identified their parents, we assign them to the individuals in our sample. Concerning the variables used in the section of robustness checks, macro-level

[6] We have also tried the age of 11 as a cut-off value because we observe a kink at age 11 in Figure 2. Our results are not sensitive to this change in the cut-off value.

origin-country characteristics are obtained from the World Development Indicators 2015.[7]

To create our instrument for English proficiency, we exploit two census variables, country of birth and age at arrival in the UK,[8] and a measure of linguistic distance between English and origin-country language. We measure linguistic distance using a variation of the Levenshtein distance computed by Isphording & Otten (2014). Following a procedure to evaluate phonetic similarity between different languages, developed by the Max Planck Institute for Evolutionary Anthropology, they compute the extent of similarity between languages in percentages. The measure of linguistic distance they construct is a standardised and continuous measure of the distance between languages based on phonetic similarity, where a higher number indicates a greater linguistic distance. Despite its purely descriptive nature that does not require any prior knowledge on language relations, this measure of linguistic distance is highly correlated with other linguistic distance measures such as those developed by linguists based on language families. We assign linguistic distance based on the predominant language in the country of birth of immigrants. In the case of immigrants born in a multilingual country, we assign the predominant native language of the country. For immigrants born in a country where English is an official language and the predominant language spoken, we assign linguistic distance of zero. The following sub-sections detail the construction of our outcome variables.

Residential segregation

After linking the ONS Longitudinal Study to the local-authority level data from the 2011 Census obtained from ONS Nomis,[9] we construct the measure of residential segregation using an index of relative clustering following Borjas (2000), defined as:

$$Relative\ Clustering\ Index_{ij} = \frac{N_{ij}/N_j}{N_i/N} \tag{4}$$

where $i = 1, ..., I$ represents the languages and $j = 1, ..., J$ represents the local authorities. N_{ij} is the total number of persons reporting language i as their main language and living in local authority j, N_j is the total number of persons living in local authority j, N_i is the total number of persons reporting language i as their main language in England and Wales, and N is the total population in England and Wales. This relative clustering index is based on the 'exposure index', corresponding to the numerator of equation (4), which gives the fraction of people in a local authority reporting a particular language as their main language. Although the exposure index is widely used in the literature that studies immigrant enclaves (e.g., Borjas 2000; Edin et al. 2003; Bauer et al. 2005), a problem with this index is that it can underweight the available contacts for small ethnic groups. The relative clustering index is a better measure (Bertrand et al., 2000), since it deflates the exposure index by the proportion of people reporting a particular language i in the

[7]The World Development Indicators 2015 are downloaded from: `http://data.worldbank.org/data-catalog/world-development-indicators`

[8]Age at arrival in the UK is derived from the date that a person last arrived to live in the UK and their age. Short visits away from the UK are not counted in determining the date that a person last arrived. The age of arrival is only applicable to usual residents who were not born in the UK and does not include usual residents born in the UK who have emigrated and since returned.

[9]The 2011 Census data for local authorities can be downloaded from ONS Nomis: `https://www.nomisweb.co.uk/`.

whole of England and Wales (i.e., the denominator of equation (4)). The relative clustering index in equation (4) captures the share of individuals reporting the same native language in the local authority where an immigrant lives in. It takes value one if the proportion of people speaking language *i* living in local authority *j* is the same as the proportion of people speaking that language in England and Wales. If the relative clustering index is greater than one, then the group of individuals speaking that language is overrepresented in that particular local authority, whereas if the index is smaller than one, the group is underrepresented in that particular local authority.

In addition to measuring immigrant segregation based on their main language, we measure it based on their ethnicity and their country of birth. Each of these measures captures residential segregation along different dimensions: An ethnic group includes anyone who reports having a particular ethnic group, irrespective of whether they were born in the UK, whereas a country-of-birth group only includes individuals born in a particular country. These different measures of segregation allow us to investigate whether and how much English language skills affect these different dimensions of immigrant residential segregation.

The geographical unit we use for our analysis is the local authority district, which is an administrative division in the UK. There were 348 local authority districts in England and Wales at the 2011 Census, with an average size of 161,138 individuals. Using this geographical unit has some advantages. First, it is large enough: This is important because an individual does not necessarily interact with his immediate neighbours, but may have different networks of people (e.g., family, friends and colleagues) with whom they can interact frequently provided they have easy access to them, which happens if they live within a reasonable distance. In addition, choosing small areas could create measurement error problems in the case of immigrant groups with few observations. The second advantage of using local authorities is that they are not too large, as is the case with regions, which are too large to allow us to make the assumption that individuals could interact and meet other individuals from their same language, country or ethnic group. The third advantage is that local authority districts are administrative divisions. This is also important as it ensures that transport communications are likely to exist and be easily accessible. This latter motive makes an administrative division better than a census division for the purpose of capturing possible interactions with other group members. In this respect, we provide an alternative approach to Bleakley & Chin (2010), who also analyse the impact of English skills on residential segregation but use a unit of geography created by the US Census that do not coincide with administrative geographic boundaries.[10] Using administrative boundaries could be a better way of defining our geographical unit as it makes it more likely that both workplace and residential interactions are taken into account, and both types of interactions can determine the decision of where to live, especially in the case of an individual with poor English skills.

Neighbourhood deprivation

We measure neighbourhood deprivation using data from the English Indices of Deprivation 2015, which are published by Ministry of Housing, Communities and Local Government (2015). These indices measure relative neighbourhood deprivation at a small-area level, called the ONS Lower-layer Super Output

[10] Bleakley & Chin (2010) base their analysis on PUMAs, which are census-created geographies that sum to at least 100,000 individuals. A PUMA can be made of various counties, but also some counties can have more than one PUMA. PUMAs and counties coincide only around 5 percent of the times.

Areas (LSOAs). LSOAs are small areas designed to be of similar population size with a minimum of 1,000 individuals and a maximum of 3,000 (between 400 and 1,200 households), which have an average of approximately 1,500 residents or 650 households. We have matched our individuals to these indices corresponding to the area in which they were living at the time of the 2011 Census. Three domains of the English Indices of Deprivation are exploited: Income deprivation, employment deprivation, and health deprivation. For each of these domains, quintiles are calculated, ranking the 32,844 LSOAs in England from least deprived to most deprived and dividing them into five equal groups. We create one variable for each domain, and each of these variables takes values 1 to 5, where 1 corresponds to the least deprived area and 5 corresponds to the most deprived area.

For the analysis of neighbourhood deprivation, we only use a sample of individuals who were living in England at the time of the 2011 Census. This is because, although there are the Welsh Indices of Deprivation, these indices measure relative levels of deprivation within Wales, and thus the Deprivation Indices of England and that of Wales are not directly comparable.

Socioeconomic class

In order to measure the socioeconomic class of immigrants, we use the NS-SEC which is aimed at capturing socioeconomic positions of individuals in the UK. The NS-SEC was developed from a socioeconomic classification known as the Goldthorpe Schema (Goldthorpe, 2007). It is an occupation-based socioeconomic classification that takes into account employment relations and conditions of occupations (ONS, 2010). The NS-SEC distinguishes three forms of employment relations: Service relationship, labour contract, and intermediate.[11] Conditions of occupation are measured by occupational category and employment conditions[12] — e.g., whether an individual is an employer, self-employed or an employee. Based on these aspects, each individual is allocated to one of 17 categories which is used as a base for the NS-SEC. We use the version of the NS-SEC that consists of five classes: Class 1 — Higher managerial, administrative and professional occupations (labelled as higher managerial and professional for brevity), e.g., solicitor, medical practitioner; Class 2 — Intermediate occupations, e.g., secretary, nursery nurse; Class 3 — Small employers or own-account workers (labelled as self-employed for brevity); Class 4 — Lower supervisory and technical occupations, e.g., plumber, motor mechanic; Class 5 — Semi-routine and routine occupations, e.g., cleaner, porter. We create one indicator variable per class for our analysis.

It is important to note that the NS-SEC is a household-level measure. A household member's own position may have less relevance to their life chances than that of another family member. In order to allow for the interdependence and shared conditions of household members, one household member is chosen as a reference person, defined as the person responsible for providing the accommodation, and that person's position is used to stand for all of the household members. To understand the assignment rule, suppose that there is person X who is a medical doctor and owns a house whose spouse is a nursery nurse. In this case, both the person X and their spouse will be assigned the class of higher managerial and

[11] In the service relationship, the employee renders service to the employer in return for compensation such as salary and job security. Under a labour contract, the employee gives discrete amounts of labour in return for a wage. The intermediate relationship combines aspects from both the service relationship and labour contract.

[12] Occupational category corresponds to occupation coded to occupational unit group of the Standard Occupational Classification 2010 (ONS, 2010).

professional (i.e., Class 1) because the person X is the household reference person in this household.

For individuals who are not in employment, the assignment rule works as follows. First, NS-SEC has a separate category for those who have been in long-term unemployment, defined as being unemployed for one year and three months or longer,[13] and who have never worked *despite that they wished to work*. This category is aimed at capturing involuntary exclusion from employment, and we create one indicator variable for this category.[14] Second, those individuals who are not classified as long-term unemployed or never worked but not in employment (e.g., individuals who are short-term unemployed) are classified according to their last main occupations. For the purpose of illustration, suppose that there are two short-term unemployed individuals, one was a medical practitioner and the other was a porter. The former and latter persons will be assigned to Class 1 and Class 5, respectively, despite the fact that both individuals are currently unemployed, because the purpose of the NS-SEC is to capture socioeconomic positions in the UK instead of capturing a temporary unemployment status.

4.2. *Sample*

Age restriction

Our sample consists of individuals in the ONS Longitudinal Study dataset who were present in the 2011 Census, aged 20 or older at the time of the 2011 Census, and are childhood immigrants, defined as individuals born outside of the UK who moved into the UK at age 16 or earlier. We impose this age-at-arrival restriction and assume that these childhood immigrants did not make a migration decision on their own, but moved into the country following their parents or guardians. For the analysis of socioeconomic class, we further restrict our sample to those aged between 20 and 60 and not in full-time education.

Country classification

To implement our identification strategy, we include two types of immigrants in our sample: (i) individuals born in a non-Anglophone country where English is not an official language (treatment group) and (ii) individuals born in an Anglophone country (control group). We classify a country as Anglophone if English is an official language and the predominant language spoken in the country.[15] We exclude from our sample individuals born in countries where English is an official language but not the predominant language spoken because it is not clear to what extent they were exposed to English prior to their arrival in the UK. This rule drops immigrants from countries such as India and Pakistan who account for significant proportions of UK immigrants.

[13] The status of long-term unemployed applies to the individuals who were last time in employment on 31 December 2009 or earlier, corresponding to those who were unemployed for one year and three months or longer at the time of the census date 27 March 2011.

[14] We are unable to create an indicator variable each for those (i) who have been in long-term unemployment and (ii) who have never worked despite that they wished to work separately because the NS-SEC bundles them in the same category. Note however that the individual who has voluntarily never worked because, for example, the person is looking after a home, will be assigned the class of their household reference person.

[15] The World Almanac and Book of Facts 2011 is used to classify countries.

Propensity-score screening

For our IV strategy to identify the causal effect of language skills, we need an assumption that those born in Anglophone and non-Anglophone countries are exposed to the same age-at-arrival effects, except for the effect of language. However, one could question the validity of this assumption because immigrants from the two sets of countries may on average have different background characteristics that might differently affect their socioeconomic outcomes. For example, Europeans account for a large proportion of immigrants from non-Anglophone countries in our sample, and European countries share similarities with the UK in various aspects due to, for instance, the presence of the European Union (EU) and a long history of cultural, social and political interactions, making it easier for migrants from Europe to adapt to the new environment in the UK. Likewise, a large proportion of immigrants from Anglophone countries come from Commonwealth countries, which share commonalities with the UK regarding, for example, culture and legal systems, also making it potentially easier for these individuals to adapt to the new environment in the UK. To deal with this type of concerns, we compute propensity score for being born in a non-Anglophone country, and use it as a tool to systematically select a sample before running regressions. The propensity score is defined as follows:

$$p(D_i = 1|X_i) = F(X_i) = \frac{1}{1+e^{-(\mu_0 + X_i'\nu)}} \tag{5}$$

where $p(\cdot)$ is the probability, D_i is an indicator function for being born in a non-Anglophone country, and $F(\cdot)$ is the logistic function. As a set of individual characteristics, X_i, dummy variables for age, sex and race are used, although we also try a different set of controls.[16] Following Crump et al. (2009), we estimate equation (5) using a pooled treatment and observational-control sample, and retain in our sample only the observations with $0.1 < p(X_i) < 0.9$ — i.e., the observations with the estimated probability of being treated is more than 0.1 but less than 0.9. Screening the sample in this way ensures that the screened sample contains only the observations that belong to the common support of covariate distributions for the treatment and control groups.[17][18]

Table A1 in the Appendix and Table 1 present a list of countries of birth for the immigrants in our sample and summary statistics, respectively, for Anglophone and non-Anglophone countries by age-at-arrival group. Panel A of Table 1 presents individual characteristics. A key observation is that, for Anglophone immigrants, mean English language skills for early- and late-arrivers are high (close to 3, corresponding to "speak very well") and similar as one would expect. In contrast, for immigrants born

[16] As a robustness check, we try additionally controlling for a measure of cultural distance to the UK, yielding a different sample. See online Appendix A for the results using this different sample specification.

[17] In fact, the restriction, $0.1 < p(X_i) < 0.9$, is stronger than common support restriction. In other words, imposing this restriction more narrowly selects sample than retaining only the observations that belong to the common support of covariate distributions for the treatment and control groups.

[18] The idea of using propensity score as a tool to systematically select a sample before running regressions is suggested by Crump et al. (2009), which is different from using propensity score as a basis for an estimator. An example of another application of this method is Angrist & Pischke (2009) who evaluate a programme to provide work experience, based on the original studies by LaLonde (1986) and Dehejia & Wahba (1999). Comparing the results based on experimental sample, unscreened observational sample, and propensity-score screened sample using the restriction of $0.1 < p(X_i) < 0.9$, Angrist and Pishke illustrate that the propensity-score screened results come very close to the experimental results.

Table 1: Immigrant characteristics and outcomes

	Born in non-Anglophone country			Born in Anglophone country		
	Arrived aged 0 - 8	Arrived aged 9 - 16	Total	Arrived aged 0 - 8	Arrived aged 9 - 16	Total
A. Individual characteristics						
English proficiency, ordinal measure	2.971 (0.200)	2.668 (0.618)	2.801 (0.505)	2.994 (0.098)	2.963 (0.210)	2.979 (0.161)
Linguistic distance	0.926 (0.108)	0.960 (0.079)	0.945 (0.094)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Age	40.049 (16.253)	34.485 (16.181)	36.926 (16.444)	44.352 (14.006)	47.106 (16.703)	45.651 (15.398)
Female	0.513 (0.500)	0.518 (0.500)	0.516 (0.500)	0.514 (0.500)	0.541 (0.498)	0.526 (0.499)
White	0.704 (0.457)	0.477 (0.500)	0.577 (0.494)	0.695 (0.461)	0.380 (0.486)	0.546 (0.498)
Black	0.105 (0.307)	0.224 (0.417)	0.172 (0.377)	0.141 (0.348)	0.348 (0.476)	0.238 (0.426)
Asian	0.130 (0.336)	0.233 (0.423)	0.188 (0.390)	0.134 (0.340)	0.234 (0.423)	0.181 (0.385)
Other single race	0.007 (0.085)	0.014 (0.118)	0.011 (0.105)	0.001 (0.036)	0.004 (0.065)	0.003 (0.052)
Multiracial	0.051 (0.220)	0.044 (0.205)	0.047 (0.212)	0.026 (0.161)	0.032 (0.176)	0.029 (0.168)
B. Residential outcomes						
B1. Enclave residency						
Language enclave	1.923 (4.256)	3.727 (6.447)	2.935 (5.663)	1.264 (3.673)	1.806 (3.925)	1.520 (3.803)
Country-of-birth enclave	3.708 (6.191)	5.202 (6.825)	4.546 (6.596)	2.277 (3.412)	3.867 (5.211)	3.027 (4.426)
Ethnic enclave	2.015 (2.099)	2.546 (2.041)	2.312 (2.083)	1.986 (2.212)	3.023 (2.908)	2.476 (2.616)
B2. Neighbourhood deprivation index (in quintiles, 5 = most deprived)						
Income deprivation	3.033 (1.464)	3.647 (1.340)	3.377 (1.429)	2.820 (1.384)	3.273 (1.385)	3.035 (1.403)
Employment deprivation	2.873 (1.405)	3.360 (1.373)	3.146 (1.408)	2.708 (1.373)	3.069 (1.394)	2.879 (1.394)
Health deprivation	2.757 (1.362)	3.167 (1.356)	2.987 (1.373)	2.671 (1.377)	2.868 (1.369)	2.764 (1.377)

Table 1: Immigrant characteristics and outcomes - continued

	Born in non-Anglophone country			Born in Anglophone country		
	Arrived aged 0 - 8	Arrived aged 9 - 16	Total	Arrived aged 0 - 8	Arrived aged 9 - 16	Total
C. Socioeconomic class (occupation based)						
Higher managerial and professional	0.420	0.252	0.330	0.460	0.354	0.414
	(0.494)	(0.434)	(0.470)	(0.498)	(0.478)	(0.493)
Intermediate	0.142	0.099	0.119	0.151	0.154	0.152
	(0.349)	(0.299)	(0.324)	(0.358)	(0.361)	(0.359)
Self-employed	0.105	0.136	0.122	0.091	0.096	0.093
	(0.307)	(0.343)	(0.327)	(0.288)	(0.295)	(0.291)
Lower supervisory and technical	0.061	0.071	0.066	0.064	0.070	0.066
	(0.239)	(0.258)	(0.249)	(0.244)	(0.255)	(0.249)
Routine and semi-routine	0.185	0.259	0.225	0.190	0.254	0.217
	(0.389)	(0.438)	(0.417)	(0.392)	(0.435)	(0.412)
Never worked and long-term unemployed	0.087	0.182	0.138	0.045	0.073	0.057
	(0.283)	(0.386)	(0.345)	(0.207)	(0.260)	(0.231)

Notes: Standard deviations are reported in parentheses. The sample consists of individuals in the ONS Longitudinal Study dataset aged 20 or over who lived in England and Wales at the time of the 2011 Census, and were born outside the UK who arrived in the UK at age 16 or before. The number of observations varies by panel and column: Panels A and B1 have 1,782; 2,280; 4,062; 3,137; 2,801 and 5,938 observations in the first to sixth columns, respectively, except for ethnic enclave (1,776; 2,261; 4,037; 3,127; 2,796 and 5,923 observations) due to 40 missing values of ethnicity. Panel B2 has 1,751; 2,235; 3,986; 3,066; 2,756 and 5,822 observations. Panel C further restricts the sample to population aged 20 to 60 not in full-time education and has 1,349; 1,540; 2,889; 2,637; 1,968 and 4,605 observations.
Source: Authors' calculations based on the ONS Longitudinal Study.

in non-Anglophone countries, late-arrivers show lower mean English language skills (2.668) than early-arrivers (2.971). This latter group has a proficiency level similar to immigrants born in Anglophone countries. Linguistic distance (from English) is zero for Anglophone countries by construction, and it takes positive values for non-Anglophone countries. Turning to residential outcomes, panel B indicates that late-arrivers born in non-Anglophone countries live in the areas with higher concentrations of people who speak their same native languages and from the countries of birth (panel B1), and in the areas that are relatively more deprived (panel B2). Socioeconomic classes also indicate a different pattern for late-arrivers born in non-Anglophone countries relative to the rest of the groups (panel C). For example, a lower probability of being in the class 'higher managerial and professional' and higher probabilities of being in the classes 'self-employed' and 'never worked and long-term unemployed'.

5. Results

We begin by estimating equation (1) using the OLS estimator.[19] Column (1) of Table 2 reports the OLS estimates of the effect of English proficiency on the socioeconomic outcomes of childhood immigrants in England and Wales, after controlling for individual characteristics and country-of-birth and age-at-arrival fixed effects. The results for enclave residency, reported in panel A, indicate that poor English skills are significantly correlated with residency in language and country-of-birth enclaves, while no significant correlation is found with ethnic-enclave residency. Turning to the impact of language on neighbourhood deprivation, panel B indicates that poorer English skills are significantly associated with living in a neighbourhood with a higher extent of deprivation measured by income, employment and health of residents. Panel C reports results for socioeconomic class. Broadly speaking, we find significant and positive correlations of better English skills with higher socioeconomic classes.

A problem with the OLS estimates of the effects of English proficiency is that they are biased if English proficiency is endogenous. To address this potential endogeneity issue, equation (1) is estimated using the IV estimator, where we use as an instrument for English skills, the *interaction* of the excess age at arrival from age eight with linguistic distance between English and origin-country language (see equation (2)). The first-stage estimates indicate that, for immigrants born in a non-Anglophone country, each year past age eight at arrival significantly decreases their English-proficiency ordinal measure by about 0.05 to 0.06 on average (column (4), Table 2). The magnitude of the coefficient implies that a person's English ordinal measure would be lowered by approximately a half of a unit if the person arrived from a non-Anglophone country at age 16 instead of at age eight. It is important for the identification that our instrument is not weak as a weak instrument is known to bias the IV estimator toward the probability limit of the corresponding OLS estimator. Stock et al. (2002) compute the critical value for the test of weak instruments based on the first-stage F-statistic, and suggest that an F-statistic above roughly 10 makes IV inferences reliable. According to their test for weak instruments, our instrument is not weak as the first-stage F-statistics on the excluded instrument range between 368 and 464.

[19] Our measure of English language skills is an ordinal variable as described in Section 4. In addition to this ordinal measure, we construct a dummy variable for speaking English "very well" to take into account possible non-linear effects of language proficiency. The results using this alternative measure of English language skills presented in online Appendix B are qualitatively similar to our main results.

Table 2: OLS, IV, reduced-form, and first-stage estimates

Dependent variable:	Enclave, deprivation and socioeconomic class			Language
	OLS	IV	Reduced-form	First-stage
	(1)	(2)	(3)	(4)
A. Enclave residency				
Language enclave	-3.044***	-3.829***	0.197**	-0.051***
	(0.676)	(1.405)	(0.076)	(0.002)
Country-of-birth enclave	-1.002***	-0.435	0.022	-0.051***
	(0.364)	(1.569)	(0.081)	(0.002)
Ethnic enclave	-0.195	0.877	-0.045*	-0.051***
	(0.128)	(0.573)	(0.026)	(0.002)
B. Neighbourhood deprivation index (in quintiles, 5 = most deprived)				
Income deprivation	-0.302***	-0.171	0.009	-0.052***
	(0.061)	(0.249)	(0.013)	(0.002)
Employment deprivation	-0.275***	-0.172	0.009	-0.052***
	(0.060)	(0.242)	(0.012)	(0.002)
Health deprivation	-0.185***	-0.198	0.010	-0.052***
	(0.057)	(0.257)	(0.013)	(0.002)
C. Socioeconomic class (occupation based)				
Higher managerial and professional	0.140***	0.146*	-0.008	-0.057***
	(0.023)	(0.083)	(0.005)	(0.003)
Intermediate	0.054***	0.084*	-0.005*	-0.057***
	(0.008)	(0.050)	(0.003)	(0.003)
Self-employment	0.008	-0.123**	0.007**	-0.057***
	(0.009)	(0.053)	(0.003)	(0.003)
Lower supervisory and technical	-0.008	-0.022	0.001	-0.057***
	(0.009)	(0.041)	(0.002)	(0.003)
Routine and semi-routine	-0.073**	0.067	-0.004	-0.057***
	(0.028)	(0.082)	(0.004)	(0.003)
Never worked and long-term unemployed	-0.122***	-0.152***	0.009***	-0.057***
	(0.017)	(0.047)	(0.002)	(0.003)

Notes: Significant at the 1% (***), 5% (**) and 10% (*) levels, respectively. Standard errors are clustered by country of birth. OLS and IV are the estimates of α_1 in equation (1). Reduced-form and first-stage are the estimates of the coefficient on the interaction of age at arrival with linguistic distance between origin-country language and English. Rows in each panel correspond to regressions for different outcomes. Every regression controls for dummies for age, gender and race, and country-of-birth and age-at-arrival fixed effects. See the footnotes of Table 1 for sample sizes of each panel. The first-stage F-statistics on the excluded instrument range between 368 and 464.
Source: Authors' calculations based on the ONS Longitudinal Study.

Column (3) of Table 2 presents the reduced-form estimates of the effects of the instrument on socioeconomic outcomes. Concerning enclave residency, panel A indicates that, for those born in a non-Anglophone country who arrived in the UK after age eight, each additional year that passes before they arrive in the UK is significantly correlated with a higher extent of language residential segregation. The IV estimate in column (2) is in line with this reduced-form estimate: Poorer English proficiency significantly leads immigrants to live in an area with a higher concentration of people who speak their native language. Interestingly, we do not find any significant impact on residency in country-of-birth and ethnic enclaves. This finding is similar to that of Bleakley & Chin (2010), who find weak causal relation between English proficiency and country-of-birth and ethnic-enclave residency using US data. It appears that, conditional on individual characteristics, immigrants with poor English proficiency congregate with their same native speakers but not necessarily with people from the same country of birth or ethnicity. Taking Spanish as an example, our results suggest that a white Spanish person with poor English proficiency congregates with other Spanish speakers but not necessarily with white people (their ethnicity) or people from Spain (their country of birth). Concerning the magnitude of the effect, to facilitate an interpretation of our estimate for the effect on language residential segregation -3.829, consider the following hypothetical situation: Suppose that there is an immigrant born in a Spanish-speaking country who does not speak English well and lives in the local authority with the relative language clustering index of 3.8 (meaning that there are roughly 3.8 times as many Spanish-speaking immigrants in the local authority as one would have expected if the Spanish-speaking population had distributed itself randomly across England and Wales). If this immigrant had spoken English "well" (instead of "not well", corresponding to a one-unit increase in our English skill ordinal measure), she would have lived in a local authority where Spanish-speaking immigrants are neither over- nor under-represented (i.e., local authority with the relative clustering index of one).

Panel B reports neighbourhood deprivation outcomes. IV estimates in column (2) indicate that poor English proficiency leads immigrants to live in the areas where residents are more deprived, but the estimates have high standard errors are insignificant. Turning to socioeconomic-class outcomes in panel C, broadly speaking, the causal estimates in column (2) show that a better English proficiency significantly raises the likelihood of being in higher socioeconomic classes, and reduces that of being in the classes 'self-employed' and 'never worked and long-term unemployed'. It might be the case that immigrants with a better command of English are sorted into occupations where a better proficiency in English improves productivity (e.g., higher managerial and professional occupations) and that those with poorer English skills are sorted into occupations where English skills do not necessarily greatly affect productivity (e.g., self-employment) or sorted out from the labour market. It is also plausible that English skills indirectly affect the socioeconomic class of immigrants via educational achievement, a channel that we will investigate in Section 5.1. The magnitudes of the effects are sizeable: The estimates indicate that, if a migrant arrived in the UK at age 16 instead of age eight, his probability of being in the class 'higher managerial and professional' would be decreased by approximately 19 per cent and that of being in the class 'self-employed' would be increased by roughly 42 per cent on average, evaluated at the mean values for non-Anglophone immigrants.

When comparing OLS and IV estimates, IV estimates tend to be larger in absolute terms for the socioeconomic-class outcomes. It is possible that an omitted variable, such as ability, biases the OLS

estimator upward, but at the same time measurement error possibly correlated with our measure of language proficiency biases the OLS estimator downward. For example, that immigrants surrounded by native English speakers tend to report their proficiency being poor, while those surrounded by other non-Anglophone immigrants may report their proficiency being fluent irrespective of their true English proficiency. In fact, self-reported categorical language measures are found to contain substantial measurement error (Dustmann & van Soest, 2001). If the downward bias caused by measurement error, known as attenuation bias, outweighs the upward bias caused by unobserved characteristics, IV estimates will be larger than OLS estimates, which can help explain the relatively larger IV effects for socioeconomic-class outcomes. For the residential outcomes, there is no clear pattern regarding the relative sizes of the two sets of estimates.

It is also plausible that there is non-classical measurement error (i.e., measurement error that is correlated with the unobserved true English proficiency) in our self-reported measure of English proficiency. For example, it is likely that there is measurement error that tends to be negative for those who are proficient because there is no room to over-report their proficiency at the top end of the categorical measure of English proficiency, and vice versa for those who are not proficient. As a result, if there are many observations at the bounds of our categorical English proficiency measure, the unobserved true English proficiency and measurement error will be negatively correlated. This is an important concern because, in our sample, a significant proportion of individuals report to speak English "very well", which is the top category in our measure of proficiency. Under the presence of non-classical measurement error, the OLS estimator is negatively biased and the IV estimator can be positively biased (Kane et al., 1999),[20] which can also help explain why IV estimates are larger than the OLS estimates for socioeconomic-class outcomes.[21]

5.1. *Role of education*

We have found that English language skills have a significant impact on immigrant residential location outcomes and socioeconomic classes. This subsection investigates a possible channel that drives these results, education. Apart from the direct effects of English proficiency on immigrant outcomes by facilitating communication with native residents and colleagues, English proficiency may also have indirect

[20] When there is non-classical measurement error in a non-binary categorical variable, the direction of the bias in the IV estimator is ambiguous as it depends on the nature of measurement error in the region of language proficiency affected by the instrument (Kane et al., 1999). In case the IV estimator is positively biased, the true effects will lie between OLS and IV estimates. In case the IV estimator is also negatively biased, the findings that both OLS and IV estimates are significant for some of the outcome variables imply that the true effects of English proficiency are possibly larger than those reported in Table 2. It is not possible to further investigate to which direction the IV estimator is biased, because we are not aware of any objective measure of English proficiency that is linked to the ONS Longitudinal Study. However, it is worth noting that the possible presence of non-classical measurement error will not invalidate our findings that language skills significantly affect some of the immigrant outcomes, although it will affect the interpretations of the sizes of the effects.

[21] When English proficiency is measured by a binary indicator, OLS and IV estimates will bound the true effect of English proficiency: Namely, the OLS and IV estimates provide the lower and upper bounds of the true effects, respectively (e.g., Brachet, 2008; Kane et al., 1999). In this regard, our OLS and IV estimates in online Appendix B, based on the model where English proficiency is measured by a binary indicator (that takes the value of one if a person speaks English "very well"), provide the lower and upper bounds of the true effects of English proficiency, respectively.

effects by improving the educational attainments of immigrants (Aoki & Santiago, 2015). To further investigate this channel, in addition to English skills, we control for the measures of education in equation (1). As education is likely to be endogenous in the estimating equation, the estimates of the impact of English skills no longer have causal interpretations. Nevertheless, we present these results in Table 3 to provide suggestive evidence of the role that education plays in explaining the effects of English skills on immigrant outcomes. The first row of each column reports the coefficient estimate on English skills, while the second to fourth rows report estimates on the measures of education. Education is measured by a set of dummy variables that take the value of one if the person has no qualifications, a post-compulsory qualification or an academic degree, respectively, and zero otherwise. The dummy variable for compulsory-level qualification is omitted from equation (1).

The results on enclave residency reported in panel A indicate that, even after controlling for education, the effect of English proficiency on language residential segregation remains significant (column (1)). This could imply that English proficiency has an independent effect on residence in a language enclave because, for example, living close to people who share a common language and culture gives them comfort and/or facilitates an exchange of information. Panel B indicates that the magnitudes of the effects of English skills on neighbourhood deprivation have been greatly diminished (i.e., roughly one tenth of the original magnitudes) after controlling for education. In contrast, education has a significant and non-negligible impact on the deprivation outcomes, suggesting that education may be a more relevant determinant in explaining the residential deprivation outcomes of immigrants. Turning to the socioeconomic-class outcomes in panel C, the impact of language tends to diminish after controlling for education, suggesting that education can be a key channel in explaining the impact of language on socioeconomic classes of immigrants. However, in the case of the class 'never worked and long-term unemployed' in column (6), English skills have a strong effect even after controlling for education, which could suggest that, irrespective of one's educational attainment, English skills have a direct impact on the feasibility to obtain a job.

6. Robustness Checks

We have found that English proficiency significantly impacts the socioeconomic outcomes of UK immigrants. This section addresses the concern that these results are driven by differences in the background characteristics of immigrants from non-Anglophone and Anglophone countries using two strategies. First, we restrict our sample to a set of countries that might be less heterogeneous from each other. Second, we explicitly control for the interactions of age at arrival with origin-country characteristics in our model. This section also addresses another concern that the main results are driven by differences in parental characteristics by controlling for parental education.

An assumption for our IV strategy to estimate the causal effect of English skills is that, aside from the effect of language, immigrants from the two sets of countries are exposed to the same age-at-arrival effects. Under this assumption, the difference in immigrant outcomes between early and late arrivers from non-Anglophone countries in excess of the corresponding difference for immigrants from Anglophone countries can be interpreted as the effect of language. However, one may question the validity of this assumption. It may be the case that, aside from language, Anglophone countries share more commonality with the UK, making it easier for immigrants from these countries to adapt to the new environment upon

Table 3: Effects of English proficiency and education on immigrant outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	A. Enclave residency			*B. Deprivation index (in quintiles)*		
Dependent variable:	Language	Country of birth	Ethnicity	Income	Employment	Health
English skills	-3.834**	-0.399	0.931	-0.017	-0.018	-0.064
	(1.518)	(1.663)	(0.605)	(0.253)	(0.251)	(0.269)
No qualifications	-0.073	0.216	0.224	0.353***	0.334***	0.291***
	(0.460)	(0.469)	(0.165)	(0.064)	(0.065)	(0.069)
Post-compulsory	-0.265**	-0.307**	-0.153**	-0.070**	-0.040	-0.005
	(0.128)	(0.152)	(0.076)	(0.032)	(0.034)	(0.048)
Academic degree	0.109	0.238	0.065	-0.382***	-0.422***	-0.391***
	(0.213)	(0.180)	(0.064)	(0.042)	(0.042)	(0.049)
	C. Socioeconomic class (occupation based)					
Dependent variable:	Higher managerial/ professional	Intermediate	Self-employed	Lower supervisory/ technical	Routine/ semi-routine	Never worked/ long-term unemployed
English skills	0.013	0.069	-0.110*	-0.008	0.161**	-0.125**
	(0.072)	(0.054)	(0.060)	(0.044)	(0.081)	(0.052)
No qualifications	-0.120***	-0.111***	-0.009	-0.005	0.181***	0.065**
	(0.025)	(0.023)	(0.022)	(0.015)	(0.042)	(0.025)
Post-compulsory	0.056***	-0.020	0.012	0.030***	-0.055***	-0.023
	(0.016)	(0.014)	(0.011)	(0.011)	(0.013)	(0.014)
Academic degree	0.386***	-0.052***	-0.069***	-0.079***	-0.159***	-0.027**
	(0.018)	(0.015)	(0.011)	(0.010)	(0.017)	(0.012)

Notes: Significant at the 1% (***), 5% (**) and 10% (*) levels, respectively. Standard errors are clustered by country of birth. The estimates shown in the first row of each panel are the IV estimates of α_1 in equation (1) including all controls described in Table 2 in addition to the following controls for education: dummy variables for having no qualifications, post-compulsory education, and an academic degree, respectively, as the highest level of qualification obtained. Estimates for these controls are shown in the second to fourth rows, respectively, of each panel.
Source: Authors' calculations based on the ONS Longitudinal Study.

arrival in the UK. So far, to address this type of concerns, we compute propensity score for being born in a non-Anglophone country and use it as the base for systematic sample selection to make immigrants in our sample less heterogeneous from each other (see Section 4.2). In addition to pre-screening the sample using the propensity score, we have been controlling for country-of-origin fixed effects in every regression, which account for systematic differences in origin-country characteristics that do not vary over age at arrival. Nevertheless, one might still be concerned that the effects of country-of-birth specific characteristics could vary by age at arrival. To address this concern, we further restrict our sample and retain only immigrants from countries which may be less heterogeneous from each other. Unavoidably we must pay the cost of losing more observations, although the more we restrict our sample, the less heterogeneous the sample becomes.

First, immigrants from Europe might find it easier to adapt to the UK environment because European countries share commonality with the UK in culture and institutions because of a long history of interactions across European countries. Similarly, immigrants from Commonwealth countries might find it easier to adapt to the UK because of, for example, a similarity in their legal systems and culture. In an attempt to make immigrants in our sample less heterogeneous, we exclude the immigrants born in these countries that have special ties with the UK. The results excluding European and Commonwealth countries are reported in columns (2) and (3) of Table 4, respectively, while column (1) presents the base results from Table 2 for comparison. The results are broadly similar when Commonwealth countries are omitted. However, interestingly, when immigrants from European countries are omitted, we observe a different pattern: Namely, better English skills *positively* impact ethnic residential segregation. It could be the case that the non-European migrants, whose customs set them apart from the majority population in the UK, when they speak English well and thereby having a wider option regarding residential location, choose to live in an ethnic enclave as a means of enhancing their economic, social, and cultural developments (Marcuse, 1997). For example, immigrants from Somalia and Turkey — which are the top non-Anglophone source countries of non-European migrants in our sample — who speak English well and have a wide option for residential locations may decide to live in ethnic enclaves because they know Somali or Turkish culture well, and therefore are able to obtain higher returns to their human capitals by living in their ethnic enclaves. It could also be the case that because non-European migrants have different culture from the mainstream culture in the UK, they face unfavorable treatment when they live outside their ethnic enclaves. As a result, those who speak English well and have a choice of residential location may choose to live in their ethnic enclaves to avoid encountering such unfavorable treatment.

Next, we divide our sample by origin-country per capita gross domestic product (GDP) to make the sample more homogeneous regarding origin-country income level, which would capture various aspects of origin-country characteristics that can affect immigrant outcomes. For example, self-employment rates tend to be higher in lower income countries. Immigrants born in these countries might be more likely to be self-employed, and this tendency may magnify as age at arrival increases because late arrivers are likely to be more affected by their origin-country characteristics. The results for below- and above-median GDP countries are reported in columns (4) and (5) of Table 4, respectively, which exhibit several interesting differences. First, we observe heterogeneity in the effects of language on residential segregation. Precisely, English proficiency appears to negatively impact residency in a language enclave and country-of-birth enclave for immigrants from high-income countries (column (5)), while language pro-

Table 4: IV effect of English proficiency using alternative samples

	All	Drop Europe	Drop common -wealth	Low GDP countries	High GDP countries
	(1)	(2)	(3)	(4)	(5)
A. Enclave residency					
Language enclave	-3.829***	-2.581*	-5.091***	-1.635	-6.505***
	(1.405)	(1.375)	(1.175)	(1.408)	(1.491)
Country-of-birth enclave	-0.435	1.227	-2.068	2.173	-3.400*
	(1.569)	(1.354)	(1.451)	(1.491)	(1.835)
Ethnic enclave	0.877	1.546**	0.371	1.576**	0.279
	(0.573)	(0.716)	(0.567)	(0.778)	(0.557)
B. Neighbourhood deprivation index (in quintiles, 5 = most deprived)					
Income deprivation	-0.171	-0.173	0.012	-0.288	-0.066
	(0.249)	(0.267)	(0.245)	(0.307)	(0.343)
Employment deprivation	-0.172	-0.336	-0.184	-0.410	-0.062
	(0.242)	(0.294)	(0.214)	(0.353)	(0.300)
Health deprivation	-0.198	-0.432	-0.187	-0.586	0.112
	(0.257)	(0.310)	(0.222)	(0.365)	(0.287)
C. Socioeconomic class (occupation based)					
Higher managerial and professional	0.146*	0.120	0.288***	0.128	0.206**
	(0.083)	(0.102)	(0.075)	(0.087)	(0.096)
Intermediate	0.084*	0.090	0.108	0.017	0.196**
	(0.050)	(0.058)	(0.066)	(0.054)	(0.088)
Self-employed	-0.123**	-0.094*	-0.143***	-0.047	-0.203**
	(0.053)	(0.053)	(0.053)	(0.059)	(0.086)
Lower supervisory and technical	-0.022	-0.055	-0.021	-0.078*	0.053
	(0.041)	(0.042)	(0.051)	(0.045)	(0.064)
Routine and semi-routine	0.067	0.097	-0.096	0.159	-0.130
	(0.082)	(0.096)	(0.067)	(0.101)	(0.088)
Never worked and long-term unemployed	-0.152***	-0.158***	-0.135***	-0.179**	-0.121**
	(0.047)	(0.053)	(0.050)	(0.089)	(0.059)

Notes: Significant at the 1% (***), 5% (**) and 10% (*) levels, respectively. Standard errors are clustered by country of birth. The estimates shown are the IV estimates of α_1 in equation (1). Each column corresponds to a different sample specification: Full sample (column 1), sample excluding European migrants (column 2), sample excluding Commonwealth migrants (column 3); and migrants born in countries with below- and above-median GDP (columns 4 and 5, respectively).
Source: Authors' calculations based on the ONS Longitudinal Study.

ficiency positively impacts residency in an ethnic enclave for those from low-income countries (column (4)). This latter result is very similar to what we found in column (2) and might suggest that immigrants from low-income countries obtain a higher return to their human capital or can avoid facing unfavourable treatment by living in an ethnic enclave for the reasons that have been previously discussed for the case where European migrants were omitted from the sample. Second, it appears that the effects of language on higher socioeconomic-class outcomes and self-employment are driven by immigrants born in high-income countries. A possible interpretation is that high-income countries have similar workplace culture to the mainstream culture to the UK, allowing immigrants from these countries to have better job opportunities if they speak English well. Another possible interpretation is that, because immigrants born in low-income countries are financially constrained and are unable to make necessary investment to obtain well-paying jobs in the UK, whether they speak English fluently or not may make less of a difference in obtaining well-paying jobs.

We now take a different approach to address the concern that the main results are driven by different background characteristics of immigrants from Anglophone and non-Anglophone countries: Namely, we control for the interactions of age at arrival with cultural distance between the UK and origin country. It may be the case that, apart from the effect of language, immigrants born in a country that is culturally more distant to the UK find it more difficult to adapt to the new UK environment, and that this adverse effect becomes more severe as age at arrival gets older. Although this is highly plausible, a challenge it to quantify cultural distance because the concept of culture is not quantitative in nature. To address this challenge, we rely on a measure of genetic distance between the origin-country population and UK population as a summary measure of cultural distance, obtained from Spolaore & Wacziarg (2016). They argue that genetic distance, measuring the closeness of populations in terms of genes, reflects time since the populations shared the same ancestors. Over time, the ancestors transmit to their descendants not only their biological traits (i.e., genes) but also their cultural traits, such as habits and values, and this transmission occurs with variation. Populations that are genetically far from each other had more time to diverge in terms of cultural traits, and this divergence can in turn create barriers to human interactions. Spolaore & Wacziarg (2016) document that genetic distance of populations is significantly positively correlated with a wide array of measures of cultural differences. Results are summarised in Table 5, where base results are copied to column (1) for comparison and column (2) controls for the interaction of age at arrival with cultural distance between the UK and origin country. The results are not sensitive to the inclusion of this additional control.

In a similar vein, in columns (3) and (4), we control for the interactions of origin-country labour market characteristics with age at arrival. For example, if non-Anglophone countries have a higher (or lower) labour force participation rate on average *and* the effects of origin-country labour force participation rate vary by age at arrival, our instrument may capture the compound effects of English skills and differential average labour force participation behaviour of the origin country. To address this type of concerns, column (3) controls for the interaction with the origin-country labour force participation rate, while column (4) controls for the interaction with the origin-country employment ratio, defined as the proportion of a country's population aged 15 or over that is employed.[22] The two columns indicate that the results

[22]Origin-country labour market variables are modeled International Labour Organisation estimates, and the values for 1990 and 1991 are used for the labour force participation rate and employment ratio, respectively, which are the

Table 5: IV effect of English proficiency using alternative regression specifications

		Additional control variables:			
	Base results (1)	Cultural distance × arrival age (2)	Labour force participation × arrival age (3)	Employment ratio × arrival age (4)	Parental education (5)
A. Enclave residency					
Language enclave	-3.829***	-4.112***	-4.184***	-3.997***	-3.112
	(1.405)	(1.352)	(1.343)	(1.372)	(2.245)
Country-of-birth enclave	-0.435	-0.659	-0.724	-0.796	0.868
	(1.569)	(1.468)	(1.462)	(1.538)	(2.265)
Ethnic enclave	0.877	1.101*	0.962*	0.978*	1.310
	(0.573)	(0.568)	(0.533)	(0.579)	(0.815)
B. Neighbourhood deprivation index (in quintiles, 5 = most deprived)					
Income deprivation	-0.171	-0.160	-0.235	-0.240	-0.347
	(0.249)	(0.209)	(0.239)	(0.254)	(0.327)
Employment deprivation	-0.172	-0.193	-0.234	-0.220	-0.252
	(0.242)	(0.213)	(0.218)	(0.239)	(0.382)
Health deprivation	-0.198	-0.237	-0.264	-0.233	-0.420
	(0.257)	(0.229)	(0.227)	(0.251)	(0.356)
C. Socioeconomic class (occupation based)					
Higher managerial and professional	0.146*	0.104	0.127	0.121	0.227*
	(0.083)	(0.090)	(0.078)	(0.082)	(0.126)
Intermediate	0.084*	0.092*	0.077	0.075	0.098
	(0.050)	(0.052)	(0.049)	(0.048)	(0.077)
Self-employed	-0.123**	-0.115**	-0.100**	-0.106**	-0.130
	(0.053)	(0.051)	(0.044)	(0.048)	(0.089)
Lower supervisory and technical	-0.022	0.003	-0.017	-0.017	-0.020
	(0.041)	(0.042)	(0.040)	(0.041)	(0.066)
Routine and semi-routine	0.067	0.061	0.061	0.079	0.021
	(0.082)	(0.082)	(0.078)	(0.078)	(0.135)
Never worked	-0.152***	-0.145***	-0.148***	-0.151***	-0.197**
and long-term unemployed	(0.047)	(0.046)	(0.043)	(0.045)	(0.087)

Notes: Significant at the 1% (***), 5% (**) and 10% (*) levels, respectively. Standard errors are clustered by country of birth. Column 1 controls for the variables specified in Table 2, while columns 2 to 5 control for an additional variable each. These additional controls are an interaction of age at arrival with cultural distance from the UK (column 2), an interaction with origin-country labour force participation rate (column 3), an interaction with origin-country employment ratio (column 4), and parental education (column 5). Sample size in column 5 is smaller than the base sample due to missing values of parental education: Panel A has 6,156 observations except for ethnic enclave residency, 6,129; panel B has 6,030; and panel C has 5,134.
Source: Authors' calculations based on the ONS Longitudinal Study.

are broadly similar after the inclusion of these additional variables.

Finally, we address another concern that can give an alternative explanation to our findings. Namely, parental characteristics of immigrants from the two sets of countries might be different, and parents with different characteristics might have made different decisions regarding the timing of migration to the UK. For example, parents from non-Anglophone countries might have recognised a possible barrier that their children would face if they migrate when their children are older, and may have chosen to migrate when their children were younger. At the same time, these parents might be different from the parents of immigrants from Anglophone countries in a way that can affect the future socioeconomic outcomes of childhood immigrants. If this is the case, our IV estimates may reflect not only the effects of English skills but also the effects of differential parental characteristics. To address this type of concerns, we control for parental education, measured by the dummy variable that takes the value of one if any of the two parents of the migrants has college education or above, and zero otherwise. We are not aware of any other studies that have accounted for parental education when analysing the causal impact of language proficiency on residential outcomes of immigrants. A limitation of this exercise is that, due to missing information on parental education, sample sizes decrease by roughly 30 to 40 per cent. Despite this limitation, we control for this potentially important confounding factor in column (5) of Table 5. A comparison of the IV estimates in columns (1) and (5) without and with parental education as a control, respectively, indicates that the two sets of results are broadly similar. A point to note is that, not surprisingly, standard errors are higher in the smaller samples in column (5) and the effects on language residential segregation and being in the class 'self-employed' are now insignificant. To investigate whether these differences in estimation results are driven by changes in sample sizes or the inclusion of parental education, we estimate the model with the smaller samples used in column (5) *without* controlling for parental education. Results (not reported) are very similar to those in column (5), implying that differences in the results are likely driven by a change in sample sizes.

7. Conclusion

Inflows of migrants have increased in the OECD and EU countries over the past two decades (OECD/EU, 2015), and the social integration of immigrants is becoming an increasingly important policy objective. Although it is widely believed that language proficiency in the language spoken in the host country is one of the important factors for promoting integration, there is limited knowledge on the causal impact of language on residential location outcomes and the socioeconomic class of immigrants in the host country. Our paper contributes to this knowledge by studying the impact on a variety of residential outcomes and the socioeconomic class of immigrants. Specifically, we construct the measures of three different types of residential segregation — language, ethnic, and country-of-birth residential segregation — and analyse the impact of English skills on these different types of residential segregation. We also analyse the impact of English skills on the extent of deprivation in the neighbourhood immigrants live in, using a unique dataset from the 2011 Census of England and Wales that is linked to the English Deprivation Indices.

To overcome a possible endogeneity issue of English skills, we rely on an IV strategy, where age

earliest years data are available for the respective variables.

at arrival in the UK is used to construct an instrument. The idea of using age at arrival is based on the critical period hypothesis of language acquisition (Lenneberg, 1967), documenting that people exposed to a new language within the critical period of language acquisition (i.e., childhood) learn it more easily relative to those exposed outside of this critical period. The hypothesis implies that immigrants born in a non-Anglophone country who arrived in the UK at a younger age would on average have better English language skills than late arrivers. Moreover, among non-Anglophone immigrants, there is variation in how close their native languages are to English, and it is plausible that immigrants whose mother tongue is linguistically closer to English (e.g., Dutch) find it easier to learn English. Following Clarke & Isphording (2017), we incorporate this idea into our analysis by using an interaction of age at arrival with linguistic distance between English and origin-country language as our instrument for English proficiency.

Our results suggest that poorer English skills significantly lead immigrants to cluster in areas where there are more individuals that speak their own native language, but not necessarily in areas where there are higher concentrations of individuals from their same ethnicity or country-of-birth. Our findings imply that, for example, a Spanish-speaking white immigrant born in Spain who does not speak English well tend to live in areas with high concentrations of Spanish speakers but not necessarily in areas with high concentrations of Spanish (their country of birth) or other white people (their ethnicity). Our results also suggest that poorer English skills lead immigrants to live in relatively more deprived neighbourhoods, but these effects are not significant and the evidence is not strong. Turning to the socioeconomic class achieved by childhood immigrants when they are adults, we find a significant impact of English skills. In particular, the negative impact of better English skills on the probability of being in the classes 'self-employed' and 'never worked and long-term unemployed' are robust and sizeable. We investigate the role of education as a potential mechanism driving these effects, and find suggestive evidence that a higher educational attainment as a result of better English skills is likely to be an important factor that drives the impact on socioeconomic-class outcomes. However, the impact on language residential segregation remains largely unchanged after accounting for education, which could imply that English proficiency has independent effects on this outcome.

The results based on our IV strategy suggest that some immigrants may be less integrated than they would like to because they arrived in the UK after the critical period of language acquisition and they do not speak English fluently enough to interact well with natives, providing evidence against the idea that a lack of integration of immigrants is merely due to their preference. An implication based on our findings is that providing specific English language courses for immigrants could be an effective policy to foster the residential and labour market integration of immigrants, since improving their English language skills makes them less likely to be in the class 'never worked and long-term unemployed', and also promotes them to live in less segregated areas with lower concentrations of people speaking their own native language. In particular, providing support to learn English to immigrants from non-Anglophone countries who arrived in the UK after age eight would be beneficial because individuals who arrived in the UK at age eight or before appear to catch up with the level of proficiency of Anglophone immigrants by the time they become adults. It is also likely to be an efficient use of resources to target younger immigrants, among those who arrived after age eight, because the earlier immigrants are exposed to English, the easier it is for them to learn the language.

Appendix

Table A1: Immigrants by country of birth

A. Anglophone countries *A-1. Arrived aged 0 - 8*	*N*	*%*	*B. Non-Anglophone countries* *B-1. Arrived aged 0 - 8*	*N*	*%*
Ireland	555	17.7	Cyprus	228	12.8
Kenya	341	10.9	Italy	121	6.8
United States	235	7.5	Somalia	118	6.6
South Africa	233	7.4	Turkey	82	4.6
Canada	222	7.1	France	78	4.4
Australia	204	6.5	Malaysia	77	4.3
Singapore	186	5.9	Germany	61	3.4
Jamaica	164	5.2	Netherlands	51	2.9
Malta	150	4.8	Vietnam	48	2.7
Uganda	147	4.7	Egypt	46	2.6
Nigeria	117	3.7	Portugal	43	2.4
Zambia	76	2.4	Spain	43	2.4
Zimbabwe	67	2.1	Belgium	41	2.3
New Zealand	60	1.9	Malawi	33	1.9
Gibraltar	51	1.6	Iran	31	1.7
Ghana	47	1.5	China	30	1.7
Guyana	34	1.1	Afghanistan	30	1.7
Isle of Man	30	1.0	Saudi Arabia	29	1.6
Trinidad and Tobago	26	0.8	Poland	27	1.5
Mauritius	22	0.7	Switzerland	25	1.4
Total top 20	2,967	94.6	Total top 20	1,242	69.7
A-2. Arrived aged 9 - 16	*N*	*%*	*B-2. Arrived aged 9 - 16*	*N*	*%*
Ireland	535	19.1	Somalia	343	15.0
Kenya	493	17.6	Cyprus	167	7.3
Jamaica	347	12.4	Turkey	161	7.1
Uganda	217	7.7	Poland	119	5.2
Nigeria	193	6.9	Afghanistan	114	5.0
South Africa	154	5.5	China	88	3.9
Zimbabwe	119	4.2	Vietnam	88	3.9
Ghana	107	3.8	Portugal	77	3.4
United States	64	2.3	Italy	65	2.9
Guyana	55	2.0	Kosovo	57	2.5
Canada	44	1.6	Germany	55	2.4
Australia	40	1.4	France	46	2.0
Singapore	39	1.4	Malaysia	44	1.9
Sierra Leone	38	1.4	Iran	41	1.8
Zambia	31	1.1	Malawi	39	1.7
Trinidad and Tobago	29	1.0	Lithuania	28	1.2
St Lucia	26	0.9	Congo (Democratic Republic)	28	1.2
New Zealand	23	0.8	Russia	28	1.2
Mauritius	23	0.8	Ethiopia	28	1.2
Barbados	22	0.8	Angola	25	1.1
Total top 20	2,599	92.8	Total top 20	1,641	72.0

Notes: Panels A and B present Anglophone and non-Anglophone countries, respectively. *N* refers to the number of individuals by country of birth for the top 20 countries present in our sample for those who arrived in the UK between age 0 and 8 (upper panels) and between 9 and 16 (lower panels).
Source: Authors' calculations based on the ONS Longitudinal Study.

References

Aldashev, A., Gernandt, J., & Thomsen, S. L. (2009). Language usage, participation, employment and earnings. *Labour Economics*, *16*(3), 330 – 341.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Aoki, Y., & Santiago, L. (2015). Education, health and fertility of UK immigrants: The role of English language skills. IZA Discussion Paper 9498, Institute for the Study of Labour.

Bauer, T., Epstein, G., & Gang, I. (2005). Enclaves, language, and the location choice of migrants. *Journal of Population Economics*, *18*(4), 649–662.

Beckhusen, J., Florax, R., Graaff, T., Poot, J., & Waldorf, B. (2013). Living and working in ethnic enclaves: Language proficiency of immigrants in u.s. metropolitan areas. *Papers in Regional Science*, *92*(2), 305–328.

Bertrand, M., Luttmer, E. F. P., & Mullainathan, S. (2000). Network effects and welfare cultures. *The Quarterly Journal of Economics*, *115*(3), 1019–1055.

Bleakley, H., & Chin, A. (2004). Language skills and earnings: Evidence from childhood immigrants. *The Review of Economics and Statistics*, *86*(2), 481–496.

Bleakley, H., & Chin, A. (2010). Age at arrival, english proficiency, and social assimilation among US immigrants. *American Economic Journal: Applied Economics*, *2*(1), 165–92.

Borjas, G. J. (2000). Ethnic enclaves and assimilation. *Swedish Economic Policy Review*, *7*, 89–122.

Brachet, T. (2008). Maternal smoking, misclassification, and infant health. MPRA Paper 21466, University Library of Munich, Germany.

Chiswick, B. R., & Miller, P. W. (1995). The endogeneity between language and earnings: International analyses. *Journal of Labor Economics*, *13*(2), 246–288.

Chiswick, B. R., & Miller, P. W. (2005). Do enclaves matter in immigrant adjustment? *City & Community*, *4*(1), 5–35.

Chiswick, B. R., & Miller, P. W. (2009). Earnings and occupational attainment among immigrants. *Industrial Relations: A Journal of Economy and Society*, *48*(3), 454–465.

Clarke, A., & Isphording, I. E. (2017). Language barriers and immigrant health. *Health Economics*, *26*(6), 765–778.

Clausen, J., Heinesen, E., Hummelgaard, H., Husted, L., & Rosholm, M. (2009). The effect of integration policies on the time until regular employment of newly arrived immigrants: Evidence from Denmark. *Labour Economics*, *16*(4), 409–417.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Dustmann, C. (1994). Speaking fluency, writing fluency and earnings of migrants. *Journal of Population Economics*, *7*(2), 133–156.

Dustmann, C., & Fabbri, F. (2003). Language proficiency and labour market performance of immigrants in the UK. *Economic Journal*, *113*(489), 695–717.

Dustmann, C., & van Soest, A. (2001). Language fluency and earnings: Estimation with misclassified language indicators. *The Review of Economics and Statistics*, *83*(4), 663–674.

Edin, P.-A., Fredriksson, P., & Aslund, O. (2003). Ethnic enclaves and the economic success of immigrants: Evidence from a natural experiment. *The Quarterly Journal of Economics*, *118*(1), 329–357.

Goldthorpe, J. H. (2007). Social class and the differentiation of employment contracts. In *On Sociology*, vol. 2, chap. 5, (pp. 101–24). Stanford: Stanford University Press, 2 ed.

Gonzalez, L. (2005). Nonparametric bounds on the returns to language skills. *Journal of Applied Econometrics*, *20*(6), 771–795.

Iceland, J., & Scopilliti, M. (2008). Immigrant residential segregation in U.S. metropolitan areas, 1990-2000. *Demography*, *45*(1), 79–94.

Isphording, I. E., & Otten, S. (2014). Linguistic barriers in the destination language acquisition of immigrants. *Journal of Economic Behavior & Organization*, *105*, 30 – 50.

Janssen, S. (2010). *The world almanac and book of facts 2011*. New York: Infobase Learning.

Kane, T. J., Rouse, C. E., & Staiger, D. (1999). Estimating returns to schooling when schooling is misreported. Working Paper 7235, National Bureau of Economic Research.

Kossoudji, S. A. (1988). English language ability and the labor market opportunities of Hispanic and East Asian immigrant men. *Journal of Labor Economics*, *6*(2), 205–228.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, *76*(4), 604–620.

Lazear, E. P. (1999). Culture and language. *Journal of Political Economy*, *107*(S6), S95–S126.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Logan, J. R., Zhang, W., & Alba, R. D. (2002). Immigrant enclaves and ethnic communities in New York and Los Angeles. *American Sociological Review*, *67*, 299–322.

Marcuse, P. (1997). The enclave, the citadel, and the ghetto: What has changed in the post-fordist U.S. city. *Urban Affairs Review*, *33*(2), 228–264.

Massey, D. S. (1985). Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research*, *69*, 315–350.

Miller, P. W., & Neo, L. M. (1997). Immigrant unemployment: The Australian experience. *International Migration*, *35*(2), 155–185.

Ministry of Housing, Communities and Local Government (2015). English indices of deprivation. Accessed on 4 July 2016.
URL `https://www.gov.uk/government/collections/english-indices-of-deprivation`

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

OECD\European Union (2015). *Indicators of immigrant integration 2015: Settling in*. Paris: OECD Publishing.

Office for National Statistics (2010). *The Standard Occupational Classification 2010 Vol 3: The National Statistics Socio-economic Classification*. Basingstoke: Palgrave Macmillan UK.

Shields, M. A., & Price, S. W. (2002). The English language fluency and occupational success of ethnic minority immigrant men living in English metropolitan areas. *Journal of Population Economics*, *15*(1), 137–160.

Spolaore, E., & Wacziarg, R. (2016). Ancestry and development: New evidence. Discussion Papers Series, Department of Economics, Tufts University 0820, Department of Economics, Tufts University.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518–29.

Online Appendix A: Alternative Ppropensity-score screened sample

Dependent variable:	Enclave, deprivation, socioeconomic class			Language
	OLS	IV	Reduced-form	First-stage
	(1)	(2)	(3)	(4)
A. Enclave residency				
Language enclave	-3.140***	-4.394***	0.227***	-0.052***
	(0.718)	(1.340)	(0.074)	(0.002)
Country-of-birth enclave	-1.084***	-1.300	0.067	-0.052***
	(0.378)	(1.558)	(0.082)	(0.002)
Ethnic enclave	-0.200	0.700	-0.036	-0.051***
	(0.127)	(0.571)	(0.027)	(0.002)
B. Neighbourhood deprivation index (in quintiles, 5 = most deprived)				
Income deprivation	-0.285***	-0.284	0.015	-0.052***
	(0.070)	(0.272)	(0.014)	(0.002)
Employment deprivation	-0.255***	-0.242	0.013	-0.052***
	(0.068)	(0.261)	(0.013)	(0.002)
Health deprivation	-0.158**	-0.256	0.013	-0.052***
	(0.064)	(0.277)	(0.014)	(0.002)
C. Socioeconomic class (occupation based)				
Higher managerial and professional	0.143***	0.178*	-0.010*	-0.056***
	(0.026)	(0.092)	(0.006)	(0.003)
Intermediate	0.051***	0.055	-0.003	-0.056***
	(0.009)	(0.050)	(0.003)	(0.003)
Self-employed	0.003	-0.114**	0.006*	-0.056***
	(0.009)	(0.057)	(0.003)	(0.003)
Lower supervisory and technical	-0.004	-0.022	0.001	-0.056***
	(0.010)	(0.045)	(0.003)	(0.003)
Routine and semi-routine	-0.080**	0.051	-0.003	-0.056***
	(0.032)	(0.083)	(0.005)	(0.003)
Never worked	-0.112***	-0.148***	0.008***	-0.056***
and long-term unemployed	(0.017)	(0.046)	(0.002)	(0.003)

Notes: *** $p<$ .01, ** $p<$ .05, and * $p<$ .10. Standard errors are clustered by country of birth. OLS and IV are the estimates of α_1 in equation (1). First-stage and reduced-form are the estimates of the coefficients on the instrument, which is an interaction of age at arrival with linguistic distance. Rows in each panel correspond to regressions for different outcomes. Refer to Table 2 for the controls included. Sample is screened on propensity score $p(X_i)$ defined in equation (5), where X_i correspond to a measure of cultural distance (cf., Section 6), in addition to dummy variables for age, sex and race. Sample size varies by outcome: Panel A has 9,038 observations except for ethnic enclave residency, 8,999; panel B has 8868; and panel C has 6,881. The first-stage F-statistics on the excluded instrument range between 352 and 438.
Source: Authors' calculations based on the ONS Longitudinal Study.

Online Appendix B: Alternative measure of English skills

Dependent variable:	Enclave, deprivation, socioeconomic class			Dummy for English ability
	OLS	IV	Reduced-form	First-stage
	(1)	(2)	(3)	(4)
A. Enclave residency				
Language enclave	-4.519***	-5.016***	0.197**	-0.039***
	(0.862)	(1.788)	(0.076)	(0.002)
Country-of-birth enclave	-1.390***	-0.569	0.022	-0.039***
	(0.492)	(2.053)	(0.081)	(0.002)
Ethnic enclave	-0.282	1.153	-0.045*	-0.039***
	(0.180)	(0.747)	(0.026)	(0.002)
B. Neighbourhood deprivation index (in quintiles, 5 = most deprived)				
Income deprivation	-0.421***	-0.225	0.009	-0.039***
	(0.084)	(0.326)	(0.013)	(0.002)
Employment deprivation	-0.377***	-0.226	0.009	-0.039***
	(0.077)	(0.319)	(0.012)	(0.002)
health deprivation	-0.274***	-0.261	0.010	-0.039***
	(0.077)	(0.338)	(0.013)	(0.006)
C. Socioeconomic class (occupation based)				
Higher managerial and professional	0.210***	0.199*	-0.008	-0.042***
	(0.031)	(0.113)	(0.005)	(0.002)
Intermediate	0.075***	0.115*	-0.005*	-0.042***
	(0.013)	(0.067)	(0.003)	(0.002)
Self-employment	-0.004	-0.168**	0.007**	-0.042***
	(0.013)	(0.071)	(0.003)	(0.002)
Lower supervisory and technical	-0.019	-0.030	0.001	-0.042***
	(0.015)	(0.056)	(0.002)	(0.002)
Routine and semi-routine	-0.130***	0.091	-0.004	-0.042***
	(0.042)	(0.111)	(0.004)	(0.002)
Never worked and long-term unemployed	-0.132***	-0.207***	0.009***	-0.042***
	(0.021)	(0.061)	(0.002)	(0.002)

Notes: *** $p<$ .01, ** $p<$ .05, and * $p<$ .10. Standard errors are clustered by country of birth. OLS and IV are the estimates of α_1 in equation (1), where an indicator for speaking English "very well" is used as a measure of English skills. First-stage and reduced-form are the estimates of the coefficients on the instrument, which is an interaction of age at arrival and linguistic distance. Refer to Table 2 for sample sizes and controls included. The F-statistics on the excluded instrument in column 4 range from 387 to 516.
Source: Authors' calculations based on the ONS Longitudinal Study.