ONLINE APPENDIX

A. Instrumental Variable Analysis

A possible concern in our empirical analysis is that while name Americanization might occur before occupational change for some individuals, the opposite might hold for others.

To address this concern, we present here an instrumental variable (IV) strategy. We calculate the "Scrabble points" for each name at birth by summing the scores attributed to each letter in the popular board game and use these points to predict name Americanization. The origin of the Scrabble point system dates to 1938 and is attributed to the architect Alfred Moscher Butts, who performed a frequency analysis of letters from the front page of various newspapers. Scrabble points capture the structure of words, measuring both their length and how uncommon their letters are. Therefore, they provide a measure encapsulating the graphemic and phonemic features of names. A name associated with high points corresponds to a complex word, whereas one with low points corresponds to a simple or euphonious word. At the same time, Scrabble points convey no information about the semantic, etymology or ethnic origin of names, or about their pronunciation.[1] Due to exposure to the U.S. linguistic system, the structure of migrants' names experienced an increase in complexity (measured by Scrabble points), leading them to adopt popular American names.

In practice, we create a measure of distance between the Scrabble points of the migrant's name at birth and the Scrabble points associated with the "American norm". Our Scrabble index S^k_{Birth} is defined as:

$$S^k_{Birth} = \frac{SP^k_{Birth}}{\overline{SP}}, \tag{A1}$$

where SP^k_{Birth} is the Scrabble points of name k and $\overline{SP}$ the median Scrabble points across American individuals living in the state of New York in 1930 (excluding name k from the computation of the median).

It is important to remember that the IV estimates measure a local average treatment effect, provided the assumptions given in Angrist et al. (1996) are satisfied. The estimate is then the average effect of Americanizing names on occupational scores for complier migrant men only, i.e., those who would change their name if they had a sufficiently high Scrabble index but would not change their name otherwise.[2]

[1]To observe this, consider that the anagram of a name has a completely different meaning and spelling, but identical Scrabble points.

[2]Our instrument identifies the causal effect of name Americanization on the sub-population of compliers. While our instrument is continuous we could give an idea of the size of this group by looking at individuals

Identification relies on the following exclusion restriction assumption: while name popularity (A_i) influences labor market outcomes – since names implicitly signal individuals' socio-economic background (e.g., Bertrand and Mullainathan, 2004, Fryer and Levitt, 2004) – names' linguistic structure does not have a direct impact in the labor market. In other words, we assume that S^k_{Birth} is uncorrelated with ϵ_{it}.

To clarify this point, consider an individual named "Guido" – a distinctively Italian name which later in the 1900s became a demeaning term to identify Italian Americans. This individual certainly has a different occupational trajectory than migrants with names of different origin and hinting at varying religious and ethnic backgrounds such as "Gunnar", "Georg", "Olaf", "Isaac" or "Moses". Nonetheless, all these names have an identical Scrabble index and hence their linguistic structure is likely to be unrelated to their labor market outcomes.[3] By contrast, "Guido" and "Salvatore" both have Italian origin, yet very different Scrabble points.

While we cannot directly test for the exclusion restriction of our instrument, we provide evidence of two key results that are suggestive of it.

Validity of the Scrabble index. To be a suitable instrument, the Scrabble index should predict earnings growth only through name Americanization, and not through a direct association with changes in labor market outcomes. In other words, the linguistic structure of names should not bear any relevant information about unobservable characteristics of migrants that affect occupational upgrading, especially after controlling for the country of birth, length of stay in the country and household characteristics. We can think of two issues that might directly affect the validity of the instrument and hence we propose two checks.

The first argument is that the linguistic structure of the name might be directly associated with labor market outcomes. It should be noted that in our specifications we already control for time-invariant factors, including a potential "distaste" with respect to certain countries. Hence, in order to invalidate the exclusion restriction, preferences for name structure should vary across names from the same country of origin. Using data from the National Opinion Research Centers General Social Survey between 1994 and 2002, Aura and Hess (2010) show that only some name characteristics correlate with socio-economic background and lifetime outcomes. For instance, they find (Table 2, p. 222) that individuals with more popular names have higher educational attainment and have more educated parents. However, the

with Scrabble index above/below the median. Using a standard calculation (Morgan and Winship, 2014), the percentage of compliers is 20%. The proportion of individuals who would have Americanized had they had a Scrabble index above the median is 8% (note, these shares are usually low in the literature).

[3] Recall that in the estimation we control for country of birth and time-invariant characteristics; therefore, the appropriate comparison should actually be between names of individuals from the same country.

Scrabble score they calculate does not correlate with any of these characteristics and both the statistical significance and magnitude of the coefficients is close to zero. Additionally, while name popularity is strongly and positively correlated with better financial status, higher family income and lower likelihood of having a child when young, Scrabble does not correlate with the key explanatory variables, aside from a statistically weak correlation with happiness and the likelihood of having a child when young (Table 3, p. 224).

To further corroborate the lack of correlation between name features and economic success, we use the 1930 census to estimate how occupational-based earnings relate to name characteristics among U.S.-born males living in the state of New York. We focus on natives only because, in a reduced form, we would expect Scrabble to affect outcomes if migrants were included in the regression.

The first panel of Table A1 shows the results of our analysis. We correlate the log-occupational scores with the Americanization index, the Scrabble index and a variable for literacy. Estimates show that having a popular name is positively associated with labor market outcomes of U.S.-born (column (a)), while the linguistic features of the name are not (column (b)). Furthermore, when we condition on name popularity, we still find no association between the Scrabble index and occupational scores (column (c)). Finally, if we further control for literacy (a proxy for higher socio-economic status), the results are unchanged (column (d)).[4] Therefore, we conclude that names matter in the U.S. market, although employers and customers do not seem to attribute any direct "price" to their linguistic structure.

The second argument that would invalidate the instrument is that linguistic structure could capture unobserved migrants' traits that are directly correlated with wage growth. Hence, in our second check we show that the Scrabble index is uncorrelated with various measures of migrant socio-economic background *within* a country. For this purpose, we extract additional characteristics of migrants obtained from the naturalization documents, and which we argue capture differences in individual backgrounds within country. Results are reported in the second panel of Table A1. We performed regressions of the Scrabble index on indicators for height deciles, for month of birth, for port of exit and for height, month of birth and port of exit taken together. In all regressions we included indicators for country of birth.[5] The panel reports $F-$tests on the joint significance of the regressors in each specification.

We start by looking at the correlation between the Scrabble index and deciles of the height distribution. Such correlation could cast doubts on the validity of the IV procedure,

[4]Estimates are unaffected when we additionally control for age, race and state of birth.
[5]We use 73 indicators to account for all different origins.

given that height correlates with economic aspects such as skills, education, income, wealth, and, consequently, earnings trajectories. We find that the correlation between linguistic structure of a name and deciles of height is both statistically insignificant and of negligible magnitude. The p-value associated with the $F-$test on these indicators is 0.78.

We continue this exercise by checking the correlation between Scrabble and indicators for month of birth. Research has shown strong effects of month of a child's birth on later outcomes involving health, educational attainment, earnings, and mortality (e.g., see Buckles and Hungerman, 2013). As in the case of height, a possible association between a name's linguistic structure and month of birth would indicate that the instrument could correlate with labor market prospects in the U.S. through channels other than name changes. The $p-$value of 0.58 indicates that the $F-$test fails to reject the null of all coefficients for month of birth being jointly equal to zero.

Table A1
Instrumental variable validity

	A. Log-Occupational Score, Natives			
	(a)	(b)	(c)	(d)
A_{it}	.014***		.015***	.015***
	(.004)		(.004)	(.004)
Scrabble Index		.004	−.001	−.001
		(.004)	(.004)	(.004)
Reads and Writes				.239***
				(.017)
R^2	.00	.00	.00	.00
N	109803	109803	109803	109803
	B. Scrabble index on indicators for			
	Height Deciles	Month of birth	Port of exit	All
$F-$test	0.58	0.39	1.18	1.12
$p-$value	0.78	0.97	0.16	0.24
R^2	.06	.06	.08	.08
N	4083	4083	4083	4083

NOTE.–Panel A. Source: 1930 Census. Sample is composed by U.S.-born males in labor force. Robust standard errors in parenthesis. Panel B. Dependent variable is the Scrabble index. Entries refer to the $F - andits$p- for the joint significance of the parameters of the variables indicated in the column headers.
*** $p < .01$.

Next, we check the correlation between the Scrabble index and port of emigration.[6] The latter variable proxies regional skill, abilities and motivational differences of migrants coming from the same country that we might not have been able to control for in our analysis. To

[6]We keep all port of exits and merge into a single category the ports with fewer than four observations.

understand the intuition, consider a migrant from Southern Italy, with lower socio-economic background and motivation to assimilate and compare him to a migrant from Northern Italy, with higher ability, motivation and earning potential in the U.S.. We assume that the migrant from Southern Italy is likely to emigrate from ports such as Naples or Palermo, while the migrant from Northern Italy is likely to leave from Genoa or Trieste. We also know that naming patterns in Southern Italy differ from naming patterns in Northern Italy. If names linguistic features were correlated with socio-economic status, we would expect the distribution of the Scrabble index to vary substantially within country and hence across ports of emigration. In the third column of this panel we show that the $F-$test fails to reject the null that all our indicators for port of exits are uncorrelated with the Scrabble index.

Finally, we report an $F-$test when all these variables are included in a regression to predict the Scrabble index. Once again, we fail to reject the null of no relationship between health, month of birth and port of exit with the Scrabble index. Taken together, the evidence from Table A1 provides further suggestive evidence that name structure might induce individuals to Americanize their name, although it is not directly associated with social status or labor market outcomes.

As a final check, we report in Table A2 a balancing test on the controls, separating our sample into those who have below and above median values of S^k_{Birth}. The table suggests that both groups of individuals are observationally similar. In particular, there are no differences in terms of years since migration, which means that the instrument is capable of purging any channel linked to human capital accumulation, including the acquisition of language skills.

IV results. Table A3 shows the instrumental variable results. In the first stage, we use the Scrabble index as a predictor of name Americanization. The first stage is reported in the bottom panel of Table A3. S^k_{Birth} is positively associated with name Americanization. This suggests that individuals with higher Scrabble points Americanize their names.[7] The instrument performs well irrespective of the controls added and remains a relevant predictor, as shown by the first stage $F-$statistics.[8] We also test whether there is evidence of endogeneity, rejecting exogeneity at the 10% significance level.

The first two columns show the payoff of name Americanization estimated from our baseline model. The last two columns report a robustness check where we consider name

[7]Full estimates about the first stage are available upon request. The most remarkable aspect is that, compared with the Germans, all nationalities with the exception of the Irish seem less likely to adopt American names over time. This seems consistent with evidence in Moser (2012) about increased discrimination towards the Germans following the First World War.

[8]Since our endogenous variable corresponds to $A_{iPet} - A_{iBirth}$, one might wonder whether the strength of the instrument might derive from a strong negative correlation between A_i at birth and the Scrabble index. Contrary to this conjecture, such correlation is weak and positive, ensuring the relevance of the instrument.

Table A2
Descriptive Statistics by Scrabble Index

	At Declaration			Difference petition-declaration		
Variable	≤ median	> median	t-test	≤ median	> median	t-test
Age	31.176	31.155	.937	5.161	5.157	.937
	(8.636)	(8.667)		(1.761)	(1.721)	
Years Since Migration	7.198	6.874	.153	4.695	4.710	.787
	(7.278)	(7.237)		(1.740)	(1.693)	
Married	.485	.465	.193	.226	.213	.330
	(.500)	(.499)		(.418)	(.410)	
Has U.S.-born spouse	.050	.052	.667	.057	.055	.841
	(.217)	(.223)		(.232)	(.229)	
Number of children	.952	.915	.444	.305	.291	.465
	(1.552)	(1.518)		(.629)	(.600)	
Has U.S.-born child(ren)	.223	.205	.150	.156	.154	.841
	(.417)	(.404)		(.376)	(.377)	
Resides outside N.Y. City	.133	.133	.995	-0.120	-0.121	.923
	(.340)	(.340)		(.343)	(.344)	
Year of arrival	1918	1918	.135			
	(6.753)	(6.760)				
Italy	.213	.221	.538			
	(.410)	(.415)				
Russian Empire	.186	.181	.662			
	(.389)	(.385)				
Central Europe (excl. DE)	.174	.162	.319			
	(.379)	(.369)				
Southern Europe (excl. IT)	.039	.025	.008			
	(.194)	(.155)				
Germany	.120	.094	.007			
	(.325)	(.292)				
Ireland	.030	.154	.000			
	(.172)	(.361)				
UK	.031	.047	.010			
	(.174)	(.212)				
Northern Europe	.105	.031	.000			
	(.307)	(.174)				
Americas	.056	.043	.044			
	(.231)	(.202)				
Other	.045	.043	.762			
	(.206)	(.202)				
N	2042	2041	4083	2042	2041	4083

NOTE.–Standard deviations in parentheses.

Americanization changes into a phonetically different name, i.e., changes in the "sound" of the name. To this aim, we use the NYSIIS algorithm. Names that sound and look similar are treated similarly and the instrument measures their complexity at arrival. The results indicate that returns to name Americanization remain positive after instrumenting for name Americanization. Overall, IV results do not reverse the main conclusions of the paper.

Table A3
The Effect of Name Americanization on Log-Occupational Score, Instrumental Variable Regression

	Without Controls	With Controls	Without Controls	With Controls
A_{it}	.559**	.500**		
	(.242)	(.227)		
NYSIIS of A_i			.776**	.722**
			(.339)	(.332)
	First stage			
S^k_{Birth}	.055***	.060***	.040***	.041***
	(.006)	(.006)	(.005)	(.005)
N	4083	4083	4083	4083
F 1^{st} stage	82.621	85.952	67.263	65.121
Partial R^2	0.019	0.022	0.016	0.016
Wooldridge test p-value	0.052	0.087	0.047	0.073
Pred. Occ. Score whole sample	0.032	0.032	0.031	0.031
Pred. Occ Score Americanizers	0.126	0.107	0.110	0.098

NOTE.–Robust standard errors in parenthesis. A_{it} = Americanization index, which varies between 0 (names with the lowest frequency) and 1 (names with the highest frequency). See text for explanation. NYSIIS of A_i= The New York State Identification and Intelligence System Phonetic Code (NYSIIS) algorithm is applied to the Americanization index. S^k_{Birth} is the Scrabble points of name k. The columns Without Controls include the variables of column (c) of Table 3. The columns With Controls include the variables of column (d) of Table 3. Wooldrdige test refers to a robust score test of endogeneity (Wooldridge, 1995).
** $p < .05$.
*** $p < .01$.

B. Additional robustness checks

In this section, we report additional robustness checks on our definitions. To obtain the occupational score, we first collected the occupation string from the naturalization papers and standardized occupation titles to match those identified in IPUMS. In the construction of our dataset we manually recoded occupations to reasonable occupational titles. When unsure about the category of the occupation, we used the Dictionary of Occupational Titles of 1949. While all occupations were standardized during this process, we flagged those for which some imputation was made for assignment into a certain category. We have performed additional checks to understand the sensitivity of our results to dropping flagged occupations. The first four columns of the top panel in Table B1 drop these flagged occupations and only report the results for the subsample of individuals for whom no imputation was necessary. Despite the smaller sample size (about two-thirds of the original sample), we obtain the same results and same patterns as in the baseline analysis.

A similar procedure was undertaken whenever addresses could not be located either due to changes in landscape or because street names were changed. The last four columns of the top panel of Table B1 show results for addresses that were correctly matched without resorting to any additional imputation. As before, our main conclusions remain unchanged.

The second panel of the table shows two additional checks. In the first four columns, we modify our Americanization index using the frequency of names in the U.S.-born population, aged 50+, resident in New York in 1930. In the second check we use a variant of the Americanization defined on the lines of the "black name index" (BNI) of Fryer and Levitt (2004). This is defined as $\frac{\sum_k \mathbb{1}(Name_{it}=Name_k)}{\sum_k \mathbb{1}(Name_{it}=Name_k)+\sum_l \mathbb{1}(Name_{it}=Name_l)}$ (for k in US-born and l in foreign-born), which varies between 0 and 1. Unlike our measure of name Americanization, which exclusively measures the popularity of American names, this is a relative index that is invariant to name popularity across minority groups. Both these alternative definitions of name Americanization reveal payoffs that are economically and statistically similar to our baseline estimates.

Table B1

The Effect of Name Americanization on Log-Occupational Score, Further Checks

	No Flagged Occupations				No Flagged Addresses			
	OLS	FD	NC	NB	OLS	FD	NC	NB
A_{it}	.035*	.126***	.235*	.223***	.031*	.125***	.258**	.233***
	(.020)	(.047)	(.138)	(.081)	(.017)	(.041)	(.125)	(.070)
N	6744	3372	1222	3372	7708	3854	1440	3854
	Names of 50+ U.S.-born				BNI			
	OLS	FD	NC	NB	OLS	FD	NC	NB
A_{it}50+/BNI	.020	.102***	.198*	.205***	.051***	.090***	.120*	.165***
	(.017)	(.037)	(.102)	(.065)	(.017)	(.025)	(.072)	(.049)
N	8166	4083	1538	4083	8166	4083	1538	4083

NOTE.–Robust standard errors in parenthesis. All models refer to the specification with all covariates in Table 3. OLS: pooled regression; FD: first-difference estimator; NC: name changers only; NB: name-at-birth fixed effects. For OLS N refers to the number of individual×period observations; for other models, N refers to the number of individuals. A_{it} = Americanization index, which varies between 0 (names with the lowest frequency) and 1 (names with the highest frequency). See text for explanation. No Flagged Occupations excludes migrants with occupations that were imputed when they could not be matched in the Dictionary of Occupational Titles of 1949. No Flagged Addresses exludes migrants with addresses that could not be located either due to changes in landscape or because street names were changed. Names of 50+ U.S.-born refers to the Americanization index constructed based on the U.S.-born population, aged 50+, resident in New York in 1930. BNI refers to an index alternative to A_i defined on the lines of that of Fryer and Levitt (2004). See text for details.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

C. Comparing 1930 naturalization records with the full population

Naturalized migrants are a subset of the foreign-born population. Naturalization rates in 1930 were higher than today. For instance, the share of naturalized foreign-born males, age 20+ living in the state of New York was 61.7%; the same share in the rest of the country was 61.4%, which is much higher than the 38.4% observed today.

While we expect the characteristics of our sample to be representative of the population of migrants who filed naturalization petitions in the Eastern and Southern districts of New York in 1930, there is no reason to expect that our sample is representative of the entire population of naturalized migrants or of foreign-born in 1930 New York or in the U.S. more generally. The 1930 census did not ask for the migrants' year of naturalization. Aiming to construct a comparable population to that of our sample, we limit ourselves to native males, aged 20+ and foreign-born males aged 20+ who arrived in 1920 or after, which corresponds to the median year of arrival in our sample.

Table C1 provides the summary characteristics for natives, aliens and foreign-born in New York and in the rest of the country. Several interesting patterns emerge.

First of all, naturalization rates for individuals arriving within the last ten years in the sample are lower than the same rate for the overall population: at about 29% in New York and 30% in the rest of the country. Second, about 57% of all naturalized migrants lived in New York state. Third, the average occupational scores between our sample and the naturalized population (both in NY and in the rest of the country) are economically very similar, and differences are statistically insignificant (at the 10% level). With the exception of the place of birth distribution for a few origin countries (Germany, Russia, Ireland), the naturalized migrants living in the rest of the country closely compare to those living in the state of New York and to our sample. We can therefore conclude that, at least in terms of average characteristics, our sample adequately captures the naturalized foreign-born population in the U.S..

Differences in average characteristics are more marked when comparing non-naturalized with naturalized migrants. Unsurprisingly, non-naturalized foreign born have lower socio-economic standing than naturalized citizens. While our results cannot be fully extended to the overall migrant population, migrants in our sample remain an interesting group, having arrived during the last surge of migration before U.S. doors were shut, and permanently settled, truly contributing to 'the making of modern America'.

In unreported results, using the same sample restrictions as above, we also run an OLS regression of the Americanization index on log-occupational score using census data. While this analysis is completely silent on the magnitude, consequences and selection patterns related to name Americanization, it is indicative of whether the baseline findings observed in our main analysis are similar to what we would observe in the foreign-born population as a whole. The point estimate from estimating this regression is 0.048 (s.e. 0.005) as compared with the point estimate of 0.045 (s.e. 0.027) using our sample. This analysis reassures that the baseline trends in our sample are also found in the overall population.

Table C1
Descriptives, IPUMS Comparison, 1930

		In New York			Outside New York		
	Sample	Natives	Aliens	Naturalized	Natives	Aliens	Naturalized
Occupational Score	26.229	27.731	24.376	26.275	23.727	23.085	25.830
	(8.699)	(10.637)	(8.420)	(9.194)	(11.043)	(8.720)	(8.939)
Age	34.278	38.187	31.979	33.171	39.301	32.666	33.655
	(8.154)	(13.503)	(9.065)	(8.826)	(13.804)	(9.620)	(9.111)
Married	.631	.637	.528	.671	.711	.555	.664
	(.483)	(.481)	(.499)	(.470)	(.453)	(.497)	(.472)
Years since migration	7.417		5.610	7.840		5.923	7.771
	(1.486)		(2.925)	(2.250)		(2.853)	(2.216)
Italy	.193	-	.154	.217	-	.089	.188
	(.395)	-	(.361)	(.412)	-	(.285)	(.391)
Russian Empire	.138	-	.059	.143	-	.029	.073
	(.345)	-	(.235)	(.350)	-	(.166)	(.260)
Central Europe (excl. DE)	.143	-	.094	.221	-	.082	.171
	(.350)	-	(.292)	(.415)	-	(.274)	(.377)
Southern Europe (excl. IT)	.025	-	.041	.027	-	.043	.036
	(.155)	-	(.199)	(.163)	-	(.203)	(.187)
Germany	.148	-	.181	.084	-	.115	.096
	(.355)	-	(.385)	(.277)	-	(.319)	(.294)
Ireland	.127	-	.117	.082	-	.047	.057
	(.333)	-	(.322)	(.274)	-	(.211)	(.231)
UK	.047	-	.088	.059	-	.115	.119
	(.211)	-	(.283)	(.236)	-	(.320)	(.323)
Northern Europe	.092	-	.075	.037	-	.085	.074
	(.289)	-	(.263)	(.188)	-	(.279)	(.262)
Americas	.052	-	.134	.093	-	.314	.107
	(.222)	-	(.341)	(.290)	-	(.464)	(.309)
Other	.036	-	.058	.039	-	.081	.079
	(.187)	-	(.234)	(.193)	-	(.273)	(.270)
N	2674	116097	13050	5425	1296052	32711	10040

NOTE.–Standard deviations in parentheses. Sources: Ancestry.com (Sample) and 1930 census (Natives, Aliens, Naturalized), only male migrants, aged 20+, arrived in 1920 or after.

D. Geocoding Procedure and Local Labor Markets Definition

The geocoding procedure to derive coordinates comprises the following steps:

1. Cataloguing the address at petition and declaration. The address typically contains the house number, the street name/number, and the name of the New York Borough.

2. Matching historical addresses with current addresses. Using a routine that scrapes addresses from Google Maps, we match the street name and house number at declaration (petition) with current street name and house number.

3. Retrieving missing addresses. The reasons for the cases in which the historical address could not be matched with the current address were either because the address no longer exists (e.g., due to a change in landscape) or because the street name/number had changed. The unmatched addresses were retrieved by relying on supplementary information on changes in streets names and house numbers available in websites and books. We mainly rely on the "One-step" web tools made freely available by Stephen Morse and colleagues (http://stevemorse.org/) and on the website http://www.oldstreets.com/. The majority of address changes are located in the Queens borough. For streets that were truncated or that have disappeared, a nearby address (within the same 5-digit postcode) was imputed. We flagged all cases in which the address could not be automatically matched through Google Maps. These correspond to 6.8% of those at declaration and 4.4% of those at petition. The exclusion of these from the sample does not change our results. At declaration, about 11% of immigrants resided in a state different than New York, while about 1% lived in the state of New York but outside New York City. We code their addresses as "Outside NYC". At petition, very few migrants live outside the city. Finally, 1.8% of addresses at declaration and 1.4% of addresses at petition could not be found, and thus corresponding observations have been excluded.

4. Obtaining coordinates. We also extract the coordinates of the address and the related postcode. While coordinates are ultimately used for mapping individuals, they also enable an important qualitative check of our geocoding procedure (e.g., whether the current matched address is in the borough reported in the petition/declaration).

5. Mapping. We use the coordinates of migrants addresses to map their location within areas of interest. For example, we are able to identify the census tract, community district or 5-digit zip-code where they live. Mapping also provides a useful visual representation of the distribution of migrants in New York City and areas within.

Local labor markets are defined by using information from the 5% census in 1930. We use this year since the sample is sufficiently large (compared to 1% in 1920) to explore the finest possible definition of local labor market. In general, the definition of a local labor market is somewhat arbitrary. One could think of it as a self-contained area in which sub-areas within are similar in terms of labor market characteristics (such as industries, occupations, employment) or patterns of commuting (i.e., the majority of people work and live in the same area). An additional challenge is to find a definition of local labor market that fits with the situation of New York in 1930. To address this issue, we ask the data about the best definition. We explore several candidates, from the highest to the lowest level of disaggregation: census enumeration districts, census tracts, community districts and New York boroughs. With a few exceptions, these areas have the feature of being fully self-contained across classifications (e.g., a finite number of enumeration districts form a tract; a finite number of tracts form a community district). Each classification has advantages and disadvantages for being used as a proxy for a local labor market. For example, the advantage of census enumeration districts is that they are numerous and hence allow us to control for very local characteristics, although their small size is prone to measurement error. On the other hand, the New York boroughs are easy to identify and there is virtually no margin of error in geocoding the location of immigrants in terms of New York boroughs; however, they might be too aggregated for a definition of "local" labor market at a time when average commuting distance was somewhat limited due to high transportation costs.

Modeling the data would inform us on the best definition. We proceed by modeling key variables that define a local labor market, focusing our attention on the unemployment rate, occupation score and a continuous variable for the industry code. We estimate OLS regressions on the individual probability of being unemployed, individual occupational scores and industry codes on a set of standard regressors (age, sex, number of children, race, country of birth, and year of immigration – for the immigrants), including indicators for the various types of areas listed above. Our aim is to find which model best "explains" labor market outcomes. To assess the best fit across various specifications, we compare their Bayesian Information Criterion (BIC), which offers the appealing feature of minimizing the log-likelihood function and penalizing the number of parameters included in the model. Our preferred choice for local labor market coincides with community districts, as these are associated with the lowest BIC across the various models for unemployment, occupation scores and industry. The results are reported in Table D1.

Exploiting the detailed geographical information in the dataset, we present the migrants' geographical distribution in Figure D1, using immigrants' place of residence at declaration and focusing on New York City. It is evident that Brooklyn and Manhattan were the two

Table D1
Defining local labor market: Tests

Dep. Variable	Unemployment			Occupational score			Industry		
Lab. Mk.	Log-l.	df	BIC	Log-l.	df	BIC	Log-l.	df	BIC
NA	-28407.9	121	58215.97	-435307.1	120	872010	-778973.3	121	1559354
BO	-28368.3	125	58183.11	-435254	125	871962.1	-778782	125	1559018
CD	-27884.4	187	57932.69	-434274.9	187	870724.9	-778184.3	186	1558532
TR	-26336.9	1272	67392.93	-432871.2	1275	880572.9	-777297.8	1275	1569426
ED	-21618.7	4759	98307.04	-428262.5	4773	912043.6	-773328.1	4773	1602175
Obs	106053			112622			112622		

NOTE.–Robust standard errors in parentheses. Columns headers indicate the dependent variable of each regression. Regressors included are: age, sex, number of children, race, country of birth, and year of immigration for the immigrants. NA: No areas; BO: NYC Boroughs; CD: Community Districts; TR: Tracts; ED: Enumeration Districts.

most populated boroughs. Figure D2 distinguishes the place of residence by major country of birth. Although some clusters by country of origin are evident, including those of Italians in the lower east side of Manhattan and Russians in the Brownsville area of Brooklyn, different ethnic groups co-existed in the same areas.

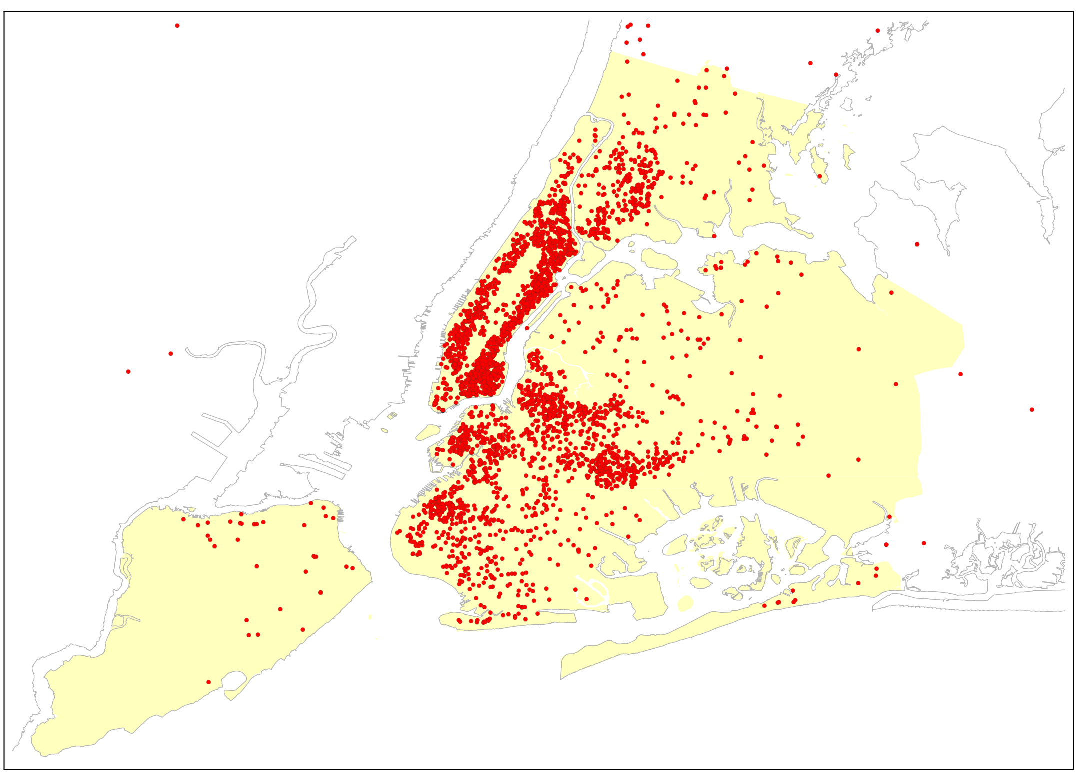

FIG. D1. Immigrants in New York City, at declaration

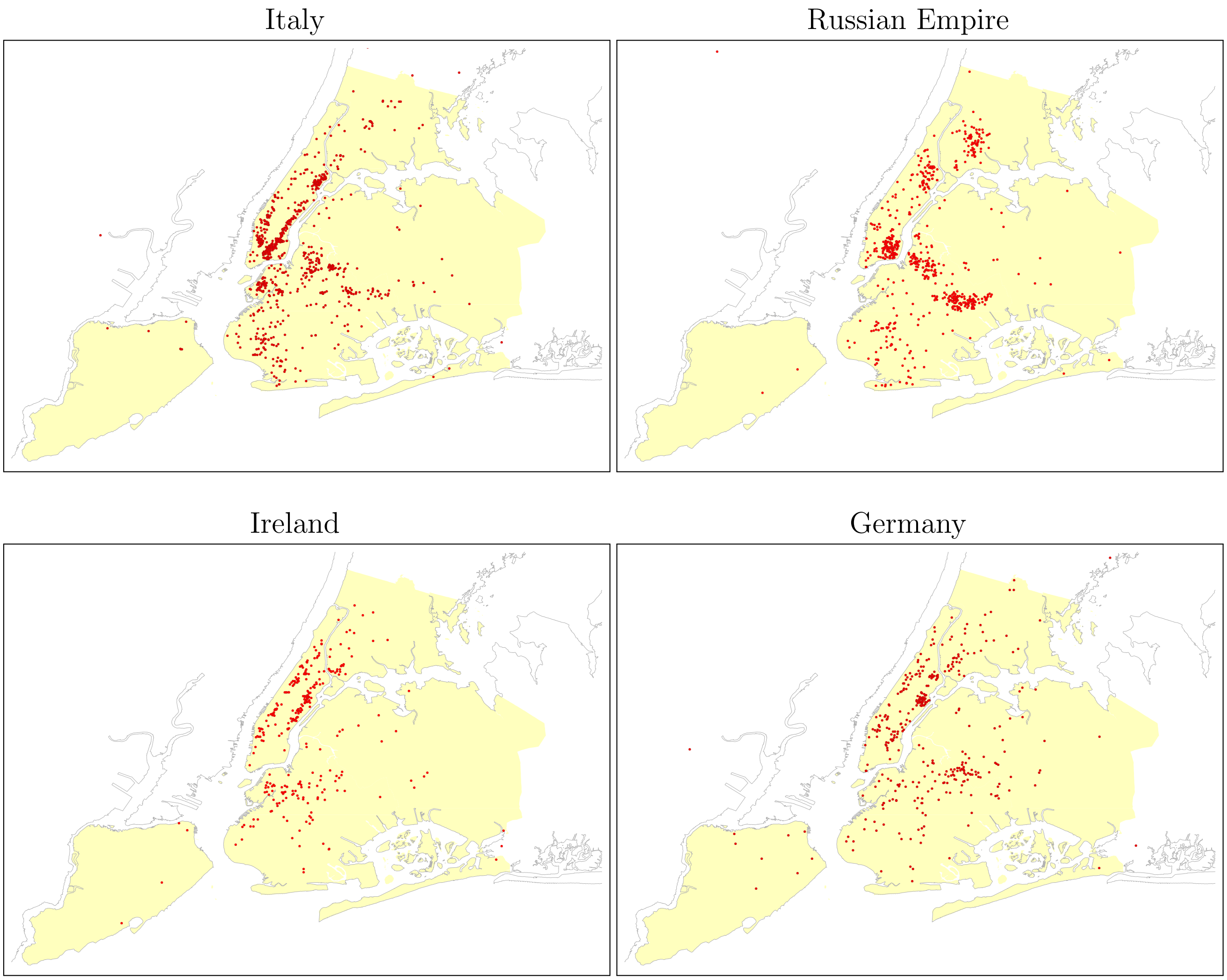

FIG. D2. Immigrants in New York City, at declaration: by country of birth

E. Full summary statistics

Table E1

Characteristics by Level of Americanization

Variable	All	Keepers	Others	Americanize	1st	2nd	3rd
	At Declaration						
Log Occupational Score	3.177	3.173	3.158	3.188	3.174	3.220	3.172
	(.436)	(.399)	(.444)	(.499)	(.510)	(.495)	(.492)
Americanization Index	.145	.113	.066	.225	.024	.104	.549
	(.261)	(.237)	(.187)	(.297)	(.054)	(.071)	(.311)
Age	31.166	31.484	30.352	30.696	30.886	30.156	31.028
	(8.651)	(8.584)	(7.962)	(8.883)	(9.322)	(8.614)	(8.674)
Years Since Migration	7.036	5.956	6.470	9.284	8.806	9.550	9.514
	(7.258)	(6.826)	(6.584)	(7.696)	(7.748)	(7.778)	(7.556)
Married	.475	.469	.451	.490	.474	.481	.516
	(.499)	(.499)	(.499)	(.500)	(.500)	(.500)	(.500)
Has U.S.-Born Spouse	.051	.044	.044	.066	.032	.093	.075
	(.220)	(.205)	(.204)	(.249)	(.176)	(.291)	(.263)
Number of Children	.934	.938	.826	.946	1.000	.854	.981
	(1.535)	(1.552)	(1.538)	(1.500)	(1.575)	(1.380)	(1.533)
Has U.S.-Born Child(ren)	.214	.195	.166	.262	.232	.268	.285
	(.410)	(.396)	(.373)	(.440)	(.423)	(.443)	(.452)
Resides outside N.Y. City	.133	.154	.111	.097	.112	.072	.105
	(.340)	(.361)	(.314)	(.295)	(.315)	(.258)	(.307)
Arrival Cohort	1918	1919	1919	1916	1917	1916	1916
	(6.757)	(6.357)	(6.103)	(7.154)	(7.109)	(7.314)	(7.037)
Italy	.217	.259	.202	.137	.084	.091	.236
	(.412)	(.438)	(.402)	(.344)	(.278)	(.288)	(.425)
Russian Empire	.183	.102	.229	.335	.403	.404	.199
	(.387)	(.303)	(.421)	(.472)	(.491)	(.491)	(.399)
Central Europe (excl. DE)	.168	.103	.237	.283	.330	.254	.264
	(.374)	(.304)	(.426)	(.451)	(.471)	(.436)	(.441)
Southern Europe (excl. IT)	.032	.029	.032	.038	.030	.041	.044
	(.176)	(.167)	(.175)	(.192)	(.170)	(.198)	(.206)
Germany	.107	.116	.138	.084	.025	.098	.131
	(.309)	(.320)	(.346)	(.278)	(.157)	(.298)	(.338)
Ireland	.092	.143	.024	.004	.005	.005	.002
	(.289)	(.351)	(.153)	(.062)	(.067)	(.069)	(.048)
U.K.	.039	.057	.024	.006	.000	.012	.007
	(.194)	(.233)	(.153)	(.079)	.000	(.109)	(.084)
Northern Europe	.068	.085	.040	.041	.064	.031	.028
	(.252)	(.279)	(.195)	(.199)	(.245)	(.174)	(.165)
Americas	.050	.068	.032	.017	.018	.017	.016
	(.217)	(.251)	(.175)	(.130)	(.134)	(.129)	(.127)
Other	.044	.039	.044	.054	.041	.048	.072
	(.204)	(.193)	(.204)	(.226)	(.199)	(.214)	(.260)
	Difference Petition-Declaration						
Log Occupational Score	.031	.016	.026	.061	.046	.060	.077
	(.478)	(.435)	(.485)	(.552)	(.652)	(.491)	(.493)
Americanization Index	.077	.000	-.037	.251	.020	.114	.621
	(.210)	(.000)	(.113)	(.304)	(.019)	(.055)	(.255)
Age	5.159	5.202	5.466	5.013	5.100	4.849	5.082
	(1.741)	(1.712)	(1.818)	(1.771)	(1.797)	(1.764)	(1.744)
Years Since Migration	4.702	4.743	5.028	4.556	4.645	4.368	4.650
	(1.716)	(1.689)	(1.763)	(1.748)	(1.800)	(1.711)	(1.719)
Married	.219	.219	.300	.204	.210	.220	.182
	(.414)	(.414)	(.459)	(.403)	(.408)	(.415)	(.387)
Has U.S.-Born Spouse	.056	.049	.079	.066	.077	.060	.061
	(.230)	(.215)	(.270)	(.249)	(.268)	(.237)	(.239)
Number of Children	.298	.298	.344	.288	.287	.275	.301
	(.615)	(.631)	(.601)	(.584)	(.604)	(.517)	(.624)
Has U.S.-Born Child(ren)	.155	.144	.233	.161	.166	.179	.138
	(.377)	(.368)	(.424)	(.382)	(.379)	(.384)	(.384)
Resides outside N.Y. City	-.121	-.140	-.091	-.088	-.107	-.067	-.089
	(.343)	(.364)	(.339)	(.297)	(.324)	(.260)	(.301)
N	4083	2545	253	1285	439	418	428

NOTE.–Standard deviations in parentheses. “Keepers” are migrants whose index did not change over time (i.e., they did not change name or changed into equally frequent names). “Others” include migrants whose index change was negative (i.e., changed into less frequent names) and migrants whose index changed twice in the period of observation. “Americanize” refers to migrants who changed names into more frequent names. 1st, 2nd and 3rd refer to terciles of the change in the Americanization index as defined in the text. Married, Has U.S.-born Spouse, Has U.S.-born Child(ren), Resides outside N.Y. City are all indicators. “Russian Empire” includes migrants born in: Estonia, Latvia, Lithuania, Armenia, U.S.S.R.. “Central Europe (excl. DE)” includes migrants born in: Austria, Austria-Hungary, Bulgaria, Czechoslovakia, Bohemia, Prussia, Hungary, Poland, Austrian Poland, Russian Poland, Romania, Yugoslavia, Montenegro. “Southern Europe (excl. IT)” includes migrants born in: Albania, Greece, Malta, Portugal, Azores, Spain. “U.K.” includes migrants born in: Scotland, Wales and the United Kingdom. “Northern Europe” includes migrants born in: Denmark, Finland, Norway, Sweden. “Americas” includes migrants born in: New York, Ohio, Pennsylvania, U.S. Virgin Islands, Canada, Nova Scotia, Newfoundland, Bermuda, Mexico, Costa Rica, Guatemala, Cuba, Jamaica, British West Indies, Antigua-Barbuda, Barbados, Tortola, Montserrat, Netherlands Antilles, Argentina, Brazil, Colombia, Guyana/British Guiana, Peru. New York, Ohio, Pennsylvania refer to three cases of second generation U.S. citizens who either lost citizenship due to naturalization in a foreign state or as a consequence of military service in a foreign state. “Other” includes migrants born in: Belgium, France, Alsace-Lorraine, Netherlands, Switzerland, East Indies, Philippines, India, Cyprus, Palestine, Syria, Turkey, Algeria, Tunisia, Australia.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91, no. 434:444–455.

Aura, Saku and Gregory D. Hess. 2010. What's in a name? *Economic Inquiry* 48, no. 1:214–227.

Bertrand, Marianne and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review* 94, no. 4:991–1013.

Buckles, Kasey S. and Daniel M. Hungerman. 2013. Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics* 95, no. 3:711–724.

Fryer, Roland G. and Steven D. Levitt. 2004. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics* 119, no. 3:767–805.

Morgan, Stephen L. and Christopher Winship. 2014. *Counterfactuals and Causal Inference.* Cambridge University Press.

Moser, Petra. 2012. Taste-based discrimination: Evidence from a shift in ethnic preferences after WWI. *Explorations in Economic History* 49, no. 2:167–188.

Wooldridge, Jeffrey M. 1995. Score diagnostics for linear models estimated by two stage least squares. In *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, eds. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, chap. 3. Oxford: Blackwell, 66–87.