Immigration and Invention: Does Language Matter?

Kirk Doran, Chungeun Yoon*

ABSTRACT

Economists have long noted that linguistically diverse immigrant flows might have a particularly large impact on innovation and creativity, through the introduction and combination of new perspectives, information, and habits (Alesina and La Ferrara, 2005). On the other hand, if innovation depends on communication, and communication depends on a common language, then linguistically uniform immigration flows may have the largest impact on innovation. In this paper, we make use of features of the 1920s U.S. immigration quotas that caused some of the "missing immigrants" to be absent from cities which had many residents who happened to speak their language, while other "missing immigrants" were absent from cities which had few residents who spoke their language. The resulting changes in innovation are consistent with a U-shaped curve for the effect of linguistic diversity on the innovativeness of a society. Too much linguistic diversity creates a "tower of babel" effect, in which people have unique things to talk about but no common language to say them in. Too little linguistic diversity creates a homogeneous population, in which people have a common language but nothing unique to share. The optimal amount of linguistic diversity for a creative society appears to be somewhere in between.

JEL classification: J61, N32, O31, Z13

*University of Notre Dame. We thank Ina Ganguli, Megan MacGarvie, and especially Shu Kahn for very helpful comments on our preliminary draft. All errors are our own.

1 Introduction

Economists have long noted several ways that immigration could affect innovation. Highly-skilled immigrants may innovate directly, while low-skilled immigrants could affect the scale of production, thereby encouraging labor-complementary inventions, and discouraging strongly-labor-saving inventions (Acemoglu, 2010). But the literature on immigration and innovation has failed to address the potential importance of one of the most obvious differences between immigrants and natives: language differences. On the one hand, immigrants may have a larger impact on innovation when there is a language similarity between the immigrants and natives. Strongly labor-complementary inventions may be incentivized more by a large homogeneous workforce that can easily work together rather than by heterogeneous labor inputs that have trouble communicating with each other. On the other hand, immigration may have a larger impact on innovation when there is a language dissimilarity between the immigrants and natives. After all, a large literature explores the possibility that a diverse ethnolinguistic mix "brings about variety in abilities, experiences, and cultures that may be productive and may lead to innovation and creativity" (Alesina and La Ferrara, 2005).

Both of the above effects of lanugage on innovation are plausible, so it is helpful to consider them in light of theoretical models of what innovation actually is. According to many theories, innovation involves making new combinations from existing ideas and experiences (Weitzman, 1998). An innovative society will thus *necessarily* involve people with diverse ideas communicating with each other to facilitate new combinations (innovations). Thus, as the number of individuals increases, innovation may increase if: (a) individuals communicate more and more; and (b) they have a larger and larger number of unique things to communicate about.

Linguistic diversity of immigrants relative to the preexisting population can affect both of these channels, but in opposite directions. Linguistic homogeneity increases the likelihood of (a) and decreases the likelihood of (b). Linguistic diversity decreases the likelihood of (a) and increases the likelihood of (b). It is therefore plausible that the optimal amount of immigrant linguistic diversity for building an innovative society will be somewhere in

between complete linguistic homogeneity and complete linguistic diversity. The results in this paper are consistent with this hypothesis.

We bring empirical evidence to bear on this question through analyzing a setting in which the language of immigrants varies independently of the language mix of the people already living in the locations that the immigrants are immigrating to. This is difficult, because shift-share style immigration instruments build on exactly that variation in immigration that is correlated with ethnolinguistic variation in the pre-existing population across locations. In this paper, we make use of immediate-onset 1920s U.S. immigration quotas that suddenly ended ongoing immigrant flows to some cities but not others. Crucially, the quotas caused cities with a circa-1920 immigrant inflow from quota-affected countries to suddenly stop receiving as many immigrants; not every such city had a native-born population descended from long-past immigrant inflows that had kept speaking their family's language. As a result, among the quota-affected cities, some lost immigrants who spoke a common language among the pre-existing population (because previous generations of immigrants had preserved their language across generations), while others lost immigrants who spoke an uncommon language among the population. These "off-diagonal" terms allow us to estimate the effects of immigrants on innovation in a city both when the immigrants speak a common language in the city and when they do not.

The results are striking. We find that native-born inventors whose cities lost immigrants who spoke uncommon languages apply for no fewer, and possibly more, patents after the quotas. Native-born inventors whose cities lost immigrants who spoke very common languages applied for somewhat fewer patents after the quotas. But native-born inventors whose cities lost immigrants with moderate linguistic diversity applied for many fewer patents after the quotas.

These results are thus consistent with a U-shaped curve for the effect of linguistic diversity on the innovativeness of a society. Too much linguistic diversity creates a "tower of babel" effect (Ballatore, Fort, Ichino, 2018) in which people have unique things to talk about but no common language to say them in. Too little linguistic diversity creates a homogeneous population (Alesina and La Ferrara, 2005) in which people have a common language but nothing unique to share. The optimal amount of linguistic diversity for a creative society

appears to be somewhere in between.

It is important to note that, as (Doran and Yoon, 2018) explains, the effect of these low-skilled immigrants on native inventors is acting through a change in the scale of production that incentivizes strongly labor complementary inventions (Acemoglu, 2010). The role of communication, therefore, is happening in the context of a low-skilled workforce, not in the context of highly skilled innovators themselves. The effects of linguistic diversity among highly-skilled immigrants may therefore differ from those reported here. We also note that here, as with many recent papers, we rely on policy variation for our identification. We refer the reader to (Doran and Yoon, 2018) for historical evidence supporting the quota identification strategy. In particular, we there argue that "far from local efforts to reduce all immigration to some locations but not others, these laws were national efforts to reduce all immigration from some sources but not others" (Doran and Yoon, 2018).

The paper is organized as follows. In Section II, we review the literature on the 1920s quotas, and explain where our results fit in the context of that literature. In Section III, we introduce the data set, referring especially to (Doran and Yoon, 2018). In Section IV, we introduce our empirical strategy and estimating equations. In Section V, we describe our results. In Section VI, we conclude.

2 Existing Economics Literature on the Quotas

Before 1921, United States law placed virtually no limitations on immigration from Europe to the United States. Starting in the 1890s, Protestant Americans of Northern and Western European descent became concerned about the increased flows of non-Protestant immigrants from Southern and Eastern Europe. These concerns eventually reached an expression in law with the 1921 and 1924 immigration quotas. The Emergency Quota Act of 1921 established annual quotas for Southern and Eastern European immigration that were considerably lower than the then-current flows, while establishing quotas for Northern and Western European immigration that were barely binding. The Immigration Act of 1924 tightened the quotas on Southern and Eastern European immigration even further.

In the last several years, a total of seven papers have emerged studying the economic

impacts of the 1920s U.S. immigration quotas. After the initial work of (Ager and Hansen, 2017), these papers have been written almost simultaneously by separate teams of authors, with subtle differences in the implementation of the identification strategies, and without a planned consistency. Nevertheless, here we argue that in fact these seven papers tell a largely consistent history, in which the reported economic impacts of the quotas correspond with those predicted by models such as (Borjas, 1987), (Acemoglu, 2010), and (Tabellini, 2018). In particular, it appears that these quotas: (1) reduced immigration from some sources but not others; (2) reduced immigration to some locations but not others; (3) induced differential wage changes among natives in affected locations; (4) induced a native migration response to affected locations that was less than one for one with the immigration reductions; (5) decreased the scale and mechanization of production in affected locations; and (6) decreased natives' inventions in affected locations, especially those inventions relevant for industries that lost a large number of immigrant workers. This set of results is not only consistent with itself, but is also consistent with the new results reported here.

In this section, we review the results of this existing literature, summarizing the results and comparing them to models such as (Borjas, 1987), (Acemoglu, 2010), and the model in Appendix B of Tabellini (2018).

One of the most important papers in this literature is "Immigration in American Economic History" (Abramitzky and Boustan (2017)). Abramitzky and Boustan (2017) review the literature on historical and contemporary immigration. They focus on three major questions in the economics of immigration. First, the paper questions whether immigrants are positively or negatively selected from their home countries over time. Second, they explore how immigrants assimilate into the US. Third, they examine the effects of immigration on the economy, especially native employment and wages. In particular, they cover the two main eras of mass immigration–the Age of Mass Migration from Europe (1850-1920), an era of unrestricted migration, and a recent period of constrained mass migration from Asia and Latin America (1965-present).

First, they find that migrant selection was mixed in the past (with some migrants being positively selected and others being negatively selected from their home countries), while migrants are positively selected in the present. Specifically, migrant selection during the

Age of Mass Migration is consistent with a Roy model (Roy, 1951), as developed by (Borjas, 1987). The Roy model would predict positive selection from northern and western Europe and negative selection from southern and eastern Europe, with differences in productive skills of migrants and income equality across sending countries. Historical evidence on income distribution supports their argument. Income distribution in western European countries was similar with that of the US at that time while income distribution in the European periphery was less equal than that of the US. Consistent with the model, historical evidence suggests that low-skilled workers from southern and eastern Europe immigrated to the US and that they were thus negatively selected. The positive selection of immigrants today can be explained by both the increase in income inequality in the US (as the model would predict) and the increasing selectivity of US immigration policy, which would favor high-skilled immigration.

Second, they find that assimilation of immigrants into US economy is not consistent with the stereotypical "American Dream", whereby poor immigrants work hard and eventually become rich. During periods of mass migration, immigrants did not catch up with US natives in the past and they do not do so today, because immigrants start behind natives, and their occupational upgrading and earnings grow at a similar pace to that of US natives over time. Although immigrants experience some earnings convergence, the immigrants themselves do not catch up with US natives in the labor market during their own life times. However, these gaps diminish across generations because many children of immigrants are educated and grow up in the US.

Third, the authors argue that immigrants during the Age of Mass Migration were more substitutable with natives in agriculture and manufacturing, and that therefore there was some effect of immigration on native wages. They also find that immigration in the past contributed to the spread of large factories used for mass production. In addition, unskilled immigrants and assembly-line machinery were complementary at that time.

The first paper to make use of the quotas as part of an identification strategy to determine the economic effects of low-skilled immigration appears to be "Closing Heaven's Door: Evidence from the 1920s U.S. Immigration Quota Acts" (Ager and Hansen (2017)).

[0]See also related papers such as Ward (2017) and Greenwood and Ward (2015).

Their first main finding is that the areas with a large decline in incoming immigrants due to the quotas experienced a decrease in the foreign-born share and lower population growth. Specifically, one additional missing immigrant per-100-inhabitants-per-year led to a decline in the foreign-born share by 1.6 percentage points and a decrease in the 10-year population growth rate by 6.7 percentage points at the county level. This suggests that any compensatory migration from non-quota-restricted immigrants or from natives was not enough to counteract the effects of the quotas on quota-affected immigration. Reinforcing the effects of this main finding is an associated decline in marriage rates in quota-affected regions. Second, they show that the quotas have a significant effect on the earnings of native workers. Natives in counties exposed to the quotas were more likely to change to lower-wage occupations, though the effect varies by gender and race. In particular, white workers experienced earning losses while black workers benefited from the quotas. Earnings of white female workers were not affected, while black female workers gained significantly. These findings suggest that immigrant workers during the 1920s had a higher elasticity of substitution to black native workers. Third, they find that labor productivity in manufacturing at the city level declined under the quotas.

A third important paper in this literature is Tabellini (2018). This paper makes two main additional contributions above and beyond the points already made in the literature described above. First, Tabellini (2018) introduces a notion of linguistic distance adapted from Chiswick and Miller (2005). The results show that the impact of immigration is tied closely to the linguistic distance of the source country language compared to English. The second main contribution is to introduce a model (in online Appendix B of Tabellini (2018)) that makes the following predictions: (1) (unskilled) immigration favors capital accumulation in the unskilled sector; (2) "immigration has a positive and unambiguous effect on high skilled wages"; and (3) immigration has an ambiguous effect on low skilled wages. This theoretical framework is consistent with Tabellini (2018) by construction, but it is clearly consistent with the results of Ager and Hansen (2017) as well.

A fourth paper in this literature is Doran and Yoon (2018). This paper addresses the question of how mass migration affects innovation. In particular, the paper questions whether low-skilled immigrants could influence innovations through labor-complementary inventions

or labor-saving inventions. The results show that incumbent inventors in cities exposed to fewer low-skilled immigration inflows due to the 1920s quotas applied for fewer patents. To be specific, inventors living in quota-exposed cities that experienced a ten percent reduction in new immigrants reduced their patent applications by 0.5 percent per year. Further, the effect of quotas on patents is driven by fewer patent applications relevant for the quota-exposed industries that lost immigrant workers.

The papers above tell a consistent history of the quotas – a history that lays the groundwork for this paper. The quotas reduced low-skilled immigration; this decrease affected the large scale manufacturing that had flourished in areas with many low-skilled immigrants; and inventors who supplied patented inventions relevant for the affected industries produced fewer such inventions after the quotas.

3 Data

Our analysis relies on a panel of individual inventors, a measure of how locations are exposed to quotas, and information on the primary languages spoken by new immigrants and the pre-existing population of U.S. Cities circa 1920 (just before the quotas were enacted).

To obtain the inventor sample, we follow the method in (Doran and Yoon, 2018). In particular, we use the European Patent Office's PATSTAT database, which provides characteristics such as inventor's full name, year of patent application, and the number of citations of each patent application granted by the U.S. Patent Office from 1899 to the present. We exploit a fuzzy matching procedure that merges patents at the individual-name level into the complete count 1920 U.S. Census with names. In the 1920 Census, 43% of the U.S. population is made up of people with a unique combination of first name, middle name, and last name. If a person from this unique-name subsample is matched to a patent application made between the years of 1919 and 1929, then, barring transcription errors, that person must be the author of the patent application unless someone with the exact same name immigrated after 1919 and patented soon after arrival. Furthermore, to increase the quality of the matches, we also restrict the matches to those with an implied age at the time of application of between 18 and 80 years old.

On this matched individual inventor sample, the variables from the complete-count 1920 U.S. Census give us each individual's birth year, birth place, citizenship, nationality, geographic location at the city/county level, as well as other characteristics.

Our second data set measures how locations were differentially exposed to the quotas over time, as well as other characteristics of these locations. In (Doran and Yoon, 2018), we digitize immigration inflows by source country and year, as well as the exact size of the quotas by country and year, from administrative data obtained from Willcox et al. (1929) and the U.S. Department Commerce (1924, 1929, 1931). Using data from the 1910 and 1920 U.S. Censuses, we collect the following aggregated characteristics of each city: total population, foreign-born population, Southern and Eastern European immigrant population, Northern and Western European immigrant population, and immigrant populations by nationality and year of immigration to the U.S..

In the next section, we explain how unique features of the implementation of the quotas allow us to identify how the impact of low-skilled immigration on American innovation varies by how closely the immigrant lanugages mirror that of the pre-existing population.

4 Empirical Strategy

Typically, a shift-share instrument for immigration relies on variation in the national origin of the pre-existing population across locations, and assumes that the new immigrants will have a tendency to choose locations where people of their ethnicity or nationality already live. In most cases, this would also imply linguistic sorting, in which immigrants who speak a given language (say, for example, Italian) end up sorting to locations full of people who already speak Italian. Given such linguistic sorting, it would be difficult to use such an instrument to determine the differential impact of immigrants who speak a relatively common language among the pre-existing population from that of immigrants who speak a relatively rare language among the pre-existing population in a given city. We would need a natural experiment in which immigrants who speak Italian, for example, are often attracted to locations with relatively few Italian speakers, and immigrants who do not speak Italian are often attracted to locations with relatively many Italian speakers. These "off-diagonal" sortings

would enable us to determine whether immigrants have a differential impact when they are located in areas with relatively many or relatively few people speaking their language.

In this paper, we exploit 1920s U.S. immigration quotas that attracted speakers of a given language to locations with both relatively many and relatively few speakers of that language. In particular, the quotas suddenly cut off immigration to many cities that were "exposed" to the quotas because they had experienced recent flows of immigrants from quota-affected countries. But these cities were not all alike: some quota-affected cities were populated by the descendents of immigrants from previous generations whose families had preserved their native tongue (to the point of that language being their primary spoken method of communication). But other quota-affected cities were populated by the descendents of immigrants who preserved an ethnic kinship with the newcomers but who had not preserved their language. While both types of cities attracted new immigrants of similar background to the pre-existing foreign-born population before the quotas were enacted, and both types of cities subsequently lost these new flows of immigrants after the quotas were enacted, only the first type of city lost immigrants who spoke a language *commonly* spoken in their destination city. The second type of city lost immigrants who spoke a language *uncommonly* spoken in their destination city.

To identify the impact of low-skilled immigration on innovation in any given subsample, we follow the method in (Doran and Yoon, 2018), which built upon (Ager and Hansen, 2018). To identify how these effects of low-skilled immigration on innovation vary depending on the linguistic distance between the new immigrants and the pre-existing population in each city, we split the sample into four subsamples: (1) one in which the languages of the new immigrants had been preserved and were spoken widely among the pre-existing population; (2) another in which the languages of the new immigrants were uncommon; (3) another in which the languages were moderately common; and (4) another in which the languages were moderately uncommon. We then replicate the main analysis in (Doran and Yoon, 2018), once for each subsample.

Our main estimation equations are:

$$Y_{ict} = \alpha + \beta(Quota_c \times Post_t) + \theta X_{it} + \tau_t + \gamma_i + \epsilon_{ict} \tag{1}$$

where Y_{ict} is the number of patents by incumbent inventor i in city c and year t. We include the quartic of age of person i in year t, individual fixed effects, and year fixed effects. The quota-exposure variable is defined as follows:

$$Quota_c = \frac{100}{P_{c,1920}} \sum_{j=1}^{J} \left(\widehat{Immig}_{j,22-30} - Quota_{j,22-30} \right) \frac{FB_{jc,1920}}{FB_{j,1920}} \tag{2}$$

where $P_{c,1920}$ is the 1920 population in city c, $\widehat{Immig}_{j,22-30}$ is the estimated average immigration inflows that would have occurred per year from country j to the United States during the post-quota period from 1922 to 1930 if the quotas had not been enacted,[1] $Quota_{j,22-30}$ is the average quota for country j during the period from 1922 to 1930, $FB_{jc,1920}$ is the foreign-born population of country j in city c in 1920, and $FB_{j,1920}$ is the total foreign-born population of country j in the 1920 Census.

When a city's predicted immigration from 1922 to 1930 (predicted from the pre-WWI annual immigration flows 1900-1914) is much higher than quotas for the years 1922 through 1930, then the quota exposure variable is high. Otherwise, it is low. This quota begins to affect quota-exposed cities sometime after the quota acts of 1921 and 1924 are implemented. We compare different options for the post-t variable, including 1922 and 1924. In this regression, β represents a difference-in-differences estimate of the effect of the quotas.

We can observe which languages the pre-existing population spoke in each location by observing the individual responses in the 1920 U.S. Census to the question in column 20: "person's mother tongue". Table 1 shows that in the U.S. during 1920, there were considerable differences between the number of people born in a given country and the number of people whose mother tongue was the language of that country. Many U.S.-born individuals continued to speak the language of their immigrant parents even though they were born in the United States. This tendency for foreign-language persistence across generations varied from city to city, and this variation allows us to divide locations into those in which the languages of the new immigrants were common and those in which the languages of the new immigrants were not.

[1]The estimates are predicted from the pre-WWI annual immigration flows 1900-1914 based on the regression model: $Immig_{jt} = \beta_1 lnt + \beta_2 (lnt)^2 + \epsilon_{jt}$ (Ager and Hansen, 2018; Doran and Yoon, 2018)

Each city c has a vector of languages in which each element, $PreLang_{lc}$, is the share of the pre-existing population in city c whose mother tongue is l:

$$PreLang_{lc} = \frac{Lang_{lc,1920}}{TotalPopulation_{c,1920}} \tag{3}$$

Each city c also has a vector of languages in which each element, $NewLang_{lc}$, is the share of the missing immigrants between 1922 and 1930 in city c whose mother tongue is l:

$$NewLang_{lc} = \frac{ImmigLang_{lc}}{TotalMissingImmig_{c}} \tag{4}$$

To calculate how close the languages of the new immigrants were to the languages spoken by the pre-existing population, we need to determine how "close" the vector $PreLang$ is to the vector $NewLang$. There is no mathematically unique way to determine how "close" two vectors are to each other. We make use of two methods used in (Borjas and Doran, 2012): the correlation coefficient and the index of similarity.

The correlation coefficient is well-known. The index of similarity of (Bojas and Doran, 2012) is based on the "Index of Dissimilarity" used by (Cutler and Glaeser 1997) and introduced by Duncan and Duncan (1955). We calculate the Index of Similarity with the following formula:

$$LangIndex_c = 1 - \frac{1}{2}\sum_{l=1}^{L} |PreLang_{lc} - NewLang_{lc}| \tag{5}$$

Clearly, the Index of Similarity will be one when the languages of the missing immigrants are distributed identically to the languages of the pre-existing population. If the pre-existing languages and the languages of the missing immigrants never match in the same city, then the Index will be zero.[2] For each of these measures of linguistic "closeness" between the missing immigrants and the pre-existing population, we divide the cities into four equal groups by quartiles. We report simple statistics in Table 2. It is clear that there are similar inflows of new immigrants and similar rates of patent applications by incumbent native inventors

[2]If two vectors never match, then
$LangIndex_c = 1 - \frac{1}{2}\sum_{l=1}^{L} |PreLang_{lc} - NewLang_{lc}| = 1 - \frac{1}{2}\sum_{l=1}^{L} |PreLang_{lc}| - \frac{1}{2}\sum_{l=1}^{L} |NewLang_{lc}| = 0$

across all quota-exposed cities, regardless of the degree of linguistic closeness of the missing immigrants to that city.

In the next section, we determine whether the quota-induced change in immigration had differential impacts on innovation depending on whether the immigrants spoke a relatively common local language or not.

5 Results

We begin our analysis by showing that the quotas decreased immigration inflows in quota-exposed locations regardless of linguistic distance. In Table 3 and 4, we report the results when the outcome variable is newly arrived immigrants rescaled by the 1920 city population in a given location in a given year, after we split the sample into the four subsamples using the correlation coefficient and the index of similarity respectively. It is apparent that regardless of which of the four subsamples partitioned by linguistic distance we consider, which of the two proxies (the correlation coefficient and the index of similarity) for linguistic distance we use to split the sample, which years we include in the sample, and which year we use as the post-treamtent year, the quota-exposed cities experienced substantial reductions to their immigrant inflows. Next, we determine how the quotas, which decreased immigration inflows to all four groups of cities, differentially affected innovation (as measured by patents) depending on the linguistic closeness of the missing immigrants to the pre-existing population.

The quota-exposure variable represents the average annual number of immigrants per-100-inhabitants in a city that were "missing" due to the quotas (Ager and Hansen 2018; Doran and Yoon 2018). Doran and Yoon (2018) find that a one-unit increase in quota exposure decreases immigration inflows by approximately 100% and decreases patent applications by incumbent native-born inventors by about 5%. Thus, for every 10% decrease in immigration inflows, patent applications per year decrease by 0.5%.

Here, we explore how these results vary with respect to the linguistic closeness of the missing immigrants to the pre-existing population. In Table 5, we report the results of estimating equation (1) on the four subsamples of cities, partitioned by linguistic closeness

measured through the correlation coefficient. It is clear that the effect of the quotas on native patenting is most significant when the missing immigrants and the pre-existing population are moderately close linguistically. Moderately far linguistic distance cities experience the second-largest and second-most significant effect on patenting. The linguistically close and linguistically far cities experience smaller and less significant effects.

In Table 6, we report the results of the same estimation when we partition the sample of cities according to linguistic closeness as measured by the Index of Similarity. Here, the results show that the effect of losing immigrants through the quotas on native patenting is positive and significant if the missing immigrants were very linguistically different from the pre-existing population. The effect becomes negative and significant when the missing immigrants are moderately linguistically close to the pre-existing population; the effect is slightly smaller for those that are very linguistically close.

Figure 1 plots the difference between patent applications per year for native inventors in quota-exposed cities and non-quota-exposed cities (above and below the median of quota exposure) over time (before and after the quotas). Each panel of Figure 1 reports this plot for one of the four groups of cities partitioned according to the linguistic closeness of the missing immigrants with the pre-existing population, as measured by the correlation coefficient. It is apparent that the largest trend breaks in patent applications at the onset of the quotas are in the moderately close and moderately far cities; the far and close cities exhibit smaller or non-existent trend breaks in patenting at the time of quota onset.

Figure 2 reports the size and confidence intervals of the estimates using the most reliable patent matching (the 1919-1929 data) and 1924 as the first post-treatment year. It is clear that linguistically far missing immigrants produce either small and insignificant or positive and significant effects on native patenting. In contrast, missing immigrants with a moderate linguistic distance to the pre-existing population have a large, negative, and significant effect on native patenting. This effect is attenuated for very linguistically-close immigrants.

To get a sense of the scale of these effects, we must compare the effect of the quotas on patenting (reported in Tables 5 and 6) with the effect of the quotas on immigrant inflows (reported in Tables 3 and 4). After all, the size of the "first-stage" effects of the quotas on immigrant inflows varied slightly across samples, and this variation could be related to the

variation across samples in the effects of the quotas on patenting which we report in Tables 5 and 6. We first rescale each of the estimated coefficients reported in Tables 3 and 4 by their respective pre-quota means, to obtain the percent declines in new immigrant inflows caused by the quotas in each subsample. We then perform the same rescaling on the effects of the quotas on patenting reported in Tables 5 and 6. Finally, we divide the latter percentages by the former to obtain the ratios reported in Table 7. The p-values are computed using the Holm–Bonferroni method (Holm, 1979).

We report in Table 7 the resulting estimates for how a 100% increase in immigration inflows effects the patenting of incumbent native inventors. In Panel A, we report the results when we measure linguistic closeness through the correlation coefficient, and in Panel B we report the results when we measure linguistic closeness through the Index of Similarity. In Panel A, we find that for every increase in immigration inflows of 10%, patent applications per year increase by about 1% in both moderately linguistically close cities and moderately linguistically far cities. In contrast, there is no significant effect for linguistically far and linguistically close cities. In Panel B, we find that for every increase in immigration inflows of 10%, patent applications per year increase by about 1% in both moderately linguistically close and linguistically close cities. In contrast, for linguistically far cities, for every increase in immigration inflows of 10%, patent applications per year decrease by 3%.

In Figure 3, we summarize these results, providing graphical evidence of a "U-shaped" curve in the effect of linguistic distance between newcomers and the pre-existing population on patent applications.

6 Conclusion

In this paper, we explore the mediating role of language in the effect of immigrants on innovation. We find, as in (Doran and Yoon, 2018), that low-skilled immigrants affect the innovation of pre-existing native inventors. But we further find that the language the immigrants speak matters.

Intuitively, if innovation is the recombination of existing ideas or experiences into new ones (Weitzman, 1998), then anything that effects this recombination could affect innovation.

Linguistic diversity could effect the number of unique ideas people have to talk about as well as the ability of people to talk about them. The first effect would make linguistic diversity favorable for innovation; the second effect would make linguistic homogeneity favorable for innovation.

It is plausible, therefore, that the optimal amount of linguistic diversity is somewhere in-between complete diversity and complete homogeneity. The results we report here are consistent with this hypothesis.

These results are of course specific to the low-skilled immigrant workforce prevalent at the time, as well as to the state of knowledge and types of inventions common during the period. Future research should determine whether the benefits of new ideas, abilities, and experiences from a linguistically diverse *highly skilled* immigrant pool outweigh any communication barriers they bring.

References

Abramitzky, R. and L. Boustan (2017). Immigration in American economic history. *Journal of Economic Literature 55 (4)*, 1311-4.

Abramitzky, R., L. P. Boustan, and K. Eriksson (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy 122(3)*, 467-506.

Acemoglu, D. (2010). When does labor scarcity encourage innovation? *Journal of Political Economy 118(6)*, 1037-1078.

Ager, P. and C. W. Hansen (2018). Closing heaven's door: Evidence from the 1920s us immigration quota acts.

Alesina, A. and Ferrara, E. L. (2005). Ethnic diversity and economic performance. *Journal of economic literature, 43(3)*, 762-800.

Ballatore, R. M., Fort, M., & Ichino, A. (2018). Tower of Babel in the Classroom: Immigrants and Natives in Italian Schools. *Journal of Labor Economics*, 36(4).

Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *The American Economic Review*, 531-553.

Borjas, G. J., & Doran, K. B. (2012). The collapse of the Soviet Union and the productivity of American mathematicians. *The Quarterly Journal of Economics*, 127(3), 1143-1203.

Chiswick, B. R. and Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development*, 26(1), 1-11.

Delgado Gómez-Flors, M. (2018). Does immigrant diversity affect productivity? The Spanish experience.

Doran, K. B. and Yoon, C. (2018). Immigration and Invention: Evidence from the Quota Acts.

Duncan, O., & Duncan, B. (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review*, 20(2), 210-217.

Greenwood, M. J. and Z. Ward (2015). Immigration quotas, world war i, and emigrant flows from the united states in the early 20th century. *Explorations in Economic History 55*, 76-96.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers*, 3(2), 135-146.

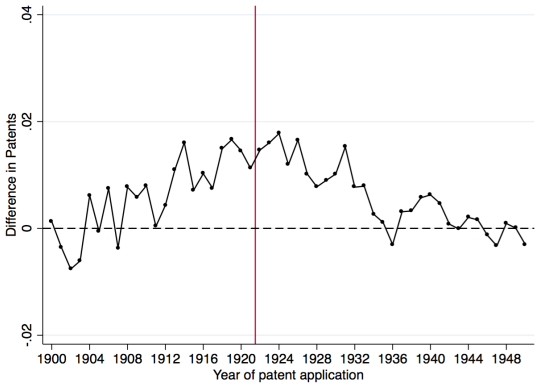
Tabellini, M. (2018). Gifts of the immigrants, woes of the natives: Lessons from the age of mass migration.

U.S. Department of Commerce (1924, 1929, 1931). Statistical Abstract of the United States.

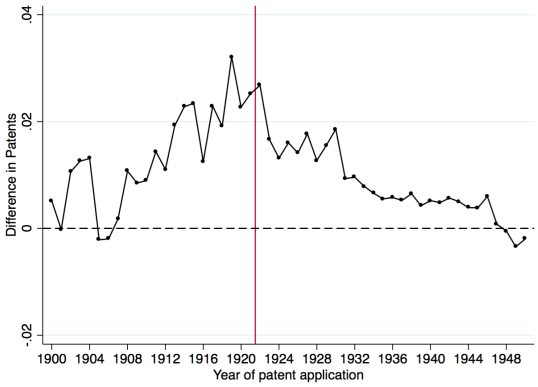
Ward, Z. (2017). Birds of passage: Return migration, self-selection and immigration quotas. *Explorations in Economic History 64*, 37-52.

Willcox, W. F. et al. (1929). International migrations, volume i: Statistics. *NBER Books.*

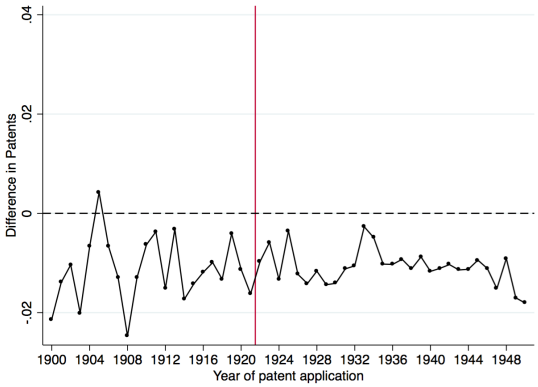
Figure 1: THE EFFECT OF THE QUOTAS ON PATENT APPLICATIONS PER YEAR



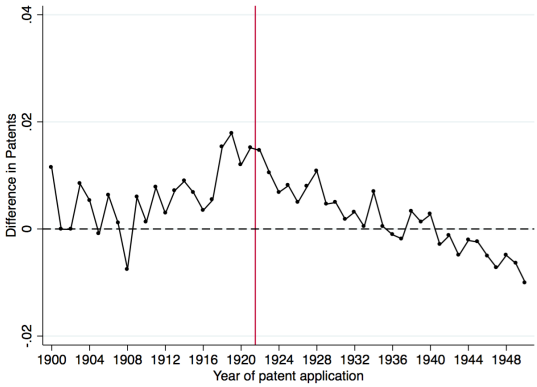
(a) Linguistically Close Cities



(b) Linguistically Moderately Close Cities



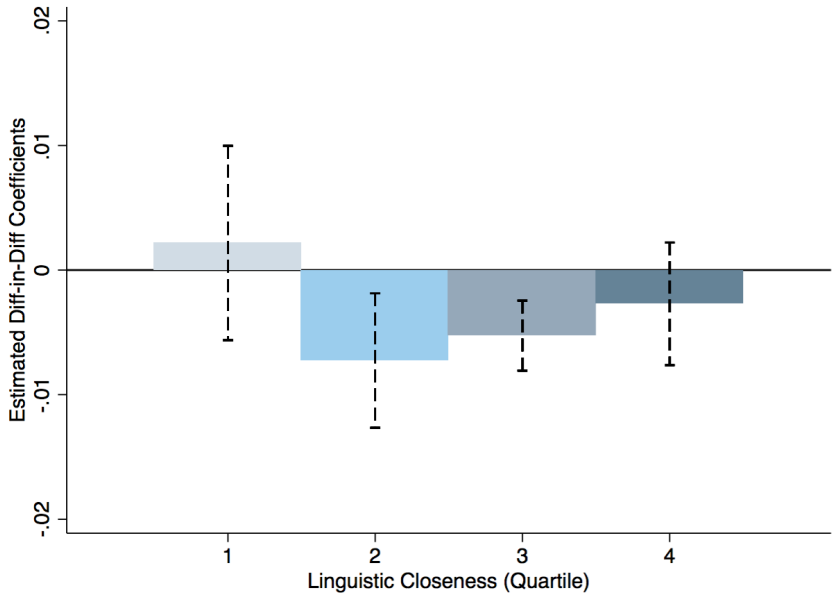
(c) Linguistically Far Cities



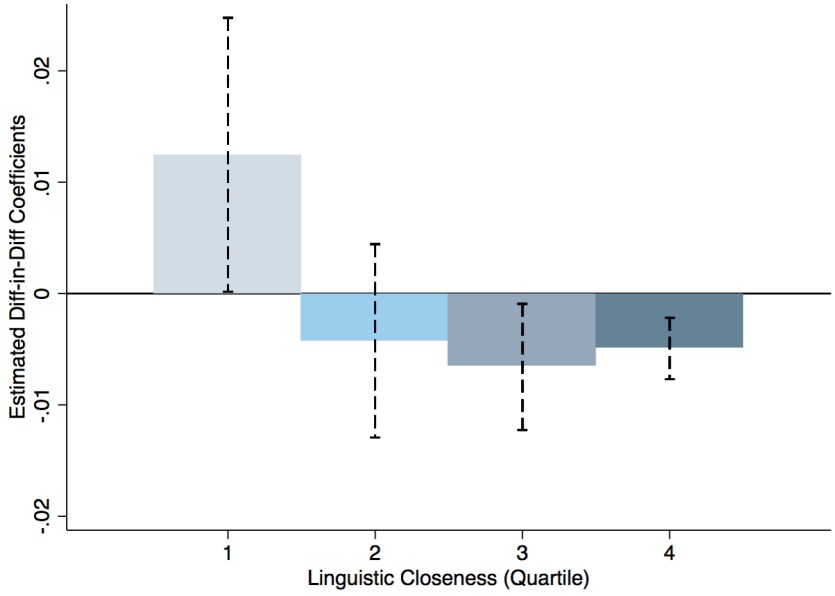
(d) Linguistically Moderately Far Cities

Note: The figures show the difference in the number of patent applications per year by incumbent inventors between quota exposed cities (those where the quota exposure variable is greater than or equal to the median) and non quota exposed cities (those where the variable is below the median). The sample is partitioned into four subsamples according to linguistic closeness between the missing immigrants and the pre-existing population, as measured through the correlation coefficient.

Figure 2: DIFFERENCE-IN-DIFFERENCES COEFFICIENTS OF THE EFFECT OF THE QUOTAS ON PATENT APPLICATIONS PER YEAR



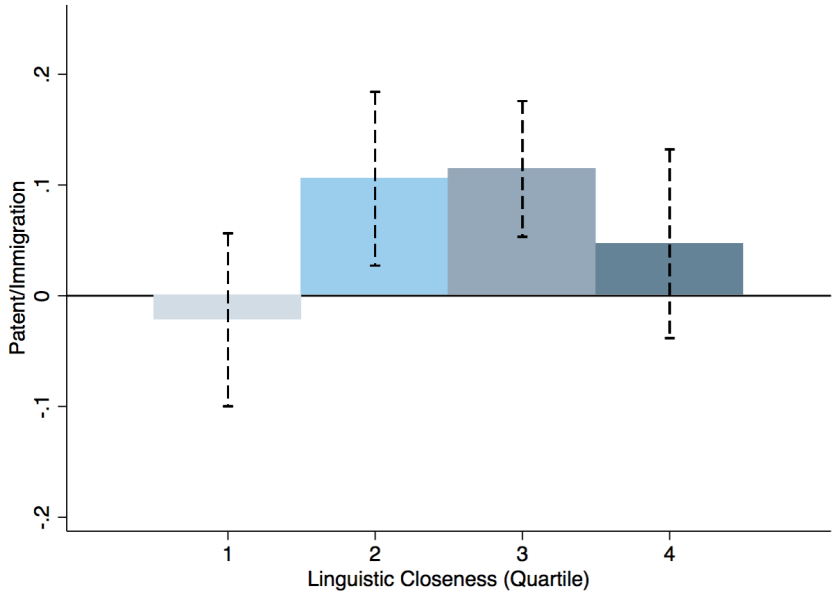
(a) Correlation Coefficient



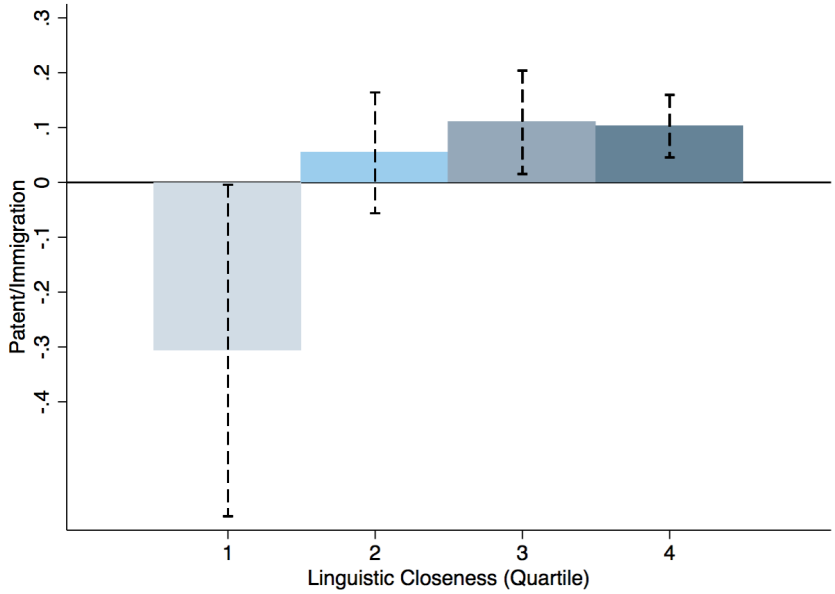
(b) Index of Similarity

Note: Panels (a) and (b) above represent the estimated diff-in-diff coefficients in column (4) of Table 5 and 6, respectively. The estimate from the linguistically far cities is located in the first quartile, while the fourth quartile shows the coefficient from the linguistically close cities. The coefficients measure the effect of the quotas on the number of patent applications per year by incumbent inventors (those who had at least one patent by 1919) during the years 1919 through 1929, using 1924 as the first post-treatment year.

Figure 3: EFFECT OF IMMIGRANT INFLOWS ON PATENT APPLICATIONS



(a) Correlation Coefficient



(b) Index of Similarity

Note: The figure graphically represents the estimated effects reported in Table 7. We rescale each of the estimated coefficients reported in Tables 3 and 4 by their respective pre-quota means, to obtain the percent declines in new immigrant inflows caused by the quotas in each subsample. We then perform the same rescaling on the effects of the quotas on patenting reported in Tables 5 and 6. Finally, we divide the latter percentages by the former to obtain the ratios reported above. The p-values are computed using the Holm–Bonferroni method (Holm, 1979).

Table 1: BIRTHPLACE AND MOTHER TONGUE IN THE 1920 U.S. CENSUS

	Birthplace							
Mother Tongue	U.S. (1)	U.K. (2)	Ireland (3)	Germany (4)	Italy (5)	Russia (6)	Poland (7)	Others (8)
English	53.12	83.38	73.37	0.93	0.33	0.52	0.25	15.99
German	17.34	0.52	0.34	95.96	0.17	8.62	7.53	8.54
Italian	4.35	0.13	0.02	0.10	99.03	0.12	0.13	0.54
Celtic	3.70	14.20	25.59	0.03	0.05	0.10	0.10	0.59
Polish	2.10	0.09	0.05	1.43	0.01	5.10	78.54	1.58
Spanish	2.51	0.04	0.02	0.05	0.05	0.05	0.02	9.67
French	2.63	0.08	0.08	0.10	0.08	0.06	0.03	7.85
Swedish	2.40	0.15	0.10	0.06	0.01	0.09	0.02	10.69
Jewish	1.52	0.83	0.07	0.23	0.04	49.98	8.56	2.68
Norwegian	1.37	0.01	0.01	0.04	0.00	0.01	0.00	5.99
Crech	1.03	0.01	0.00	0.08	0.00	0.08	0.07	3.63
Russian	0.79	0.11	0.00	0.04	0.00	23.91	1.15	0.14
Dutch	0.84	0.10	0.03	0.13	0.01	0.05	0.03	2.90
Danish	0.79	0.02	0.05	0.30	0.01	0.02	0.01	2.98
Hungarian	0.69	0.01	0.01	0.09	0.01	0.06	0.07	4.57
Others	4.80	0.32	0.26	0.43	0.20	11.23	3.49	21.64
Total	91,683,696	1,153,841	1,049,330	1,631,480	1,608,841	1,450,734	1,133,710	6,033,502

Notes: This table shows the relationship between birthplace and mother tongue in the 1920 U.S. Census. Each number indicates the percentage of people reporting a given mother tongue out of those born in a given birthplace. The U.K. numbers in column (2) exclude people born in Ireland. The last row shows the total number of individuals born in a given birthplace.

Table 2: SUMMARY STATISTICS

Variables	Quota Exposed Cities	Non Quota Exposed Cities
Number of Cities	1668	1669
Quota Exposure	0.5805 (0.6110)	0.0237 (0.0217)
Index of Similarity	0.3021 (0.1356)	0.2617 (0.2061)
Correlation Coefficient	0.2015 (0.2259)	0.2446 (0.3050)
Population in 1920 Census	46128 (130596)	19902 (12276)
Southern and Eastern Foreign Born in 1920	3826 (24679)	34 (54)
New Immigrants per year and city as a Fraction of 1920 Population, 1900-1921	0.0039 (0.0052)	0.0007 (0.0027)
Patents per year and inventor, 1900-1921	0.1200 (0.1270)	0.1295 (0.1794)
Cities With Linguistically Close Missing Immigrants		
Quota Exposure	0.9248 (0.8258)	0.0204 (0.0212)
New Immigrants	0.0054 (0.0060)	0.0006 (0.0031)
Patents	0.1230 (0.1103)	0.1273 (0.1819)
Cities With Linguistically Moderately Close Missing Immigrants		
Quota Exposure	0.5351 (0.4926)	0.0304 (0.0240)
New Immigrants	0.0039 (0.0051)	0.0010 (0.0032)
Patents	0.1218 (0.1188)	0.1183 (0.1359)
Cities With Linguistically Moderately Far Missing Immigrants		
Quota Exposure	0.3486 (0.3132)	0.0268 (0.0224)
New Immigrants	0.0031 (0.0044)	0.0007 (0.0026)
Patents	0.1212 (0.1453)	0.1207 (0.1548)
Cities With Linguistically Far Missing Immigrants		
Quota Exposure	0.4764 (0.5201)	0.0203 (0.0188)
New Immigrants	0.0029 (0.0046)	0.0004 (0.0022)
Patents	0.1073 (0.1359)	0.1426 (0.2096)

Notes: This table presents means and standard deviations (in parenthesis) of variables used in our analysis, in subsamples defined by quota exposure of cities (above and below the median) and by linguistic closeness (as measured through the correlation coefficient).

Table 3: HOW THE EFFECT OF THE QUOTAS ON IMMIGRANT INFLOWS VARIES WITH LINGUISTIC CLOSENESS, AS MEASURED BY THE CORRELATION COEFFICIENT

	Year of Immigration			
	1900-1929		1919-1929	
	Post-Treatment Year			
	1922	1924	1922	1924
	(1)	(2)	(3)	(4)
Dependent Variable: New Immigrants as a Fraction of 1920 Population				
A. Linguistically Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0028***	-0.0029***	-0.0008***	-0.0010***
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Dependent Variable Mean	0.0025	0.0024	0.0018	0.0016
Number of Observations	23190	23190	8503	8503
Number of Cities	773	773	773	773
R-squared	0.5447	0.5409	0.6364	0.6422
B. Linguistically Moderately Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0024***	-0.0025***	-0.0006***	-0.0009***
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Dependent Variable Mean	0.0028	0.0027	0.0021	0.0020
Number of Observations	23850	23850	8745	8745
Number of Cities	795	795	795	795
R-squared	0.5576	0.5558	0.6972	0.7010
C. Linguistically Moderately Far Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0030***	-0.0032***	-0.0008***	-0.0011***
	(0.0004)	(0.0004)	(0.0002)	(0.0001)
Dependent Variable Mean	0.0024	0.0023	0.0016	0.0016
Number of Observations	22890	22890	8393	8393
Number of Cities	763	763	763	763
R-squared	0.5633	0.5604	0.6757	0.6821
D. Linguistically Far Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0034***	-0.0035***	-0.0005***	-0.0009***
	(0.0003)	(0.0002)	(0.0002)	(0.0002)
Dependent Variable Mean	0.0016	0.0015	0.0009	0.0008
Number of Observations	22170	22170	8129	8129
Number of Cities	739	739	739	739
R-squared	0.5279	0.5227	0.6340	0.6439

Notes: The outcome variable of new immigrants is constructed by combining information from the 1910, 1920, and 1930 U.S. Census. Specifically, new immigrants per year between the years 1900 and 1909 are obtained from the 1910 Census data, those between 1910 and 1919 from the 1920 U.S. Census, etc. We restrict data to cities that exist in all three censuses to obtain a balanced panel. The sample of cities is partitioned into four equally-sized subsamples according to the quartile of linguistic closeness, as measured through the correlation coefficient.

Table 4: HOW THE EFFECT OF THE QUOTAS ON IMMIGRANT INFLOWS VARIES WITH LINGUISTIC CLOSENESS, AS MEASURED BY THE INDEX OF SIMILARITY

	Year of Immigration			
	1900-1929		1919-1929	
	Post-Treatment Year			
	1922	1924	1922	1924
	(1)	(2)	(3)	(4)
Dependent Variable: New Immigrants as a Fraction of 1920 Population				
A. Linguistically Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0026*** (0.0002)	-0.0027*** (0.0002)	-0.0008*** (0.0001)	-0.0010*** (0.0001)
Dependent Variable Mean	0.0033	0.0032	0.0022	0.0020
Number of Observations	22650	22650	8305	8305
Number of Cities	755	755	755	755
R-squared	0.5905	0.5864	0.7680	0.7761
B. Linguistically Moderately Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0027*** (0.0003)	-0.0029*** (0.0003)	-0.0007*** (0.0002)	-0.0012*** (0.0001)
Dependent Variable Mean	0.0028	0.0028	0.0021	0.0020
Number of Observations	23730	23730	8701	8701
Number of Cities	791	791	791	791
R-squared	0.5351	0.5346	0.6228	0.6299
C. Linguistically Moderately Far Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0036*** (0.0003)	-0.0039*** (0.0003)	-0.0004* (0.0002)	-0.0012*** (0.0003)
Dependent Variable Mean	0.0020	0.0020	0.0014	0.0013
Number of Observations	23400	23400	8580	8580
Number of Cities	780	780	780	780
R-squared	0.5127	0.5115	0.6444	0.6470
D. Linguistically Far Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0034*** (0.0003)	-0.0033*** (0.0002)	-0.0001 (0.0001)	-0.0003*** (0.0001)
Dependent Variable Mean	0.0011	0.0011	0.0007	0.0006
Number of Observations	22320	22320	8184	8184
Number of Cities	744	744	744	744
R-squared	0.4990	0.4905	0.6076	0.6084

Notes: The outcome variable of new immigrants is constructed by combining information from the 1910, 1920, and 1930 U.S. Census. Specifically, new immigrants per year between the years 1900 and 1909 are obtained from the 1910 Census data, those between 1910 and 1919 from the 1920 U.S. Census, etc. We restrict data to cities that exist in all three censuses to obtain a balanced panel. The sample of cities is partitioned into four equally-sized subsamples according to the quartile of linguistic closeness, as measured through the Index of Similarity.

Table 5: HOW THE EFFECT OF THE QUOTAS ON PATENTS VARIES WITH LINGUISTIC CLOSENESS, AS MEASURED BY THE CORRELATION COEFFICIENT

	Year of Patent Application			
	1900-1950		1919-1929	
	Post-Treatment Year			
	1922	1924	1922	1924
	(1)	(2)	(3)	(4)
Dependent Variable: Patents by Incumbent Inventors in 1919				
A. Linguistically Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0011	-0.0031*	0.0004	-0.0027
	(0.0020)	(0.0019)	(0.0035)	(0.0025)
Dependent Variable Mean	0.1215	0.1173	0.1010	0.0906
Number of Observations	1217491	1217491	292122	292122
Number of Inventors	27170	27170	27170	27170
Number of Cities	845	845	845	845
R-squared	0.2540	0.2540	0.4292	0.4292
B. Linguistically Moderately Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0022*	-0.0033***	-0.0047***	-0.0053***
	(0.0012)	(0.0012)	(0.0015)	(0.0014)
Dependent Variable Mean	0.1291	0.1244	0.1105	0.0975
Number of Observations	2370644	2370644	565794	565794
Number of Inventors	52385	52385	52385	52385
Number of Cities	813	813	813	813
R-squared	0.2302	0.2302	0.3999	0.3999
C. Linguistically Moderately Far Missing Immigrants				
Quota Exposure × Post-Treatment	0.0002	-0.0020	-0.0060*	-0.0073***
	(0.0023)	(0.0022)	(0.0031)	(0.0027)
Dependent Variable Mean	0.1250	0.1204	0.1060	0.0933
Number of Observations	2018139	2018139	482816	482816
Number of Inventors	44749	44749	44749	44749
Number of Cities	816	816	816	816
R-squared	0.2149	0.2149	0.3850	0.3850
D. Linguistically Far Missing Immigrants				
Quota Exposure × Post-Treatment	0.0042	0.0031	0.0035	0.0022
	(0.0031)	(0.0028)	(0.0054)	(0.0040)
Dependent Variable Mean	0.1203	0.1157	0.1008	0.0886
Number of Observations	965134	965134	231378	231378
Number of Inventors	21398	21398	21398	21398
Number of Cities	813	813	813	813
R-squared	0.2494	0.2494	0.3945	0.3945

Notes: The outcome variable is the number of patent applications per year by native-born incumbent inventors who already had at least one patent in 1919. The sample of cities is partitioned into four equally-sized subsamples according to the quartile of linguistic closeness, as measured through the correlation coefficient.

Table 6: HOW THE EFFECT OF THE QUOTAS ON PATENTS VARIES WITH LINGUISTIC CLOSENESS, AS MEASURED BY THE INDEX OF SIMILARITY

	Year of Patent Application			
	1900-1950		1919-1929	
	Post-Treatment Year			
	1922	1924	1922	1924
	(1)	(2)	(3)	(4)
Dependent Variable: Patents by Incumbent Inventors in 1919				
A. Linguistically Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0016	-0.0028***	-0.0037**	-0.0049***
	(0.0011)	(0.0010)	(0.0016)	(0.0014)
Dependent Variable Mean	0.1277	0.1233	0.1084	0.0970
Number of Observations	1892525	1892525	452132	452132
Number of Inventors	41902	41902	41902	41902
Number of Cities	823	823	823	823
R-squared	0.2400	0.2401	0.4147	0.4147
B. Linguistically Moderately Close Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0039*	-0.0054**	-0.0066**	-0.0066**
	(0.0020)	(0.0021)	(0.0026)	(0.0029)
Dependent Variable Mean	0.1253	0.1209	0.1103	0.0971
Number of Observations	2534367	2534367	605425	605425
Number of Inventors	56022	56022	56022	56022
Number of Cities	823	823	823	823
R-squared	0.2175	0.2175	0.3949	0.3949
C. Linguistically Moderately Far Missing Immigrants				
Quota Exposure × Post-Treatment	-0.0076**	-0.0085**	-0.0047	-0.0042
	(0.0036)	(0.0034)	(0.0065)	(0.0044)
Dependent Variable Mean	0.1216	0.1169	0.0990	0.0871
Number of Observations	1436744	1436744	344294	344294
Number of Inventors	31986	31986	31986	31986
Number of Cities	822	822	822	822
R-squared	0.2464	0.2464	0.3978	0.3978
D. Linguistically Far Missing Immigrants				
Quota Exposure × Post-Treatment	0.0099*	0.0119**	0.0082	0.0125**
	(0.0052)	(0.0050)	(0.0071)	(0.0063)
Dependent Variable Mean	0.1250	0.1196	0.0977	0.0851
Number of Observations	707772	707772	170259	170259
Number of Inventors	15792	15792	15792	15792
Number of Cities	819	819	819	819
R-squared	0.2391	0.2391	0.3783	0.3783

Notes: The outcome variable is the number of patent applications per year by native-born incumbent inventors who already had at least one patent in 1919. The sample of cities is partitioned into four equally-sized subsamples according to the quartile of linguistic closeness, as measured through the Index of Similarity.

Table 7: EFFECT OF IMMIGRATION INFLOWS ON PATENTING

	Linguistic Closeness			
	Far	Moderately Far	Moderately Close	Close
	(1)	(2)	(3)	(4)
A. Correlation Coefficient				
Patent/Immigration	-0.0217	0.1057	0.1145	0.0469
p-value	0.5859	0.0083	0.0003	0.2802
B. Index of Similarity				
Patent/Immigration	-0.3064	0.0539	0.1095	0.1025
p-value	0.0469	0.3377	0.0227	0.0004

Notes: This table shows the effect of quotas on patents relative to its effect on immigration inflows by dividing the estimated coefficients on patents relative to its mean (in column (4) on Table 5 and 6 respectively) by the estimated coefficients on immigration inflows relative to its mean (in column (4) on Table 3 and 4 respectively). The p-values are computed using Holm–Bonferroni method. The estimated effects are graphically shown on Figure 3.