A NATIONS OF EMIGRANTS: TRANSATLANTIC EMIGRATION AND ITALIAN NATIONAL IDENTITY *

Davide M. Coluccia† Giampaolo Lecce‡ Laura Ogliari§

This version: March, 2026

Preliminary and incomplete. Please do not circulate.

Abstract

We study how intergroup contact among immigrants shapes national identity formation within migrant communities and in their places of origin. We assemble a novel dataset on Italian migrants in the U.S. between 1892 and 1924 by linking population census records to Ellis Island ship passenger lists, which contain migrants' municipalities of origin. In the U.S., exposure to more diverse Italian communities fostered the emergence of a shared national identity, as measured by a novel index of national naming choices and intergroup marriages, while reducing assimilation into American society. In Italy, municipalities of origin of migrants who have been exposed to more diverse communities abroad exhibit lower cultural particularism. Our findings document the novel nation-building effect of intergroup contact within immigrant communities.

Keywords: National Identity, Migration, Age of Mass Migration
JEL Classification: J15, N32, N33, Z13

* We are grateful to Philipp Ager, Elliott Ash, Toman Barsbai, Sascha Becker, Julia Cagé, Eleonora Guarnieri, Yanos Zylberberg, and to seminar audiences at the University of Bergamo and the University of Bristol for useful feedback. This research received financial support from the University of Bristol. Any error remains our own.

†University of Bristol. Email: davide.coluccia@bristol.ac.uk. Website: dcoluccia.github.io.

‡University of Bergamo and CEPR. Email: giampaolo.lecce@unibg.it. Website: sites.google.com/giampaololecce.

§University of Bergamo and CEPR. Email: laura.ogliari@uning.it. Website: sites.google.com/laura-ogliari.

"The Italians Abroad to the Fatherland"
– Altar of the Fatherland, Rome[1]

INTRODUCTION

The rise of nation-states fundamentally transformed the scale at which political and economic life is organized. A prominent view in economics and political science holds that national identities arise from intentional state interventions, such as mass schooling, bureaucratic centralization, military conscription, and propaganda deployed by political elites (Mylonas and Tudor, 2021; Rohner and Zhuravskaya, 2024). But can shared national identities instead emerge from decentralized social interactions in the absence of coordinated state intervention?

We examine this question in the context of the Italian migration to the United States during the Age of Mass Migration (1892–1924). During this period, Italy was a recently unified country characterized by deep linguistic and cultural fragmentation, famously described by Austrian Chancellor Klemens von Metternich as little more than a "geographical expression." Italian migrants arriving in the United States were highly heterogeneous and embedded in local identities shaped by place of origin, language, and customs. The 4.5 million Italians who settled across American cities at the turn of the 20th century encountered and interacted with one another in shared urban environments. This context thus provides a unique setting to examine whether decentralized social interactions among heterogeneous individuals can foster the emergence of a shared national identity.

We construct a novel dataset linking U.S. census records of Italian immigrants to ship manifests recorded at the Ellis Island immigration station. This linkage allows us to trace the municipality of origin for nearly half a million migrants, yielding an unprecedentedly detailed view of the within-country composition of an immigrant community.

Our first finding documents that Italian immigrants exposed to more diverse Italian communities in the U.S. were more likely to abandon local identities and adopt a "national" one at the expense of slower cultural assimilation into American society. Migrants living among Italians from different parts of the country give their children names that are less characteristic of local identity—our novel measure of cultural particularism—and exhibit a higher probability of marrying migrants from different municipalities. At the same time, we document slower assimilation as measured by the Foreign Name Index of their children (Abramitzky et al., 2020). In the second part of the paper, we show that identity formation within migrant communities had lasting effects on their places of origin. Italian municipalities whose emigrants were exposed to more diverse communities in the United States exhibit lower name particularism and greater commemoration of national events, as reflected in street-naming patterns and in monument construction. Moreover, exposure to diversity abroad is strongly associated with higher World War I death rates among returning migrants, which likely indicates higher volunteer rates among migrants from those municipalities. Taken together, these findings suggest that decentralized social interactions among migrants can foster the consolidation

[1]The original epigraph reads *"Gli Italiani all'estero alla Madrepatria"* and is engraved on the two ceremonial braziers of the *Vittoriano* monument in Rome, whose perpetual flames symbolize national unity.

of national identity and reshape both host communities and societies of origin.

Between 1892 and 1924, approximately 4.5 million Italian migrants settled in the U.S. (Choate, 2008). These communities reflected the deep cultural heterogeneity of Italy itself (Banti, 2014). Migrants clustered along their local origins, creating "Little Palermos" and "Little Naples" within "Little Italies" (Gabaccia, 2013). Over time, however, interaction across local groups fostered forms of collective identity (Vecoli, 1983; Orsi, 1985). While previous literature largely focused on immigrants' assimilation into American society, historical accounts emphasize that Italians developed identities distinct from both native-born Americans and their communities of origin. Given that more than half of Italian migrants eventually returned, this identity shift plausibly affected migrants' places of origin (Luconi, 2011; Chiaricati, 2020; Mourlane et al., 2022).

Despite the historical accounts, there is little systematic evidence on whether interactions among migrants generated national identity, or whether such identity formation had measurable consequences for their places of origin. Existing data sources are not well-suited to addressing this question, as they do not provide information on migrants' municipalities of origin and their settlement locations in the U.S. To overcome this limitation, we develop an algorithm that links U.S. census records to ship passenger lists recorded at the Ellis Island immigration station, which report migrants' municipalities of origin. Our approach augments standard record-linkage methods (Abramitzky et al., 2021) with neural-network embeddings to flexibly match names that were frequently Anglicized after arrival. In our preferred specification, we link approximately 470,000 migrants, equivalent to a 20% matching rate. We assess the reliability of the linked dataset in three ways. First, linked migrants constitute an observationally representative sample of the full population. Second, 90% of the established links are classified as highly likely by a large language model. Third, 95% of a random subsample of links are independently validated by human reviewers.

We use this novel dataset to test two hypotheses. First, we examine whether exposure to Italians from different parts of the country shaped the formation of national identity in the United States. Second, we study whether inter-group contact among migrants influenced the formation of national identity in their communities of origin.

We conduct an individual-level analysis to estimate how exposure to a more diverse Italian community within an immigrant's neighborhood affects national identity and assimilation. To measure national identity, we exploit the linguistic and cultural divides within Italy documented by De Mauro (1963) to construct a novel index of name particularism, following recent work (e.g., see Bazzi et al., 2019; Fouka, 2020). Names that are geographically concentrated reflect stronger local identities, whereas names that are more evenly distributed across the country capture national identity. We complement this measure with information on marriage patterns with individuals from other municipalities, provinces, and regions. We measure assimilation into American society using the name-based indicator developed by Abramitzky et al. (2020).

Estimating the effect of neighborhood diversity on cultural identity raises two distinct identification challenges. First, local diversity may reflect the selective sorting of immigrants from specific municipalities across locations in the United States. Second, local diversity may be correlated with

omitted, time-varying local factors that independently influence identity formation and assimilation.

We address this concern using two approaches. First, we construct an instrumental variable that leverages variation in ship composition, following the logic of Dippel and Heblich (2021). Under the hypothesis that interactions during the transatlantic voyage shaped subsequent settlement patterns, quasi-random variation in ship composition generates exogenous differences in immigrants' exposure to diversity upon arrival. Individuals traveling on more diverse ships are more likely to settle near their shipmates, and therefore to be exposed to more diverse Italian neighborhoods. Empirically, we show that ship-level diversity is a strong predictor of neighborhood diversity and is uncorrelated with observable individual and local characteristics, suggesting that the instrument isolates plausibly exogenous variation in exposure to within-Italian diversity.

Second, we implement a "movers" design in the spirit of Chetty and Hendren (2018). We estimate a difference-in-differences specification that compares individuals who move to more diverse neighborhoods with non-movers before and after the move. This design allows us to assess the parallel trends assumption by showing that movers and non-movers exhibit statistically indistinguishable trends in identity outcomes before relocation.

We find that exposure to more diverse Italian communities decreases name particularism—our measure of national identity—and reduces Americanization. Italians exposed to more diverse immigrant communities are more likely to marry individuals born in other municipalities and provinces, and less likely to marry natives. Individuals who settle in more versus less diverse neighborhoods are similar upon arrival, but diverge sharply in national and American identity as they spend time in the United States. Quantitatively, ten years after arrival, immigrants exposed to more diverse neighborhoods exhibit a 10% increase in national identity and show no evidence of assimilation into American society. In contrast, immigrants exposed to less diverse neighborhoods do not experience an increase in national identity and give their children 9% more American-sounding names.[2]

These findings suggest that immigrants in more diverse communities developed a shared Italian group identity. A natural implication is that such immigrants would be less likely to return to Italy, where local identities remained more salient. We test this implication using a newly developed estimator of return migration that leverages longitudinal census links (Price et al., 2021; Abramitzky et al., 2025). We find that Italian immigrants residing in more diverse communities are approximately 15% more likely to remain in the United States than otherwise comparable immigrants in less diverse areas. Therefore, while roughly half of Italian migrants returned to Italy overall, only about 35% of migrants exposed to highly diverse Italian communities did so. These results suggest that exposure to diversity fostered a distinct form of national identity, distinct from both local particularism in Italy and assimilation into U.S. society.

In the second part of the paper, we examine whether the national identity that formed within immigrant communities spilled over to migrants' places of origin in Italy. We use three indicators to

[2]We formalize these results in a model of immigrant identity choice along the lines of Akerlof and Kranton (2000) and Shayo (2009). Under mild and plausible assumptions, this simple framework, displayed in Appendix Section D, yields predictions that closely align with our empirical findings.

proxy for the cultural relevance of national identity at the local level: naming patterns, street naming choices, and the construction of national monuments.

Using given names as our primary proxy, we show that Italian municipalities whose migrants were more exposed to diverse Italian communities in the United States experienced a persistent decline in name-based particularism. To address the absence of detailed historical microdata on names, we draw on individual-level casualty records for soldiers who died during World War 1 and on a newly constructed dataset of World War 2 casualties. In a difference-in-differences setting, we find that municipalities whose migrants were exposed to higher diversity in the United States shifted from locally concentrated names toward nationally shared names, consistent with the transmission of national identity from migrant communities back to their places of origin.

We corroborate these patterns using street naming as an alternative proxy for national identity. Municipalities with greater historical exposure to diversity among their U.S. emigrants are significantly more likely to commemorate national figures and events. Conversely, we find no systematic relationship between emigrants' diversity exposure and placebo categories such as religion, art, or literature. This contrast suggests that the results are not driven by a general propensity for commemoration or naming activity.

Third, we examine public monuments as a third, long-run indicator of national sentiment. Focusing on Southern municipalities, we find that exposure to diversity through U.S. migration predicts a higher incidence of monuments commemorating national events, but only in the decades following mass emigration. Exposure to diversity, by contrast, is not associated with national monuments in pre-migration decades, nor with monuments that do not specifically commemorate national events.

Fourth, we document a positive correlation between municipality-level death rates of U.S. emigrants serving as Italian soldiers during World War I and exposure to diversity in the United States. Since the enlistment of migrants in the United States by Italian authorities was largely voluntary, we interpret this pattern as consistent with stronger identification with the Italian nation among migrants who had been exposed to more diverse environments abroad.

Taken together, evidence from names, streets, monuments, and military enlistment consistently indicates that interaction among diverse Italian migrants in the United States contributed to the emergence and diffusion of national identity in Italy.

Benedict Anderson (1983) argues that nations are "imagined communities" formed through shared representations that do not rely on interpersonal contact among their members. This paper documents a complementary perspective: social interactions *can* foster national identity among fragmented groups and, specifically, migrant group identity contributes to national cohesion in their country of origin. Our findings suggest that today's diasporic communities—often originating from highly fractionalized societies (Alesina et al., 2003)—can play an important and underappreciated role in fostering nation-building in their countries of origin.

Contributions to the Literature. First, this paper contributes to a growing body of work on the economics of nation-building (Rohner and Zhuravskaya, 2024).[3] A unifying approach in this literature is that national identity emerges in response to top-down interventions, including mass education (Cantoni et al., 2017; Bandiera et al., 2019; Alesina et al., 2021; Blanc and Kubo, 2025), foreign state intervention (Dell and Querubin, 2018; Dehdari and Gehring, 2022), media propaganda (DellaVigna et al., 2014; Esposito et al., 2023), welfare spending (Caprettini and Voth, 2023), forced population relocation (Bazzi et al., 2019), and military conflict (Lecce et al., 2022). Our contribution is to study a setting in which individuals from highly heterogeneous communities were largely insulated from coordinated policies. In this sense, our findings are related to Depetris-Chauvin et al. (2020), who show that shared experiences contribute to national cohesion, and Cagé et al. (2023), who link social interactions with charismatic leaders to the rise of nationalism. We provide novel evidence that sustained inter-group contact, even in the absence of explicit state intervention, can foster the formation of national identity.

Second, we contribute to the literature on immigrant assimilation (e.g., Abramitzky et al., 2020; Fouka et al., 2022; Carlitz et al., 2025; Gagliarducci and Tabellini, 2025). A common focus of this work is on how minority groups assimilate into a majority (native) population. We take a different perspective by studying how exposure to migrants from the same origin country, but from different regions, shapes group identity formation.[4] Guided by a simple model of identity choice in the spirit of Akerlof and Kranton (2000), we show that greater within-origin diversity among migrants strengthens national identity and slows cultural assimilation into the host society.

Third, our findings relate to the literature on how inter-group contact shapes identity. Most studies in this literature focus on majority–minority contact. For example, Bazzi et al. (2019) study the nation-building effects of an extensive relocation program in Indonesia, while Bursztyn et al. (2024) show that proximity to immigrants reduces hostility toward subsequent immigration. Related work examines other settings, such as workplace and educational interaction (Billings et al., 2021; Corno et al., 2022) and ethnic or political cleavages (Dahl et al., 2021; Schindler and Westcott, 2021; Okunogbe, 2024). Building on the contact hypothesis, Lowe (2021) distinguishes between collaborative and adversarial contact and shows that only the former fosters integration. Our work is related to Bagues and Roth (2023), who show that exposure to interregional diversity in Spain fosters national identity formation. We inform this literature by showing that *within*-minority interactions (among migrants) can shape group identity formation and reallocate identity away from majority–minority assimilation.

Fourth, we contribute to the literature on the effects of out-migration on emigration countries (Batista et al., 2025). A growing body of work shows that emigrants can benefit their places of ori-

[3] The economics of nation-building is a relatively recent field, but it builds on an established literature at the intersection of sociology, political science, and history. "Modernist" theses argue that nation-states emerged in response to industrialization (Gellner, 1983; Anderson, 1983), whereas Smith (1986) maintains that modern nations reflect pre-modern identities. Classic case studies of top-down nation-building processes within this tradition include Weber (1976) on education in France and Hobsbawm and Ranger (1983) on the use of invented tradition as a legitimizing device for national identity.

[4] In this sense, this work is related to the multidisciplinary field of diaspora studies (among others, see Gilroy, 1993; Edwards, 2003). Hall (2015) emphasizes that cultural identity in diasporic communities is produced in response to displacement and internal heterogeneity. Our findings echo these qualitative narratives, indicating that Italians in the United States developed an identity that is partially distinct from the local identity they acquired in Italy.

gin through monetary remittances (Khanna et al., 2026), the diffusion of new technologies (Coluccia and Dossi, 2025), increased technology adoption and entrepreneurship (Anelli et al., 2023; Coluccia and Spadavecchia, 2025), and human capital formation (Fernández Sánchez, 2022). Recent studies, in particular, emphasize the political and social consequences of emigration, documenting how migrants diffuse social norms (Levitt, 1998; Levitt and Lamba-Nieves, 2011; Beine et al., 2013; Melki et al., 2024; Godlonton and Theoharides, 2025) and political preferences (Spilimbergo, 2009; Barsbai et al., 2017; Karadja and Prawitz, 2019; Manacorda et al., 2024) to origin communities, and contribute to increased collectivism among stayers (Knudsen, 2024). We document a novel effect of emigration: its contribution to the formation of national identity in migrants' countries of origin.

Outline of the Paper. Section I provides a concise overview of the historical background. Section II introduces our novel dataset and the additional data used in the analysis. Section III focuses on the United States and documents that Italian immigrants developed a distinctive group identity, which we refer to as "national identity." Section IV examines how historical out-migration contributed to the rise of national identity in migrants' communities of origin. Section V concludes.

I HISTORICAL BACKGROUND

Section I.A outlines Italian mass migration to the United States and its regional composition. Section I.B describes Italy's cultural fragmentation, while Section I.C documents cultural convergence within Italian immigrant communities in the United States and its feedback effects on Italy. Additional historical detail and rich anecdotal evidence are provided in Appendix Section C.

I.A The Italian Mass Migration to the United States, 1892–1924

Between 1876 and the early 1920s, approximately 26 million Italians emigrated abroad, constituting the largest episode of voluntary emigration in recorded history (Choate, 2008). By 1913, Italy was the world's second-largest source of emigrants, and its emigration rate was comparable to that of high-emigration countries such as Ireland and Norway in earlier decades (Foerster, 1919). The United States absorbed roughly 40 percent of this flow (Livi-Bacci, 1961). Migration was facilitated by liberal migration regimes on both sides of the Atlantic and remained largely unrestricted until the introduction of U.S. literacy tests in 1917 and quota restrictions in 1921–1924, which abruptly curtailed Italian inflows (Abramitzky and Boustan, 2017).

Italian migrants to the United States were predominantly young men from rural areas of Southern Italy (Foerster, 1919). Although migrants were negatively selected at the national level, they were positively selected within origin communities in terms of height, education, and wealth (Spitzer and Zimran, 2018). Migration was frequently temporary: more than half of Italian migrants eventually returned (Bandiera et al., 2013). Both historical narratives and recent empirical evidence emphasize the central role of local networks in shaping migration decisions and settlement patterns (Vecoli, 1983; Spitzer and Zimran, 2023).

Chain migration and strong localism generated distinctive patterns of geographic concentration across and within U.S. cities (Vecoli, 1983). Migrants from specific towns and provinces often clustered

together: stonecutters from the marble-producing area of Massa settled in Barre, Vermont; migrants from Molfetta concentrated in Hoboken, New Jersey; and emigrants from Floridia, Sicily, in Hartford, Connecticut (Corradini, 1923). Residential segregation by origin persisted even within neighborhoods. In East Harlem, for example, Neapolitans clustered between 106th and 108th Streets while migrants from Basilicata settled primarily between 108th and 115th Streets (Orsi, 1985). In Chicago, Italians settled across numerous neighborhoods according to their origins (Vecoli, 1983). These settlement patterns reproduced Italy's localism and initially fragmented Italian communities abroad. Over time, however, interaction within shared urban spaces contributed to the gradual emergence of a broader Italian-American identity.

I.B Cultural Fragmentation in Italy After Unification

Italian national identity emerged largely in response to political unification in 1861 and diffused only gradually across the population (Riall, 2002; Banti, 2014). This delay is captured by a famous epigram by Piedmontese statesman Massimo d'Azeglio, who remarked in 1867 that "Italy has been made; Italians remain to be made." Regional divisions rooted in centuries of political fragmentation persisted along linguistic, cultural, and economic dimensions well into the twentieth century (Dickie, 1999).

At the time of unification, only a small minority of the population, estimated between 2.5 and 6.5 percent, spoke standard Italian (De Mauro, 1963). Italian functioned primarily as a written language accessible to literate elites, and literacy rates were low nationwide, particularly in the South (Ciccarelli and Weisdorf, 2019).

Educational expansion gradually promoted linguistic homogenization but proceeded unevenly. Despite the introduction of compulsory elementary schooling (1861 and 1877), attendance was not universal, and dialects remain widespread, including as a language of instruction at schools. Linguistic fragmentation, therefore, remained pervasive on the eve of World War I, especially in rural areas from which most migrants originated. A contemporary public inquiry documented that, as late as 1906, only 20% of the population could use Italian for everyday communication, and this proportion was likely lower in Southern rural areas (Corradini, 1910).

Geography and economic divergence reinforced these divisions. Mountainous terrain and limited transport infrastructure historically constrained interaction across regions, while the divide between the Center–North and the South widened after unification. Quantitative evidence indicates that the ratio of Southern to Northern GDP per capita declined substantially between unification and the interwar period (Felice, 2019). This divergence reflected distinct development trajectories and was accompanied by social and political tensions, including the post-unification conflict known as *brigantaggio* (Molfese, 1964; Dickie, 1999). Linguistic, economic, and cultural fragmentation thus remained central features of Italian society during the Age of Mass Migration.

I.C Cultural Convergence in Immigrant Communities in the United States

Italy's internal fragmentation was initially reproduced within Italian immigrant communities in the United States (Gabaccia, 2013). Over time, however, these localized identities converged into a broader Italian-American identity. Historians emphasize the role of shared experiences in the host country,

including labor market interactions, associational life, religious institutions, and exposure to common media, in fostering this convergence (Vecoli, 1985; Luconi, 2003).

Language convergence played a central role. Most Italian immigrants did not speak Italian upon arrival and instead communicated in regional dialects. Immigrants developed a creole dialect that mixed Neapolitan and other regional dialects with English as a *lingua franca* (Carnevale, 2010). The Italian-language press circulated widely within immigrant communities and helped standardize language use across regional groups (Deschamps, 1996).

Mutual aid societies also illustrate cultural convergence. Early associations were typically organized at the village or provincial level and restricted membership to migrants from specific localities. Covello (1967), for example, counted sixty-four associations with origin-based admission criteria in East Harlem alone. Beginning in the early twentieth century, these organizations increasingly federated across regional lines, culminating in national associations with large memberships (Vecoli, 1985).

Religious institutions further reinforced cultural convergence. Although nearly all Italian migrants were Catholic, devotional practices varied substantially across regions. In St. Louis, Sicilians boycotted worship of St. Ambrose, a cult that was predominantly widespread in Milan (Luconi, 2003). Over time, shared parishes and religious celebrations promoted interaction across regional groups and contributed to the emergence of common practices within increasingly heterogeneous communities. Homogeneous religious practices emerged both endogenously, as a consequence of between-groups interactions, as in the case of religious festivals in Manhattan, and in response to homogenizing efforts by the Catholic Church, particularly the Scalabrinians (Orsi, 1985; Choate, 2008).

High rates of return migration created sustained links between migrants and their communities of origin. Anecdotal evidence suggests that cultural practices developed within immigrant communities diffused back to Italy through return migrants and transnational networks (Wyman, 1993). A group of returnees in the remote village of Montefegatesi, for example, in 1902, financed the construction of a statue to Giuseppe Garibaldi, a symbol of national unity. Italy's emergence as an "emigrant nation" therefore reflected not only mass departure but also the circulation of identities and practices across the Atlantic (Choate, 2008).

II DATA

This section presents the novel linked dataset of Italian migrants (Section II.A), our methodology to measure national identity (Section II.B), and the additional data we employ in the analysis (Section II.D). We conclude by discussing how we construct the final analysis samples (Section II.E).

II.A Linking Ellis Island Ship Passenger Lists to Census Records

II.A.1 Motivation and Overview

Our empirical analysis requires information on the origins and destinations of Italian immigrants in the United States at a fine level of spatial granularity. While the U.S. population census reports the enumeration district where individuals resided, it does not report information on migrants' places of

origin beyond the country level.[5]

To circumvent this limitation, previous studies have linked U.S. census records to origin-country censuses for the United Kingdom (Coluccia and Dossi, 2025) and Norway (Abramitzky et al., 2012). In this paper, we link detailed microdata from ship rosters recorded at the Ellis Island immigration station, compiled by Coluccia and Spadavecchia (2025), to the U.S. population census. These rosters report individuals' names, surnames, years of birth, and years of immigration for approximately 4.2 million Italian migrants. Of these, about 2.5 million records are georeferenced to the coordinates of the municipality of origin in Italy and constitute our analysis sample. Coluccia and Spadavecchia (2025) show that migrants with a valid origin municipality are representative of the broader population of Italian immigrants in the United States.

Relative to prior approaches, this strategy offers two methodological advantages. First, all immigrants should appear in both datasets.[6] Second, and more importantly, the data include an additional blocking variable (year of immigration), which reduces the risk of false positives and multiple matches when name, surname, and year of birth do not uniquely identify individuals.

II.A.2 Description of the Linking Algorithm

Linking Ellis Island ship manifests to U.S. census records poses two challenges for existing record-linkage algorithms (e.g. Abramitzky et al., 2021). First, census enumerators and Ellis Island officials typically did not speak Italian, leading to frequent transcription errors. Second, many Italians Anglicized their names after arrival (Biavaschi et al., 2017). While spelling errors can be addressed using fuzzy string–matching techniques, name translations are more problematic, as original and translated names are semantically equivalent but orthographically distant.

To address this issue, we exploit the flexibility of neural networks to construct embeddings that preserve semantic proximity using bidirectional transformers (`SBERT`, Reimers and Gurevych, 2019).[7] We train a multilingual encoder on a subset of names, which we translate manually. Then, we fine-tune a classifier that uses the embeddings to identify name translations. We complement the `SBERT`-based classifier with a traditional Jaro-Winkler classifier to handle spelling mistakes. In our application, we consider two names a match if either their Jaro-Winkler proximity exceeds .95, a threshold commonly used in the literature, or their `SBERT` distance exceeds .6, the fine-tuned optimal threshold. Surnames are not subject to translation. Hence, we consider two surnames a match if their Jaro-Winkler similarity exceeds .95.

Armed with this classification rule, we link the records in the Ellis Island dataset to the U.S. census. Consider an entry i in the Ellis Island dataset, and let n_i and s_i denote i's first and last names. Let b_i and y_i denote i's year of birth and year of immigration, respectively. Define $\mathcal{M}_i = \{j \mid b_j \in [b_i - 2, b_i + 2] \wedge y_j \in [y_i - 5, y_i + 5]\}$ as the set of *candidate* matches among U.S. census records. Following

[5] Appendix Section B provides additional details on the data sources, the implementation of the linking algorithm, and the validation exercises used to assess its reliability.

[6] Approximately 95% of Italians who entered the United States between 1895 and 1924 passed through the Ellis Island station (Unrau, 1984).

[7] Figure F.1 provides a visual representation of the linking algorithm.

(Abramitzky et al., 2021), we allow for measurement error in reported birth and immigration years. Let $\overset{n}{\sim}$ denote the name-equivalence classifier defined above, such that $n_i \overset{n}{\sim} n_j$ if and only if the SBERT similarity between n_i and n_j exceeds 0.6 or their Jaro–Winkler similarity exceeds 0.95. Analogously, $\overset{s}{\sim}$ is the surname-equivalence classifier, such that $s_i \overset{s}{\sim} s_j$ if and only if their Jaro–Winkler similarity exceeds 0.95. The set of *effective matches* is then $\widetilde{\mathcal{M}}_i = \{j \in \mathcal{M}_i \mid n_i \overset{n}{\sim} n_j \wedge s_i \overset{s}{\sim} s_j\}$. The set $\widetilde{\mathcal{M}}_i$ may be empty, in which case i is unmatched; it may contain a single element (a "single" match); or it may contain multiple elements (a "multiple" match). To avoid double-counting, multiple matches are weighted by the inverse of the cardinality of $\widetilde{\mathcal{M}}_i$.

II.A.3 Evaluation of the Linking Approach and Validation

Figure I presents visual diagnostics to assess the performance of the linking algorithm.[8] Figure Ia shows the number of immigrants by year of arrival in the full sample (gray bars) and in the linked sample (red dots). The two series exhibit a strong positive correlation, indicating that the algorithm does not systematically over-sample migrants in particular periods. Out of 2,526,024 immigrants in the sample, the algorithm successfully links 473,922 individuals to at least one census record.

The average matching rate, shown in Figure Ib, is 19%. Matching rates are lower before 1900, primarily because the quality of recorded names and surnames in the Ellis Island data improves between 1895 and 1900. Thereafter, the matching rate remains stable, except during the war years, when overall immigrant inflows contracted. The average matching rate is comparable to that reported in earlier record linkage studies, such as Abramitzky et al. (2021). However, when focusing on non-English individuals, Abramitzky et al. (2021) report substantially lower rates (e.g., 10% for the Germans). More recent approaches, such as Althoff et al. (2025) and Green (2025), achieve higher matching rates. These figures are not directly comparable to ours, as both papers focus on samples composed predominantly of English-speaking individuals. Turning to non-English populations, Postel (2023) constructs a longitudinal linked sample of Chinese immigrants using U.S. census records and attains a 9% matching rate. Our algorithm thus performs on the higher end of match rates for relevant populations.

A high incidence of multiple matches may indicate spurious links. Figure Ic shows that nearly 96% of the matches produced by the algorithm are single, meaning Ellis Island records linked to one census record. The share of records associated with more than two matches is below 1%.

Figure Id explores the representativeness of the linked sample. We plot the log number of emigrants by municipality in the full sample on the horizontal axis and in the linked sample on the vertical axis. The solid line shows the linear fit. The correlation between the two series is positive and statistically significant, indicating that the distribution of municipalities of origin in the linked sample mirrors that of the full sample. The estimated slope ($\hat{\beta} = 0.825$) suggests that migrants from smaller municipalities are slightly more likely to be linked, but the magnitude of this difference is modest.

Table E.1 reports selected characteristics of migrants with a match (column 1) and without a match (column 2). Columns 3 and 4 present differences between the two groups. Panel A compares linked and non-linked migrants in the Ellis Island data, while Panel B reports the corresponding compari-

[8] Appendix Section B.IV provides complementary evidence to evaluate the internal reliability of the linking algorithm.

son using U.S. census records. In the Ellis Island data, linked migrants are younger than non-linked migrants. They are also slightly less likely to originate from Northern Italy and more likely to come from the Islands, primarily Sicily, although these differences are small and not statistically significant. In the U.S. census, linked migrants are more likely to be male. This pattern is expected, as matching rates for women are lower due to surname changes upon marriage. Conditional on labor force participation, the industry and occupation distributions of linked migrants closely resemble those of non-linked migrants. The geographic distribution of linked and non-linked migrants is also similar. Overall, the linked sample appears broadly representative of Italian immigrants in the U.S. census.

Validating the linked data against external sources is infeasible due to data limitations. In Figure F.4, we therefore implement two alternative procedures to assess the quality of the links.[9] We first construct an alternative sample, displayed in red, by providing `GPT5` with the same information available to the linking algorithm and asking it to evaluate the reliability of each match. In 91% of cases, the model classifies the matches as correct or very likely to be correct. All results are robust to restricting the analysis to this subset of matches. Second, we restrict the accepted matches to the highest-similarity match, and again find quantitatively unchanged results.

II.B Measuring Local, National, and American Identity of Immigrants in the United States

National identity in contemporary settings is typically measured using survey data (Rohner and Zhuravskaya, 2024). Such data are rarely available in historical contexts. To address this limitation, prior work has proxied national identity using first names, focusing on subsets of names with clear ideological or ethnic connotations (Bazzi et al., 2019; Fouka, 2020; Assouad, 2025; Lindfors et al., 2025).

In this paper, we propose a novel, disaggregated measure of national identity based on first names. Our approach builds on historical research documenting the extreme geographic fragmentation of naming practices in Italy (Delille, 1985; Viazzo, 1989). First names were tightly linked to local kinship norms, patron-saint cults, and dialectal traditions, resulting in pronounced spatial concentration. We interpret names that were more geographically concentrated as denoting particularistic identities, relative to names that were more evenly distributed across the national territory.

The first measure of particularism is the Local Name Index (LNI), defined as

$$LNI_n \equiv 100 \times \sum_m (s_{mn} - p_m)^2, \tag{1}$$

where $s_{mn} = C_{mn}/N_n$, C_{mn} denotes the total number of individuals named n in municipality m, and $N_n = \sum_m C_{mn}$. Thus, s_{mn} is the share of individuals named n residing in municipality m. Further, let $p_m = B_m/B$, where $B_m = \sum_n C_{mn}$ and $B = \sum_m B_m$, so that p_m represents the population share of municipality m. Intuitively, p_m is the share that name n would exhibit in municipality m if it were evenly distributed across the territory. By construction, the LNI equals zero if name n is distributed proportionally to population across all municipalities (i.e., $s_{mn} = p_m$ for all m), and it approaches 100 as the mass of name n becomes increasingly concentrated.

We construct the LNI using municipalities of origin reported in the Ellis Island records. Figure IIa

[9] Appendix Section B.V describes these validation procedures in detail.

shows the resulting distribution. To illustrate the index, we highlight two names, "Alfio" and "Raimondo," chosen because they occur with similar frequencies. Figure IIb plots the spatial distribution of "Alfio," while Figure IIc reports that of "Raimondo." The former, which has an LNI of 4.518, is highly concentrated in western Sicily, plausibly reflecting the local cult of Saint Alfio, patron saint of Taormina, near Siracusa. By contrast, "Raimondo" ($LNI = 0.712$) is more evenly distributed across municipalities throughout both Northern and Southern Italy.

The baseline particularism index captures the spatial concentration of names. By construction, however, the LNI does not account for spatial proximity between municipalities.[10] To address this limitation, we construct a distance-adjusted index (DLNI):

$$DLNI_n \equiv \sum_m \sum_{m'} s_{mn} d_{m,m'} s_{m'n} \ , \tag{2}$$

where $d_{m,m'}$ denotes the geodesic distance between municipalities m and m'. The index DLNI_n is equal to the expected distance between two randomly drawn individuals sharing name n. Accordingly, higher values of DLNI_n indicate more geographically diffuse, and hence more "national," names, while lower values indicate stronger local particularism.

To measure assimilation into U.S. society, we follow Abramitzky et al. (2020) and assign to each name its associated Foreign Name Index (FNI). The FNI is normalized between zero and 100 and takes higher values for names that are relatively more prevalent among immigrants than natives.[11]

Following Bazzi et al. (2019), we complement the name-based indices with measures of the probability of marrying Italians from other municipalities, provinces, and regions.

II.C Measuring Local and National Identity in Italy

Unlike in the United States, individual-level records from Italian censuses did not survive. To the best of our knowledge, no existing sources allow us to observe the universe of names. Here, we outline one strategy to address this data limitation and present two complementary measures constructed for the Italian case.

II.C.1 World War 1 and World War 2 Casualty Data as Individual-Level Repositories of Names

We collect names from two alternative sources: the repositories of World War 1 (WW1) deaths and World War 2 (WW2) casualties. Italy suffered approximately 650,000 military deaths in WW1 and 500,000 casualties in WW2.

WWI casualty data report first and last name, father's name, year of birth, municipality of origin, military rank, unit, year, place, and cause of death (Acemoglu et al., 2022).[12] We georeference the municipality of origin and disambiguate first and last names. For 570,355 individuals, we observe complete information on name, year of birth, and municipality of birth.

---

[10] Consider two names, n_1 and n_2, each observed in exactly two municipalities. If n_1 is present in two distant municipalities while n_2 is present in two nearby municipalities, it may be desirable for a measure of particularism to classify n_2 as more local than n_1.

[11] Formally, $FNI_n \equiv 100 \times \Pr(\text{Name} = n \,|\, \text{Immigrant}) / \left(\Pr(\text{Name} = n \,|\, \text{Immigrant}) + \Pr(\text{Name} = n \,|\, \text{Native})\right)$.

[12] The dataset was compiled by the Fascist regime for propaganda purposes and is available at `cadutigrandeguerra.it`.

WW2 casualty data have been recently digitized by the Italian Ministry of Defense. To the best of our knowledge, this paper is the first to use these data.[13] We construct a dataset covering the universe of recorded casualties. The records report first name, surname, date and municipality of birth, and date and region of death or disappearance. As with the WWI data, we georeference the municipality of origin and clean first names to merge them with the baseline name indices. The final dataset contains information on name, municipality of origin, and year of birth for 408,940 individuals.

We use these data to construct a yearly series of municipality-level name-based indices. The time span of the resulting series corresponds to the birth cohorts eligible for military service during the two World Wars. We restrict the analysis to cohorts 1875 to 1925. The resulting sample consists almost exclusively of men, since women did not serve in combat roles in the regular army.

In addition, we link the WW1 casualty data to the Ellis Island database to measure the number of U.S. emigrants who fought and died as soldiers in the Italian army.[14] While conscription was universal, the enlistment of migrants in the United States was essentially voluntary (Choate, 2008). We thus view the voluntary willingness to serve in the Italian army as an additional variable capturing national sentiment, as in Caprettini and Voth (2023).

II.C.2 Street Names

We use street names as a first complementary indicator of national and local identities at the municipality level (Azaryahu, 1996; Oto-Peralías, 2018). In Italy, municipal councils have near-complete discretion in naming new streets, and the process typically involves local communities. Street names thus plausibly reflect the preferences of local legislators and community members.

We collect the universe of municipal streets in Italy in 2025 from `OpenStreetMap`. The resulting dataset contains information on street names, georeferenced locations, and lengths for 849,124 streets.[15] We use a large language model (`GPT5`) to classify street names that reflect national and local identities. Examples of national identity street names include figures associated with the independence movement, such as Giuseppe Garibaldi and Giuseppe Mazzini, prominent historical figures, such as Cristoforo Colombo and Giuseppe Verdi, and references to patriotic events, such as Giacomo Matteotti or the Battle of Isonzo. By contrast, local identity street names honor notable individuals from the municipality or the surrounding area.

In addition, we identify street names belonging to three placebo categories: arts, literature, and religion. The same name may belong to more than one category. For example, Giosuè Carducci, a central figure in Italian literature at the end of the nineteenth century and winner of the Nobel Prize in 1906, was widely regarded as the "poet laureate" of the newly unified Italian state. As such, his name is classified as both national and literary. Moreover, the classification of a given street name generally depends on its location. A street named San Zeno is typically associated with religion, but

[13] Individual searches within the dataset are available at `difesa.it`.

[14] See Appendix Section A.IV for additional details on the linking approach and for an overview of the results.

[15] We exclude provincial, regional, and national roads, as well as highways, for two reasons. First, their names are not determined by municipal councils. Second, their names are typically administrative codes, as in the case of highways, or reflect only the places where they are located or the locations they connect.

in Verona, it is also classified as local, since San Zeno is the city's patron saint.[16]

II.C.3 Monuments

We use monuments as a third indicator of national sentiment (Johnson, 1995). We construct a dataset covering the universe of nationally recognized monuments compiled by the Italian Ministry of Culture.[17] The data report the year of designation and the location of each monument. We restrict the sample to monuments designated between 1861 and 2000, since coverage prior to unification depends on the legislation of pre-unitary states. We georeference each monument and assign it to a municipality using `GIS` boundary data.

Each monument is accompanied by an artistic and a historical description. We apply a large language model (`GPT5`) to these descriptions to distinguish monuments that signify national identity from all other monuments.[18] National-identity monuments typically commemorate soldiers from the municipality in which the monument is located who died in combat, often during World War I and World War II. The dataset includes information on 32,325 monuments designated between 1861 and 2000, of which 36% (11,774) are classified as reflecting national identity sentiment.

II.D Other Data

II.D.1 United States

We employ the individual-level U.S. census records to construct multiple economic and demographic variables—population, immigration, racial composition, age structure, and occupational and industry structures—that we include in the regressions as controls and use to check instrument balance (Ruggles et al., 2025). All census-based controls are available at the census-decade frequency, i.e., decennially, between 1900 and 1930.

II.D.2 Italy

The Italian Statistical Office (ISTAT) compiled the population, land area, and altitude, of each municipality from population censuses between 1861 and 2000. In addition, we digitize the literacy rate from historical population censuses (1881, 1901, 1911, 1921, and 1931). Using the data described in Section II.C.1, we compute the number of deaths at the municipality level during WW1 and WW2. Lastly, we employ the railway dataset digitized by Coluccia and Spadavecchia (2025) to measure the timing of railway access.

We complement these with coarser data at the levels of *circondario* and *provincia* compiled by Coluccia and Spadavecchia (2025).[19] For each *circondario*, we have information on the agricultural and manufacturing employment and urbanization, defined as the share of individuals living in towns with at least 5,000 inhabitants. For each *provincia*, we have employment data by manufacturing sector and proxies for technology adoption (the number and horsepower of engines by power source).

---

[16] The full prompt used for classification is reported in Appendix Section A.II.

[17] The dataset is available at `dati.cultura.gov.it`.

[18] As with streets, the prompt used for classification is reported in Appendix Section A.III.

[19] In 1926, there were 8,176 municipalities, 290 *circondari*, and 69 *province*.

II.E Construction of the Samples

II.E.1 Immigrants in the United States

First, we compile a panel of Italian migrants observed between 1900, or the first census in which they appear, and 1930. We link migrants over time using the longitudinal links produced by the Census Linking and Tree projects (Price et al., 2021; Abramitzky et al., 2025; Buckles et al., 2025). In this sample, name-based indices are computed by averaging them across all children born in the ten years preceding each census decade. Migrants' locations correspond to their place of residence in each census year.

Second, we construct a panel in which migrants are observed repeatedly at each child's birth. In this case, a migrant's location is taken from the census record closest to the child's birth year.

In both samples, we measure the diversity of the Italian community at the enumeration district level. Enumeration districts were approximately ten times smaller than counties and, therefore, better suited to capture local immigrant communities, particularly since Italians clustered in urban areas (Eriksson and Ward, 2019). A limitation of enumeration districts is that their boundaries changed across censuses. To control for consistent sub-county characteristics, we construct a hexagonal spatial grid and assign individuals to grid cells using data from Berkes et al. (2023). Because the hexagons are time-invariant, hexagon-fixed effects control for geographic variation at a finer level than counties and remain consistent over time. Panel A in Table I provides descriptive statistics at the neighborhood (Panel A.1) and individual (Panel A.2) levels.

II.E.2 Panel Dataset of Italian Municipalities

The analysis for Italy is conducted at the municipality level. We construct three datasets that differ in their temporal granularity.

First, name-based indicators are available at the cohort level. We therefore assemble a yearly panel tracking municipalities between 1870 and 1925. We collapse the data to coarser time intervals to reduce short-term noise and to reduce the number of empty cells, that is, municipality-year observations with no recorded individuals. Second, the analysis of street names relies on a cross-sectional dataset in which each municipality is observed once. Third, for monuments, we observe both their location and year of designation. We aggregate these data at the municipality level by decade and follow municipalities from 1870, corresponding to the 1860 to 1869 decade, through 2000.

In all samples, we use georeferenced coordinates to assign observations to municipalities defined by their 1991 boundaries, as historical `GIS` data on municipal borders are unavailable. Panel B in Table I lists descriptive statistics for the main outcome and explanatory variables employed in the analysis on national sentiment in Italy.

III LOCAL, NATIONAL, AND AMERICAN IDENTITIES OF ITALIAN IMMIGRANTS

This section presents the main results we document on the emergence of national identity among Italians in the United States. Section III.A discusses the empirical approach. In Section III.B, we present

the main results. Section III.C explores how exposure to diversity influenced return rates of Italians. Section III.D summarizes the robustness analyses.

III.A Empirical Strategy

III.A.1 Measuring the Diversity of Italian Immigrant Communities

We measure diversity of the Italian community by municipality of origin at the enumeration district level. Formally, let $d(t)$ be a district at time t. Denote with $n^m_{d(t)}$ the number of immigrants from municipality m located in district $d(t)$, and let $p^m_{d(t)} = n^m_{d(t)} / \sum_m n^m_{d(t)}$ be the share of migrants from that municipality. The diversity of the Italian community is the Shannon entropy:

$$H_{d(t)} = -\sum_m p^m_{d(t)} \log(p^m_{d(t)}) \,. \tag{3}$$

Entropy is weakly positive and increasing in the diversity of the Italian community. Shannon entropy reflects the idea that relatively more infrequent encounters are more informative than more frequent encounters. In our context, this property is appealing because exposure to a small number of immigrants from relatively distant municipalities is plausibly more salient than repeated exposure to many immigrants from the same place of origin. Results, however, are robust to using the fractionalization index, a more standard metric that does not embed this feature.

In Figure F.6, display Italian segregation in the United States. The blue line shows the distribution of the average distance between each immigrant's municipality of origin and the municipality of origin of other Italian immigrants residing in the same enumeration district. We compare this distribution to counterfactual benchmarks generated under random sorting.[20] The observed distribution exhibits excess mass at shorter distances, indicating that migrants from geographically proximate municipalities were more likely to settle together. However, segregation remains modest: the average observed distance between neighbors' municipalities is only 14% lower than the counterfactual mean under random sorting, implying that Italian immigrant communities were considerably diverse.

III.A.2 Baseline Specification

The baseline analysis employs the individual-level panel dataset described in Section II.E.1. We estimate variations of the following specification:

$$y_{ihdt} = \alpha_{m(i)} + \alpha_h + \alpha_t + \beta \cdot H_{d(t)} + X_d' \Gamma_1 + X_i' \Gamma_2 + \varepsilon_{ihdt} \,, \tag{4}$$

where i, h, d, and t denote an individual, the hexagon of residence, the enumeration district, and the census year, respectively. Municipality-of-origin fixed effects ($\alpha_{m(i)}$) control for time-invariant origin characteristics such as average population size, emigration intensity, and geographic location. Area-level fixed effects (α_h) absorb time-invariant local characteristics, including location and proximity to

[20] To benchmark observed segregation against random sorting, we construct counterfactual neighbor-distance distributions via permutation within districts. For each individual i and district d, we define the pool of potential neighbors as all Italian immigrants observed in the same census year and district, excluding members of i's household. Let n_{id} denote the number of such neighbors in i's observed district. In each simulation draw, we sample n_{id} individuals at random from the pool and compute the mean distance between i's municipality of origin and those of the sampled individuals. We run 100 simulations per individual, yielding a counterfactual distribution under random assignment that preserves group sizes and household structure.

major ports of entry. Time-fixed effects (α_t) capture aggregate shocks, such as business-cycle fluctuations, that affect migration flows to the United States.

Because enumeration district boundaries change over time, we cannot include district fixed effects. We therefore estimate variants of (4) that include hexagon-by-time fixed effects ($\alpha_{h\times t}$). These absorb time-varying unobserved heterogeneity at sub-county spatial units and purge confounding variation driven by local factors such as economic conditions that evolve within counties. We also report specifications that include municipality-by-time fixed effects ($\alpha_{m(i)\times t}$). These absorb time-varying heterogeneity at the origin level, including population growth, which mechanically generates larger emigrant flows, as well as local economic conditions in Italy. In addition, they account for the expansion over time of origin-specific migrant networks.

In various specifications, we include district-level controls (X_d) capturing demographic characteristics (population, the share of women, the share of African Americans, the share of immigrants, and the share of Italians) and economic characteristics (employment shares across occupations). In addition, some specifications include individual-level controls (X_i), namely occupation, gender, birth year, and year of immigration.

The coefficient of interest (β) captures how the outcome y responds to changes in the diversity of the local Italian immigrant community. We consider four classes of outcomes: the local name index and its distance-adjusted variant; the foreign name index; and a set of indicators capturing the probability that individual i married an Italian migrant originating from a different municipality, province, or region.

III.A.3 Identification Challenge and Instrumental Variable Design

Identification Challenges. Identification of the causal effect of diversity in (4) is potentially problematic for two reasons. First, the origin mix within a district may reflect selective sorting of immigrants from specific municipalities across U.S. locations. If these groups differ in unobserved traits related to identity, then $\hat{\beta}$ may partly capture omitted origin- and destination-level heterogeneity.

Second, $H_{d(t)}$ may respond endogenously to time-varying local conditions. Destination shocks can simultaneously affect outcomes and the composition of the Italian community by shifting migration flows and cohort retention. Relatedly, residential mobility and marriage-market choices may feed back into measured diversity, inducing simultaneity.

Instrumental Variable Design. To address these concerns, we construct an instrument based on quasi-random exposure to Italian diversity induced by ship-level composition. Building on Dippel and Heblich (2021), we hypothesize that interactions during the transatlantic voyage, which could last more than 50 days, as documented by Hatton (2024), shaped subsequent settlement patterns. Under this hypothesis, we can isolate plausibly exogenous ship-level variation in the origin mix of passengers to estimate the causal effect of exposure to diversity.

Consider individual i from municipality $m(i) = m$ who traveled on ship $s(i) = s$.[21] Let x_{os} denote

[21] A "ship" is defined as a unique ship-year pair. We omit the year variable and simply call it a "ship" to avoid notational burden.

the number of immigrants from municipality o who traveled on ship s, and let p_{os} be the share of immigrants from municipality o on s, so that $\tilde{p}_{os} = p_{os}/(1 - p_{ms})$ is the associated normalized share.[22] The instrument for individual i is then given by

$$\hat{H}_i = -\sum_{o \neq m} \tilde{p}_{os} \log \tilde{p}_{os} \,. \tag{5}$$

We compute $\hat{H}_i$ only on ships with at least 100 immigrants (95% of the total) to avoid numerical instability. We estimate regression (4) using the instrument in (5) via two-stage least squares (2SLS).

The instrument associated with each individual excludes shipmates born in the same municipality. This leave-out methodology ensures that the instrument does not mechanically reflect pre-existing social networks. Results remain quantitatively unchanged when excluding immigrants from the same province, indicating that residual social networks not captured by excluding own-municipality shipmates are unlikely to drive the results.

Instrument Relevance. We implement two analyses to verify that ship diversity predicts diversity exposure at the neighborhood level.

First, we regress the endogenous variable (3) on the instrument (5), conditional on the controls we include in the baseline regression specification. Figure IIIa displays the result. We find a positive, strongly statistically significant association between neighborhood and ship entropy.

Second, we implement a simple simulation approach building on Basse et al. (2024).[23] We test whether shipmates from different municipalities are disproportionately likely to co-locate after arrival. For each ship, we compute the number of passengers who end up in the same district. We compare this probability to a simulated benchmark in which the same passengers are assigned to districts at random. Excess cross-municipality co-location relative to this random benchmark indicates that co-traveling on the same ship predicts subsequent residential clustering.

Figure IIIb reports the distribution of the ship-level excess co-location statistics against its asymptotic distribution under the null of no co-location (in blue). We find that passengers of the same ship are substantially more likely to settle in the same district. For example, we find more than four times as many ships with a p-value beyond 5% than what would be implied by random co-location.

Evidence in Support of the Identification Assumption. Identification relies on the assumption that, conditional on fixed effects and controls, ship composition influences outcomes only by shifting post-arrival exposure to other Italian origins. Under this exclusion condition, the IV estimates recover the causal effect of exposure to a more diverse mix of Italian origins on identity formation.

We implement two exercises to assess the empirical plausibility of this assumption. First, we examine instrument balance by regressing predetermined area characteristics on the instrument, con-

[22] Municipality shares on ship s are computed as $p_{os} = (x_{os} + \lambda)/(N_s + \lambda K)$, where x_{os} is the number of passengers from municipality o, $N_s = \sum_o x_{os}$ is the total number of passengers on the ship, and K is the number of distinct municipalities represented. Absent smoothing, these shares are the empirical frequencies from a multinomial distribution over municipalities induced by the ship's passenger list. Because $\log(p_{os})$ appears in Shannon entropy, small counts generate extreme or undefined values, especially in leave-one-out transformations. We therefore apply Laplace smoothing ($\lambda = 1$), which amounts to adding one observation per municipality. This regularizes entropy in finite samples without affecting large ships, as the influence of smoothing vanishes as N_s grows.

[23] Appendix Section A.V discusses in greater detail the simulation-based approach.

ditioning on the same fixed effects and controls as in the baseline specification. Figure F.7a reports the resulting coefficients. Because enumeration districts cannot be tracked consistently over time, we construct these predetermined characteristics at broader geographic levels and report results at the county (in blue) and the hexagon (in red) levels. Across a wide set of variables capturing the diversity of the Italian community, demographics, and employment shares by industry and occupation, the estimates are small and statistically insignificant, indicating no correlation between the instrument and pre-migration characteristics.

Second, we conduct a placebo exercise by regressing the instrument on one-decade leads of individual characteristics that should not be affected by ship diversity. These include demographic variables, which are fixed by construction, and measures of employment status by industry and occupation. Figure F.7b reports the resulting estimates. We find no statistically significant associations, suggesting that the instrument is not systematically related to individual characteristics that may influence location choices or to forward-looking outcomes.

Movers' Design. As a complementary approach, we implement a mover's design along the lines of Chetty and Hendren (2018). This strategy takes advantage of individuals who move between neighborhoods and exploits differences in the diversity of Italian communities they are exposed to.[24] Because we compare movers and stayers over time, this framework allows us to evaluate the empirical plausibility of the parallel trends assumption through event-study graphs. We find that exposure to more diverse communities decreases name particularism (see Table E.2) and document no evidence of statistically significant pre-move differences between treated and control individuals (Figure F.8).

III.B Local, National, American Identity and Neighborhood Diversity

III.B.1 Descriptive Patterns of Assimilation After Arriving in the U.S.

We begin our analysis with a descriptive exercise. In Figure IV, we display the evolution of the local name index (Figure IVa) and the foreign name index (Figure IVb) for children of immigrants by the number of years since migration to the United States. We split the baseline sample into two groups: red dots represent immigrants who settle in neighborhoods with above-median Italian community diversity, whereas blue dots represent immigrants who settle in neighborhoods with below-median diversity.

Migrants were similar in terms of their LNI and FNI at the time of migration, regardless of the neighborhood they settled in. We detect no statistically significant difference in LNI or FNI for migrants who had been living in the United States for just one year, indicating that we are comparing individuals who were not differentially selected on levels of particularism or assimilation at arrival.

The trajectories of the LNI and FNI, however, diverge markedly over time. Migrants who settle in low-diversity areas exhibit no change in name particularism and experience substantial name-based assimilation. Ten years after arrival in the United States, their children bear names that are no longer differentially local, but are 5% more similar to English-sounding names. This pattern of gradual assimilation is consistent with evidence by Abramitzky et al. (2020).

[24] Appendix Section A.VI describes the empirical specification and provides a detailed discussion of the results.

Migrants who settle in high-diversity areas, by contrast, experience a pronounced decline in name-based particularism. Ten years after migration, their children bear names that are 11% less local. Differences in name particularism emerge as early as two years after arrival and remain relatively stable thereafter. In contrast, even ten years after arrival, the children of migrants in high-diversity areas do not display significantly more foreign-sounding names than at arrival. This pattern suggests substantially slower, or even absent, cultural assimilation into U.S. society among migrants exposed to more diverse Italian communities.

III.B.2 Name-Based Assimilation: OLS and 2SLS Results

To provide more formal evidence for these descriptive patterns, Table II reports estimates of specification (4). Panels A, B, and C present results using the local name index (1), the distance-adjusted local name index (2), and the foreign name index as dependent variables, respectively. Columns (1–6) report OLS estimates. Columns (7–8) present the reduced-form effects of the instrument in (5) on each outcome, while columns (9–10) report the corresponding 2SLS estimates.

Depending on the specification, the OLS estimates indicate that a unit increase in the entropy of the Italian community is associated with a 0.12–0.17 decrease in the local name index. To put these magnitudes into perspective, a one-standard-deviation increase in Italian community diversity corresponds to a 4% or 5%-standard-deviation decline in the local name index. Panel B reports results using the distance-adjusted LNI. Because higher values of the DLNI indicate names that are more geographically widespread within Italy, increases in this index capture stronger national identity. A unit increase in Italian community diversity is associated with a 0.02 increase in the DLNI, implying that a one-standard-deviation increase in diversity yields a 10%-standard-deviation increase in the distance-adjusted local name index. Column (3) reports the complementary effect of Italian diversity on the foreign name index, which, following Abramitzky et al. (2020), we interpret as a proxy for cultural assimilation in U.S. society. A unit increase in Italian diversity is associated with a 0.02 increase in the FNI. Analogously, a one-standard-deviation increase in Italian diversity corresponds to an 11%-standard-deviation increase in the foreign name index.

These estimates are stable across a range of specifications. In column (2), we replace county fixed effects with hexagon fixed effects to control for more granular time-invariant local characteristics. In column (3), we include municipality-of-origin–by–time fixed effects, which absorb time-varying unobserved heterogeneity at the origin level. For instance, population growth may increase migrant outflows from a given municipality, potentially affecting the likelihood of settling near migrants from the same origin. In column (4), we add individual-level time-invariant controls from the U.S. census, including occupation, gender, birth year, and year of immigration. Finally, columns (5) and (6) incorporate neighborhood-level time-varying controls to account for demographic factors–population size, the share of African Americans, immigrants, and Italian immigrants–and economic conditions, measured by employment shares across industries, that may independently influence assimilation.

These estimates are consistent with the predictions of a simple model of identity choice, along the lines of Akerlof and Kranton (2000) and Shayo (2009), which we outline in Appendix Section D. Under mild and plausible assumptions, higher Italian diversity reduces the prevalence of local iden-

tities because the returns to national identity increase. Similarly, higher Italian diversity decreases the adoption of American identity, because identity serves as a coordination device whose value depends on the level of diversity. Higher diversity thus reallocates identity *within* the immigrant group, rather than toward the host society.

As discussed in Section III.A.3, OLS estimates may not recover the causal effect of exposure to Italian diversity due to selective sorting and simultaneity concerns. We therefore report reduced-form estimates based on the ship-level instrument defined in (5) in columns (7–8), and the corresponding 2SLS estimates in columns (9–10). The instrumental-variable estimates corroborate the OLS results and are substantially larger in magnitude. Specifically, the 2SLS coefficients are approximately an order of magnitude larger than their OLS counterparts.

There are two reasons why the causal effect for this group may exceed the average effect captured by OLS. First, migrants whose residential choices were influenced by shipmate networks were likely more socially embedded within Italian communities, making them particularly responsive to within-group diversity. Second, selective sorting and measurement error in neighborhood composition are likely to attenuate OLS estimates, as migrants with stronger pre-existing identity or assimilation preferences may self-select into particular locations. By contrast, the IV isolates variation in exposure that is orthogonal to such preferences and therefore recovers the effect for individuals whose social environment was plausibly exogenously assigned.

Under this interpretation, the IV and OLS estimates are not directly comparable: the latter capture an average association in the population, whereas the former identify a treatment effect for a specific subpopulation. Nonetheless, the IV results indicate that exogenous exposure to Italian diversity causally increased national identity formation and reduced cultural assimilation into U.S. society, under the maintained exclusion restriction.

These results highlight two findings. First, the composition of the immigrant community matters for assimilation into the host country's society: exposure to greater within-origin diversity strengthens national identity and slows assimilation. As a result, focusing solely on the size of immigrant groups by country of origin misses the role played by the *composition* of diasporic communities. Second, inter-group contact among migrants reallocates identity within the immigrant population, fostering a shared national identity rather than accelerating convergence toward native cultural norms, as predicted by the theoretical framework in Appendix Section D.

III.B.3 Out-Group Marriage Patterns

Marriage patterns provide an additional proxy for group identity formation (Abramitzky et al., 2020). In our context, Gordon (1964) documents the central role of marriages between migrants and natives, as well as between migrants from different origin groups, as indicators of immigrant assimilation. We distinguish three types of marriages, based on whether spouses originate from different municipalities, different provinces, or different regions.

In more diverse immigrant communities, individuals are more likely to marry outside their municipality of origin simply because fewer potential spouses from the same origin are available. As a result, raw inter-municipality marriage rates may mechanically increase with diversity. To address this

issue, we follow Bazzi et al. (2019) and construct supply-adjusted measures of inter-origin marriage that net out the role of local marriage-market composition. Specifically, we compare observed inter-origin marriages with the intergroup marriage probabilities that would arise under random matching between spouses irrespective of origin. We interpret the residual as reflecting identity-driven integration rather than mechanical supply effects.[25]

Figure Va displays the results. Each dot reports estimates from equation (4), comparing specifications that do not correct for the mechanical association between diversity and intergroup marriage probabilities (red dots) with those that account for it (blue dots). Exposure to more diverse Italian communities is associated with a higher probability of intergroup marriage. When controlling for supply-side effects, a unit increase in Italian community diversity raises the probability of marrying an immigrant from another municipality by 4%. Given that such marriages were relatively rare, this effect corresponds to approximately 22% of the mean. Exposure to more diverse communities also has a positive, statistically significant effect on the probability of marrying an Italian from another province, but not from other regions.[26]

III.C Return Migration and Neighborhood Diversity

A final indicator of group identity formation is the propensity of migrants to return to their origin country. Among all emigration countries, Italians had the highest return rates (Bandiera et al., 2013). However, if exposure to diversity within immigrant communities fostered a shared national identity that was distinct from local particularism in Italy, migrants in more diverse communities should have been less inclined to return.

Data on return migration are notoriously scarce. Neither the U.S. nor Italian authorities collected systematic, spatially disaggregated data on return migrants. We propose a methodology to partially address this limitation that uses longitudinal links of U.S. census records. We derive a differential return estimator and use it to examine whether return rates among Italian immigrants varied with the diversity of the Italian communities they were exposed to.

Let $I \in \{0,1\}$ indicate nationality of an individual, where $I = 1$ denotes Italians and $I = 0$ denotes natives, and let $D \in \{H, L\}$ indicate whether the individual lives in a high- or low-diversity neighborhood. We observe whether an individual present in the census at time t can be linked forward to the U.S. census at $t + h$; let $L_{t+h} = 1$ if the individual is successfully re-observed, and $L_{t+h} = 0$ otherwise. We define the forward-link probability as $\ell_{I,D} = \Pr(L_{t+h} = 1 \mid I, D)$.

Linking forward in the U.S. census requires not returning to Italy and, separately, not disappearing due to other factors such as death, internal migration within the U.S., or linkage error. Let $R = 1$

[25] For each married individual i, define the marriage market $\mathcal{M}(i)$ by census year and enumeration district, and let $\mathcal{S}_{-i}$ denote the set of opposite-sex individuals in $\mathcal{M}(i)$. The supply-only benchmark is constructed by repeatedly drawing a spouse $j \in \mathcal{S}_{-i}$ uniformly at random and computing $\mathbb{E}[I\{g(i) \neq g(j)\} \mid \mathcal{M}(i)]$, where $g(\cdot)$ denotes origin (municipality, province, or region). We net out this benchmark from observed inter-origin marriage outcomes in the estimating equation.

[26] Two reasons may explain why we do not find an effect of diversity on out-of-region marriages. First, regions may be too coarse to capture the identity. Given that there are only 21 regions in Italy and that most migrants originated from Southern regions, there is limited variation in marriages with individuals from other regions. Second, the supply-side correction is strongest at coarse aggregation levels, implying that once this benchmark is netted out, there is little residual variation in marriage patterns left.

denote a return to Italy between t and $t+h$, and let $M=1$ denote disappearance without return. An Italian individual in a D-neighborhood remains observable in $t+h$ if and only if $R=0$ and $M=0$. The forward-link probability for Italians is therefore $\ell_{1,D}=1-r_{1,D}-m_{1,D}$, where $r_{1,D}=\Pr(R=1 \mid I=1,D)$ is the (latent) return probability and $m_{1,D}=\Pr(M=1 \mid I=1,D)$ captures all other forms of attrition. For natives, $r_{0,D}=0$, so $\ell_{0,D}=1-m_{0,D}$.

We compare how much more Italians disappear relative to natives in high-diversity areas versus low-diversity ones. Consider the "difference-in-differences" object $(\ell_{1,H}-\ell_{0,H})-(\ell_{1,L}-\ell_{0,L})$. Plugging in the decomposition above yields our differential return estimator:

$$\varrho \equiv (\ell_{1,H}-\ell_{0,H})-(\ell_{1,L}-\ell_{0,L}) = -(r_{1,H}-r_{1,L})-[(m_{1,H}-m_{0,H})-(m_{1,L}-m_{0,L})] \tag{6}$$

Estimator ϱ recovers the difference in return rates between high- and low-diversity neighborhoods, $r_{1,H}-r_{1,L}$, provided that the Italian–native attrition gap unrelated to return, $m_{1,D}-m_{0,D}$, is stable across neighborhood types.[27] Under this assumption, ϱ isolates the extent to which the additional disappearance of Italians in high-diversity areas reflects true return migration rather than differential mortality, internal migration, or linkage failure.

Figure Vb displays the estimated values of ϱ across age bins, where H and L denote, respectively, districts with above- and below-median diversity of Italian migrants.[28] We find that return rates are between 10% (for ages 0–10) and 5% (for ages 25–29) lower in high-diversity neighborhoods relative to low-diversity neighborhoods. Bandiera et al. (2013) estimate that more than 60% of Italians returned over the course of a single decade, implying that exposure to relatively more diverse communities reduced the propensity to return by approximately 8%.[29]

Lower return rates among migrants in high-diversity districts are consistent with the interpretation that diversity fostered a durable sense of national belonging, thereby weakening ties to local communities in Italy.

III.D Robustness Analysis

A first concern regarding our name-based indicators is that naming decisions for firstborn children may reflect family tradition rather than parents' identity choices. In Table E.3 and Table E.6, we therefore recompute all name-based measures excluding firstborn children. The results remain quantitatively unchanged.

The local name index is computed at the municipal level to maximize the granularity of infor-

[27] Our approach requires weaker identification assumptions compared to a more naïve estimator equal to the difference in forward-link rates between Italians and natives within neighborhoods, $\ell_{1,D}-\ell_{0,D}=\ell_{1,D}-\ell_{0,D}=-r_{1,D}-(m_{1,D}-m_{0,D})$. This difference identifies the return rate $r_{1,D}$ only if $m_{1,D}=m_{0,D}$, i.e., Italians and natives must have the same non-return disappearance rate. This is a demanding condition: if, as is plausible, Italians and natives have different linkage success rates, then $\ell_{1,D}-\ell_{0,D}$ cannot be interpreted as a return probability.

[28] To estimate the differential return ϱ, we construct 30-year forward links. Results are qualitatively similar when using alternative horizons. Standard errors are computed using a district-level cluster bootstrap with 100 replications. For each age group, we resample districts with replacement and recompute the collapsed difference-in-differences estimand; reported standard errors are the bootstrap standard deviation.

[29] This comparison should be interpreted with caution. The methodology in Bandiera et al. (2013) is not directly comparable to our pseudo-return estimator, and we focus on long-run return decisions over a 30-year horizon, whereas their estimates concern ten-year return spells.

mation contained in first names. However, one may be concerned that municipalities are too small to meaningfully capture identity. In column (1) of Panels A and B in Table E.5 and Table E.8, we therefore recompute the LNI at the province-of-origin level. Provinces are substantially larger administrative units—Italy has only 107 provinces, compared to more than 8,000 municipalities—so the province-level index reflects a coarser layer of identity. The results based on province-level and municipality-level LNI are quantitatively similar.

In the baseline definition of Italian community diversity in (3), we measure diversity at the enumeration district and municipality levels. Enumeration districts are small and, we argue, better suited than counties to proxy for the local communities immigrants interacted with in the U.S.. Municipalities, in turn, provide an appropriate proxy for local social networks in the origin country. In Table E.5 and Table E.8, we replicate the analysis along two dimensions. First, in columns (3–4) and (5–6), we recompute the diversity measure at the province and region levels, respectively. Second, in columns (2), (4), and (6), we recompute diversity at the U.S. county level. Across these alternative definitions, we consistently find a negative association between Italian diversity and name-based particularism. Estimates are less precise when diversity is measured at the county level, which is expected: because counties are substantially larger than the relevant social units migrants interacted with, aggregating diversity at this level introduces attenuation bias due to measurement error.

A related concern regarding the measure of the diversity of the Italian immigrant community in (3) is that Shannon entropy is not the most commonly used diversity metric in economics. In Table E.4–Table E.7, we therefore replicate the baseline estimates using the fractionalization index to measure the diversity of the Italian immigrant community and obtain qualitatively unchanged results.[30]

Italian immigrants were concentrated in a small number of states—most importantly New York, Pennsylvania, New Jersey, and Massachusetts—raising the concern that our results may be driven by idiosyncratic characteristics of specific locations rather than by the diversity of Italian communities per se. In Figure F.9 and Figure F.10, we therefore replicate the baseline analysis while sequentially excluding one state at a time from the estimation sample. The OLS and 2SLS estimates remain stable and statistically indistinguishable from the baseline coefficient regardless of the state we exclude.

Similarly, selective sorting across U.S. locations by migrants from specific areas of Italy may induce spatial autocorrelation and potentially inflate statistical significance. To address this concern, in Figure F.11 and Figure F.12 we report confidence intervals obtained using a range of alternative variance estimators. In particular, we show that statistical significance is unchanged when adopting different clustering strategies across U.S. locations and when applying the Conley (1999) spatial correction with varying distance cutoffs. Similarly, the estimates remain statistically significant when clustering the standard errors by the origin of immigrants, at different layers of spatial granularity, and when accounting for spatial autocorrelation at the origin level.

[30]The fractionalization index is defined as one minus the Hirschman-Herfindahl index. Formally, let s_g denote the share of migrants originating from municipality g in a given district. The fractionalization index is given by $F = 1 - \sum_g s_g^2$. This measure captures the probability that two randomly drawn individuals belong to different origin groups and is widely used as a measure of ethnic or group diversity (e.g., see Alesina et al., 2003).

IV EMIGRATION AND THE RISE OF NATIONAL IDENTITY IN ITALY

We now turn to Italy. We explore how the sentiment of national identity that emerged among U.S. emigrants contributed to the rise of national identity in Italy. Section IV.A reports the results obtained using names as a measure of identity. In Section IV.B and Section IV.C, we present the results using street names and monuments as ancillary proxies. Section IV.D employs WW1 death rates among emigrants as an additional high-stakes measure of national sentiment. Section IV.E discusses the robustness analyses.

IV.A Name-Based National Identity

Our primary measure of national identity formation in Italy relies on given names, as described in Section II.C.1. The idea is to trace how name-based particularism in each municipality evolved after migrants from that origin were exposed, while in the United States, to Italians from other parts of the country. To this end, we implement two empirical strategies.

IV.A.1 Panel Regression

First, we estimate a set of panel regressions of the following form:

$$y_{mt} = \alpha_m + \alpha_t + X'_{mt}\Gamma + \sum_{\tau=1880}^{1930} \beta_\tau \cdot Diversity\ Exposure_m \cdot I(t = \tau) + \varepsilon_{mt}\ , \tag{7}$$

where m indexes municipalities and t indexes years, with municipality and year fixed effects given by α_m and α_t. The vector X_{mt} collects time-varying municipality-level controls. In some specifications, X_{mt} additionally includes the number of emigrants from m during the Mass-Migration period interacted with year dummies, allowing outcomes to flexibly evolve with emigration intensity. The coefficients of interest are β_τ, which trace the differential evolution of y_{mt} over time as a function of diversity exposure. We define $Diversity\ \ Exposure_m$ as the share of migrants from municipality m during the Mass-Migration period who settled in U.S. neighborhoods in the top 20% of the baseline entropy distribution in (3).

Because Italian emigration to the United States was concentrated in the period 1895–1921, the coefficients for the pre-migration years in (7) provide a falsification test. If the treatment effect coefficients capture the effect of exposure to diversity through migration, there should be no systematic differential evolution between high- and low-exposure municipalities before large-scale emigration began.

IV.A.2 Difference-in-Differences Regression

To implement a formal difference-in-differences design, we define a municipality's treatment time as the first year in which its annual emigration rate exceeds 1‰ of the 1880 population. This choice is motivated by Spitzer and Zimran (2023), who show that U.S.-bound emigration followed an S-shaped diffusion pattern. We interpret the 1‰ threshold as the point at which out-migration gains momentum. While the threshold is arbitrary, results are robust to alternative cutoffs between 1‰ and 5‰ (see Table E.9).

Formally, we estimate variations of:

$$y_{mt} = \alpha_m + \alpha_t + X'_{mt}\Gamma + \sum_{\substack{\tau=-10 \\ \tau\neq 0}}^{20} \beta_\tau \cdot Diversity\ Exposure_m \cdot I(t - t^*_m = \tau) + \varepsilon_{mt}\,, \tag{8}$$

where t^*_m denotes the first year in which the emigration rate exceeds 1‰. The control vector X_{mt} always includes an indicator $I(t \geq t^*_m)$ that equals one once municipality m enters the mass-migration period. The coefficients β_τ therefore trace event-time differences in outcomes for municipalities with higher *vis-à-vis* lower diversity exposure, before and after the mass emigration.

Identification in (8) relies on a parallel-trends assumption. Absent exposure to diverse Italian communities in the United States, municipalities with higher $Diversity\ Exposure_m$ should have exhibited the same evolution in y_{mt} around t^*_m as other municipalities. We assess this implication by reporting the pre-treatment coefficients β_τ for $\tau < 0$, which are small in magnitude and statistically indistinguishable from zero.

IV.A.3 Results

In Italy, the mass migration period coincided with persistent localism and fragmentation, notwithstanding sustained efforts by Fascist authorities to promote national identity formation (De Mauro, 1963). Figure VIa plots the average local name index of Italian men born between 1880 and 1920. Over this period, the LNI increased by more than 30%, indicating a shift toward more locally concentrated naming practices.

Figure VIb reports the estimates from equation (7). Italian municipalities whose migrants later settled in more diverse U.S. locations do not differ from other municipalities between 1885 and 1900. Because large-scale emigration began in the mid-1890s, this pattern indicates that treated and control municipalities were comparable in terms of name particularism prior to the onset of mass migration. Beginning in the 1910s, however, municipalities whose migrants were exposed to relatively more diverse Italian communities in the United States exhibit significantly lower levels of local naming than control municipalities. The magnitude of this difference increases over time and is largest by the late 1920s.

In Table III, we report alternative specifications. Column (1) includes municipality and year fixed effects. Column (2) additionally controls for interactions between the total number of U.S. emigrants during the mass migration period and year fixed effects. This specification thus compares municipalities with similar emigration intensity and exploits variation in the destinations where migrants settled in the United States. The results confirm that the LNI declined in municipalities whose migrants were exposed to more diverse Italian communities in the U.S., while the pre-migration coefficients remain small in magnitude and statistically indistinguishable from zero. Column (3) documents a similar pattern when using the distance-adjusted local name index as the outcome.

Figure VIc presents the dynamic treatment effects estimated from equation (8). Prior to the onset of mass emigration, differences in the average LNI between treated and control municipalities are small and statistically insignificant. Following the start of the mass migration, naming practices in treated municipalities became progressively less local relative to those in control municipalities. As

in Figure VIa, this gap widens over the subsequent two decades. On average, municipalities whose migrants settled in relatively more diverse areas of the United States experience a 52% decline in the local name index relative to the mean.

In Table III, we report results from alternative specifications of equation (8). Column (4) includes municipality and year fixed effects. Column (5) further controls for the total number of U.S. emigrants between 1892 and 1924, interacted with period dummies, thereby exploiting variation in emigrants' settlement destinations among municipalities with comparable emigration intensity. Column (6) documents a similar pattern when we employ the distance-adjusted local name index as the outcome variable.

Across all specifications, we also report the coefficient on a secondary treatment indicator that equals one after a municipality enters the mass migration period. The coefficient of this term remains consistently small in magnitude and statistically insignificant, indicating that transatlantic emigration *per se* did not affect local identity, thereby suggesting that increased national identity in treated municipalities was driven by intergroup contact between emigrants from each municipality and Italians from other parts of the country.

In Figure VId, we examine heterogeneity in the effect of exposure to diversity through migrant communities by the size of the origin municipality. We estimate the baseline mean-shift specification in (8) and interact the DiD term with indicators for quintiles of municipality population.[31] The gray bars report the average local name index in the pre-treatment period for municipalities in each quintile. As expected, smaller municipalities exhibit higher levels of name particularism, reflecting stronger local identities in smaller and more isolated communities. Exposure to more diverse Italian communities in the United States reduces the LNI across all population quintiles. However, the decline is quantitatively larger in larger municipalities, indicating that transatlantic emigration amplified pre-existing differences in particularism between large and small centers.

IV.B Street Naming Patterns

Street names provide an alternative proxy for national identity formation, as municipal councils exercise discretion over the naming of new streets. We classify street names into two identity-related categories, local and national, and three "placebo" categories, art, literature, and religion. Our dataset consists of a cross-section covering the universe of municipal streets in Italy as of 2025.

We estimate variations of the following specification:

$$y_m = \alpha_{p(m)} + X_m'\Gamma + \beta_1 \cdot Diverse_m + \beta_2 \cdot Homogeneous_m + \varepsilon_m\,, \tag{9}$$

where $p(m)$ denotes the province of municipality m. The variable $Diverse_m$ measures the number of U.S. emigrants from m who settled in neighborhoods in the top 20% of the entropy distribution, while $Homogeneous_m$ captures the number of emigrants from m who settled in less diverse neighborhoods. The control vector X_m includes population levels before (1881) and after (1991) the mass-migration

[31] In this specification, we additionally include population-quintile-by-year fixed effects to absorb time-varying shocks common to municipalities of similar size that may affect naming practices.

period, as well as geographic characteristics such as municipal area and altitude. Depending on the specification, we further include province or region fixed effects.

Table IV reports the results. In columns (1–3), the dependent variable is the share of streets bearing a national-identity designation, while columns (4–6) focus on streets associated with local identity. Columns (1) and (4) present the baseline specification; columns (2) and (5) add province fixed effects; and columns (3) and (6) further control for municipality characteristics. We find that the number of emigrants who were exposed to diverse Italian communities in the United States is positively associated with the share of streets commemorating national figures or events. By contrast, the number of emigrants exposed to relatively more homogeneous migrant communities is negatively associated with national-identity street naming.

Local-identity naming follows the opposite pattern. Exposure of emigrants to more diverse Italian communities is negatively correlated with the share of streets named after local figures or motives, whereas exposure to more homogeneous communities is associated with a higher prevalence of local-identity street names.

These patterns are quantitatively stable across specifications and underscore sharp differences in commemorative practices across municipalities that experienced similar emigration intensity but whose emigrants settled in more *vis-á-vis* less diverse communities. Table IV separately reports the p-value of a test of equality between the coefficients of emigration flows to diverse and homogeneous communities. Across all models, we find strong evidence that the two coefficients are statistically different ($p < 0.01$).

The estimates from (9) are correlational. However, the inclusion of placebo categories provides a diagnostic for the plausibility of a causal interpretation. While we find a large and statistically significant association, with opposite signs, between historical exposure to diversity and national versus local identity in street names, we detect no systematic relationship with placebo categories. In columns (7–9), we employ the share of streets named after artistic, literary, and religious figures as the outcome variables. None of these outcomes is significantly associated with either the number of U.S. emigrants who settled in high or low-diversity communities. These patterns suggest that exposure to diversity is not merely correlated with a general propensity for street naming or commemorative practices, thus restricting the scope for alternative explanations based on unobserved municipal characteristics.

IV.C National Monuments

We use public monuments as a third proxy for national identity sentiment. We rely on a municipality-level dataset and aggregate the number of monuments to the decade level.

We estimate the following specification:

$$y_{mt} = \alpha_m + \alpha_{r(m)\times t} + X'_{mt}\Gamma + \sum_{\tau=1870}^{1990} \beta_\tau \cdot Diversity\ Exposure_m \cdot I(t=\tau) + \varepsilon_{mt}\,, \tag{10}$$

where, as in the rest of the analysis, $Diversity\ Exposure_m$ denotes the share of U.S. emigrants from municipality m who settled in districts in the top quintile of the entropy distribution. Because this analysis spans the entire twentieth century, we include region-by-decade fixed effects ($\alpha_{r(m)\times t}$) to ab-

sorb time-varying shocks at a relatively coarse geographic level, such as differential exposure to wars, reconstruction, and institutional change.

As in (7), because U.S. migration began in the late 1890s, we interpret the coefficients for the 1870s, 1880s, and 1890s as evidence on pre-treatment differences between treated and control municipalities. In this analysis, we restrict the sample to Southern municipalities for two reasons. First, national identity monuments were largely erected to commemorate casualties of the World Wars, and Northern regions were substantially more exposed to combat, so monument construction there is more likely to reflect proximity to warfare rather than identity formation. Second, given the long time horizon of the analysis, comparisons between Northern and Southern regions would confound the estimates with multiple region-specific developments that unfolded over the twentieth century.

Figure VII reports the estimated effects. In both panels, the left y-axis plots the coefficients from equation (10), while the gray line on the right y-axis shows the number of monuments by year of construction.

Figure VIIa focuses on national monuments. Most national monuments were erected in the 1920s and 1930s and commemorate soldiers who died during the First World War, with a second, quantitatively smaller wave following the Second World War. Prior to the onset of mass emigration, we find no statistically significant differences between municipalities whose U.S. emigrants later settled in more versus less diverse Italian communities. The estimates, therefore, do not conflate pre-existing differences across origin municipalities correlated with subsequent exposure to diversity in the U.S.

Beginning in the interwar period, municipalities whose migrants settled in relatively more diverse Italian communities in the United States exhibit more national monuments than control municipalities. We also estimate a statistically significant effect of exposure to diversity on the number of national monuments erected between 1950 and 1959, when most monuments commemorate WW2 casualties. This persistence suggests that national-identity values formed through emigrants' historical exposure to diversity were subsequently activated by the experience of the Second World War.

Figure VIIb replicates the analysis for non-national monuments, which we interpret as a placebo category. We find no association between exposure to Italian diversity in the United States and the construction of non-national monuments prior to the mass migration period. Likewise, exposure to diversity has no effect on the construction of non-national monuments in the post-migration period. This pattern suggests that our estimates are unlikely to capture general shifts in commemorative activity unrelated to exposure to diversity.

Table E.10 reports the estimates in tabular form under a range of alternative specifications where we control for time-varying shocks at the region (columns 2 and 7) and province (columns 3 and 8) levels. In columns (5) and (9), we control for the number of U.S. emigrants over the migration period interacted with period dummies to test whether the results are driven by the overall intensity of the migration outflows rather than by its direction. In columns (6) and (10), we add municipality-level time-invariant controls—including the number of deaths during WW1 and WW2. The baseline results remain quantitatively robust.

IV.D Participation in World War I of U.S. Emigrants

As a final measure of national identity, we follow Caprettini and Voth (2023) and use World War I death rates among individuals who emigrated to the United States and served in the Italian army. Enlistment constituted a high-stakes expression of national attachment, exposing migrants to the risk of death in the service of the Italian state. Although military service was formally universal and mandatory, Italian consulates in the United States had limited enforcement capacity, implying that enlistment for emigrants in the United States was largely voluntary in practice (Choate, 2008).

We estimate variants of the following cross-sectional regression:

$$y_m = \alpha_{p(m)} + X_m'\Gamma + \beta \cdot US\ Emigrants_m + \gamma \cdot Diversity\ Exposure_m + \varepsilon_m\,, \tag{11}$$

where, as in the previous specifications, $Diversity\ Exposure_m$ denotes the share of U.S. emigrants from municipality m who settled in districts in the top quintile of the entropy distribution. In our preferred specification, we control for province-level fixed effects, as in (9), and for the number of U.S. emigrants.

Table E.11 displays the results. Exposure to more diverse environments in the United States is strongly associated with higher death rates among returning migrants serving in the Italian army, across various specifications (columns 1–4). The magnitude is economically meaningful and stable across specifications: moving from the 25th to the 75th percentile of diversity exposure is associated with a 1.5% increase in the death rate among migrants, which is roughly equivalent to 80% of the mean.

By contrast, diversity exposure is uncorrelated with overall municipal death rates (column 5), suggesting that the relationship is specific to migrants rather than reflecting broader differences in wartime mortality across municipalities. Moreover, the intensity of U.S. emigration is uncorrelated with World War I death rates (columns 3–5), indicating that the estimates plausibly reflect the effects of exposure to diversity rather than differences in the overall scale of emigration. This pattern is consistent with greater national attachment among migrants exposed to more diverse environments abroad.

IV.E Robustness Analysis

Most Italian migrants in the United States originated from Southern regions. A potential concern, therefore, is that our results may be driven by specific origin areas. In Figure F.13, Figure F.14, and Figure F.15, we replicate the baseline analyses for names, streets, and monuments while sequentially excluding one region at a time from the estimation sample. In each panel, the first row reports the full-sample baseline estimate, and subsequent rows report the treatment effects obtained when excluding each region. Across all outcomes and specifications, the estimates remain statistically comparable to the baseline, indicating that the results are unlikely to be driven by idiosyncratic characteristics of particular regions.

A related concern is that spatial clustering of emigrants across regions or municipalities may induce spatial autocorrelation, potentially inflating statistical significance. In Figure F.16, Figure F.17,

and Figure F.18, we therefore replicate all analyses using a range of alternative standard error estimators, including the spatial correction proposed by Conley (1999) with different distance cutoffs. Statistical significance remains largely unchanged across outcomes and specifications.

We assess the sensitivity of our naming and monument results to potential violations of the parallel trends assumption using the framework developed by Rambachan and Roth (2023), which formalizes how pre-treatment deviations from inform post-treatment violations rather than assuming exact parallel trends. In our application, robust confidence intervals reported in Figure F.19 and Figure F.20 remain statistically significant for values of the tuning parameter up to one, meaning that the post-treatment departure from parallel trends would have to be at least as large as the largest pre-treatment deviation to overturn the results. These patterns indicate that the results remain robust to sizable departures from the parallel trends assumption.

V CONCLUSIONS

This paper investigates whether national identity, a key legitimizing element of nation-states, can emerge in response to intergroup contact among individuals from fragmented, heterogeneous communities. We study the Italian diaspora at the turn of the 1900s, when 5 million emigrants from diverse linguistic, political, and socioeconomic backgrounds settled in a handful of neighborhoods across America.

We exploit this unique setting to study how intergroup contact among members of a fractionalized country can foster the emergence of national identity. We construct a novel, highly detailed dataset that allows us to observe the municipality of origin of half a million Italian migrants in the United States by linking the U.S. census records and the Ellis Island ship passenger lists. We document that American neighborhoods were areas where Italians from different parts of the country met and interacted. These contacts contributed to the emergence of a sense of shared national identity even in the absence of coordinated state interventions. Increased national identity among migrants, however, hampered cultural assimilation into U.S. society.

We then show that the national identity forged within immigrant communities diffused back to Italy. Municipalities more exposed to interregional interaction through their emigrants experienced an increased sense of national identity. These patterns are difficult to reconcile with purely domestic explanations and instead point to a transnational process through which identity formation abroad reshaped cultural outcomes at home.

Taken together, our findings suggest that today's diasporic communities can play an important and underappreciated role in fostering nation-building and contribute to social cohesion in migrants' countries of origin.

REFERENCES

Abramitzky, Ran and Leah Boustan, "Immigration in American Economic History," *Journal of Economic Literature*, 2017, *55* (4), 1311–1345.

_ **, Katherine Eriksson, Santiago Pérez, and Myera Rashid**, "Census Linking Project: Version 3.0," *[dataset]*, 2025.

_ , **Leah Boustan, and Katherine Eriksson**, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 2012, *102* (5), 1832–1856.

_ , _ , **and** _ , "Do Immigrants Assimilate More Slowly Today than in the Past?," *American Economic Review: Insights*, 2020, *2* (1), 125–141.

_ , _ , _ , **James Feigenbaum, and Santiago Pérez**, "Automated Linking of Historical Data," *Journal of Economic Literature*, 2021, *59* (3), 865–918.

Acemoglu, Daron, Giuseppe De Feo, Giacomo De Luca, and Gianluca Russo, "War, Socialism, and the Rise of Fascism: An Empirical Exploration," *The Quarterly Journal of Economics*, 2022, *137* (2), 1233–1296.

Akerlof, George A. and Rachel E. Kranton, "Economics and Identity," *The Quarterly Journal of Economics*, 2000, *115* (3), 715–753.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg, "Fractionalization," *Journal of Economic Growth*, 2003, *8* (2), 155–194.

_ , **Paola Giuliano, and Bryony Reich**, "Nation-Building and Education," *The Economic Journal*, 2021, *131* (638), 2273–2303.

Althoff, Lukas, Harriet Brookes Gray, and Hugo Reichardt, "America's Rise in Human Capital Mobility," *Working Paper*, 2025.

Anderson, Benedict, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, London and New York: Verso, 1983.

Anelli, Massimo, Gaetano Basso, Giuseppe Ippedico, and Giovanni Peri, "Emigration and Entrepreneurial Drain," *American Economic Journal: Applied Economics*, 2023, *15* (2), 218–252.

Assouad, Lydia, "Charismatic Leaders and Nation Building," *Working Paper*, 2025.

Azaryahu, Maoz, "The Power of Commemorative Street Names," *Environment and planning D: Society and Space*, 1996, *14* (3), 311–330.

Bagues, Manuel and Christopher Roth, "Interregional Contact and the Formation of a Shared Identity," *American Economic Journal: Economic Policy*, 2023, *15* (3), 322–350.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo, "The Making of Modern America: Migratory Flows in the Age of Mass Migration," *Journal of Development Economics*, 2013, *102*, 23–47.

_ , **Myra Mohnen, Imran Rasul, and Martina Viarengo**, "Nation-Building Through Compulsory Schooling During the Age of Mass Migration," *The Economic Journal*, 2019, *129* (617), 62–109.

Banti, Alberto Mario, *Sublime Madre Nostra: La Nazione Italiana Dal Risorgimento Al Fascismo*, Gius. Laterza & Figli Spa, 2014.

Barsbai, Toman, Hillel Rapoport, Andreas Steinmayr, and Christoph Trebesch, "The Effect of Labor Migration on the Diffusion of Democracy: Evidence from a Former Soviet Republic," *American Economic Journal: Applied Economics*, 2017, *9* (3), 36–69.

Basse, Guillaume, Peng Ding, Avi Feller, and Panos Toulis, "Randomization Tests for Peer Effects in Group Formation Experiments," *Econometrica*, 2024, *92* (2), 567–590.

Batista, Catia, Daniel Han, Johannes Haushofer, Gaurav Khanna, David McKenzie, Ahmed Mushfiq Mobarak, Caroline Theoharides, and Dean Yang, "Brain Drain or Brain Gain? Effects of High-Skilled International Emigration on Origin Countries," *Science*, 2025, *388* (6749), eadr8861.

Bazzi, Samuel, Arya Gaduh, Alexander D Rothenberg, and Maisy Wong, "Unity in Diversity? How Intergroup Contact Can Foster Nation Building," *American Economic Review*, 2019, *109* (11), 3978–4025.

Beine, Michel, Frédéric Docquier, and Maurice Schiff, "International Migration, Transfer of Norms and Home Country Fertility," *Canadian Journal of Economics/Revue canadienne d'économique*, 2013, *46* (4), 1406–1430.

Berkes, Enrico, Ezra Karger, and Peter Nencka, "The Census Place Project: A Method for Geolocating Unstructured Place Names," *Explorations in Economic History*, 2023, *87*, 101477.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique, "The Economic Payoff of Name Americanization," *Journal of Labor Economics*, 2017, *35* (4), 1089–1116.

Billings, Stephen B, Eric Chyn, and Kareem Haggag, "The Long-Run Effects of School Racial Diver-

sity on Political Identity," *American Economic Review: Insights*, 2021, *3* (3), 267–284.

Blanc, Guillaume and Masahiro Kubo, "The Making of France," *Working Paper*, 2025.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley EB Wilbert, "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project," *Explorations in Economic History*, 2025, p. 101717.

Bursztyn, Leonardo, Thomas Chaney, Tarek A Hassan, and Aakaash Rao, "The Immigrant Next Door," *American Economic Review*, 2024, *114* (2), 348–384.

Cagé, Julia, Anna Dagorret, Pauline Grosjean, and Saumitra Jha, "Heroes and Villains: The Effects of Heroism on Autocratic Values and Nazi Collaboration in France," *American Economic Review*, 2023, *113* (7), 1888–1932.

Cantoni, Davide, Yuyu Chen, David Y Yang, Noam Yuchtman, and Y Jane Zhang, "Curriculum and Ideology," *Journal of Political Economy*, 2017, *125* (2), 338–392.

Caprettini, Bruno and Hans-Joachim Voth, "New Deal, New Patriots: How 1930s Government Spending Boosted Patriotism During World War II," *The Quarterly Journal of Economics*, 2023, *138* (1), 465–513.

Carlitz, Ruth, Ameet Morjaria, Joris Mueller, and Philip Osafo-Kwaako, "State Building in a Diverse Society," *Review of Economic Studies*, 2025, *92* (6), 3704–3740.

Carnevale, Nancy C, *A New Language, a New World: Italian Immigrants in the United States, 1890-1945*, University of Illinois Press, 2010.

Chetty, Raj and Nathaniel Hendren, "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," *The Quarterly Journal of Economics*, 2018, *133* (3), 1107–1162.

Chiaricati, Federico, "Nationalism and Nation-Building in the Dietary Consumption of Italian Migrants in the United States: A Transnational Perspective," *Modern Italy*, 2020, *25* (4), 403–419.

Choate, Mark I., *Emigrant Nation: The Making of Italy Abroad*, Harvard University Press, 2008.

Ciccarelli, Carlo and Jacob Weisdorf, "Pioneering into the Past: Regional Literacy Developments in Italy Before Italy," *European Review of Economic History*, 2019, *23* (3), 329–364.

Coluccia, Davide M and Gaia Dossi, "Return Innovation: The Knowledge Spillovers of the British Migration to the United States, 1870-1940," *Working Paper*, 2025.

_ **and Lorenzo Spadavecchia**, "Emigration Restrictions and Economic Development: Evidence from the Italian Mass Migration to the United States," *Working Paper*, 2025.

Conley, Timothy G, "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Corno, Lucia, Eliana La Ferrara, and Justine Burns, "Interaction, Stereotypes, and Performance: Evidence from South Africa," *American Economic Review*, 2022, *112* (12), 3848–3875.

Corradini, Camillo, "Relazione Presentata a S. E. Il Ministro della Pubblica Istruzione Dal Direttore Generale per L'istruzione Primaria E Popolare," in "L'istruzione primaria e popolare in Italia, con speciale riguardo all'anno scolastico 1907-1908, vol. I," Roma, 1910.

Corradini, Enrico, "L'emigrazione Italiana Nell'america del Sud (1909)," *Discorsi politici (1902–1923)*, 1923.

Covello, Leonard, *The Social Background of the Italian American School Child*, Leiden, 1967.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth, "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams," *The Quarterly Journal of Economics*, 2021, *136* (2), 987–1030.

De Mauro, Tullio, *Storia Linguistica dell'italia Unita*, Gius. Laterza & Figli Spa, 1963.

Dehdari, Sirus H and Kai Gehring, "The Origins of Common Identity: Evidence from Alsace-Lorraine," *American Economic Journal: Applied Economics*, 2022, *14* (1), 261–292.

Delille, Gérard, *Famille Et Propriété Dans le Royaume de Naples*, Vol. 259, Ecoles Françaises d'Athènes et de Rome, 1985.

Dell, Melissa and Pablo Querubin, "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies," *The Quarterly Journal of Economics*, 2018, *133* (2), 701–764.

DellaVigna, Stefano, Ruben Enikolopov, Vera Mironova, Maria Petrova, and Ekaterina Zhuravskaya, "Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia," *Amer-*

ican Economic Journal: Applied Economics, 2014, *6* (3), 103–132.

Depetris-Chauvin, Emilio, Ruben Durante, and Filipe Campante, "Building Nations Through Shared Experiences: Evidence from African Football," *American Economic Review*, 2020, *110* (5), 1572–1602.

Deschamps, Bénédicte, "De la Presse "Coloniale"' à la Presse Italo-Américaine, le Parcours de Six Périodiques Italiens aux Etats-Unis (1910-1935)." PhD dissertation, Université Paris 7 1996.

Dickie, John, *Darkest Italy: The Nation and Stereotypes of the Mezzogiorno, 1860–1900*, Springer, 1999.

Dippel, Christian and Stephan Heblich, "Leadership in Social Movements: Evidence from the "forty-Eighters" in the Civil War," *American Economic Review*, 2021, *111* (2), 472–505.

Edwards, Brent Hayes, *The Practice of Diaspora: Literature, Translation, and the Rise of Black Internationalism*, Harvard University Press, 2003.

Eriksson, Katherine and Zachary Ward, "The Residential Segregation of Immigrants in the United States from 1850 to 1940," *The Journal of Economic History*, 2019, *79* (4), 989–1026.

Esposito, Elena, Tiziano Rotesi, Alessandro Saia, and Mathias Thoenig, "Reconciliation Narratives: The Birth of a Nation After the US Civil War," *American Economic Review*, 2023, *113* (6), 1461–1504.

Felice, Emanuele, "The Roots of a Dual Equilibrium: GDP, Productivity, and Structural Change in the Italian Regions in the Long Run (1871–2011)," *European Review of Economic History*, 2019, *23* (4), 499–528.

Fernández Sánchez, Martín, "Mass Emigration and Human Capital over a Century: Evidence from the Galician Diaspora," *Working Paper*, 2022.

Foerster, Robert Franz, *The Italian Emigration of Our Times*, Harvard University Press, 1919.

Fouka, Vasiliki, "Backlash: The Unintended Effects of Language Prohibition in US Schools After World War I," *The Review of Economic Studies*, 2020, *87* (1), 204–239.

_ **, Soumyajit Mazumder, and Marco Tabellini**, "From Immigrants to Americans: Race and Assimilation During the Great Migration," *The Review of Economic Studies*, 2022, *89* (2), 811–842.

Gabaccia, Donna R, *Italy's Many Diasporas*, Routledge, 2013.

Gagliarducci, Stefano and Marco Tabellini, "Faith and Assimilation: Italian Immigrants in the US," *The Economic Journal*, 2025, *135* (670), 1814–1851.

Gellner, Ernest, *Nations and Nationalism*, Blackwell Publishers, 1983.

Gilroy, Paul, *The Black Atlantic: Modernity and Double Consciousness*, Harvard University Press, 1993.

Godlonton, Susan and Caroline Theoharides, "Diffusion of Reproductive Health Behavior Through International Migration: Effects on Origin-Country Fertility," *American Economic Review*, 2025, *115* (10), 3597–3637.

Gordon, Milton M, *Assimilation in American Life: The Role of Race, Religion, and National Origins*, Oxford University Press, 1964.

Green, A, "Networks and Geographic Mobility: Evidence from World War II Navy Ships," *Working Paper*, 2025.

Hall, Stuart, "Cultural Identity and Diaspora," in "Colonial Discourse and Post-Colonial Theory," Routledge, 2015, pp. 392–403.

Hatton, Timothy J, "Time on the Crossing: Emigrant Voyages Across the Atlantic, 1853–1913," *European Review of Economic History*, 2024, *28* (1), 120–133.

Hobsbawm, Eric and Terence Ranger, *The Invention of Tradition*, Cambridge University Press, 1983.

Johnson, Nuala, "Cast in Stone: Monuments, Geography, and Nationalism," *Environment and planning D: society and space*, 1995, *13* (1), 51–65.

Karadja, Mounir and Erik Prawitz, "Exit, Voice, and Political Change: Evidence from Swedish Mass Migration to the United States," *Journal of Political Economy*, 2019, *127* (4), 1864–1925.

Khanna, Gaurav, Emir Murathanoglu, Caroline B Theoharides, and Dean Yang, "Abundance from Abroad: Migrant Income and Long-Run Economic Development," *American Economic Review*, 2026, *Forthcoming*.

Knudsen, Anne Sofie Beck, "Those Who Stayed: Selection and Cultural Change in the Age of Mass Migration," *Working Paper*, 2024.

Lecce, Giampaolo, Laura Ogliari, and Tommaso Orlando, "State Formation, Social Unrest and Cul-

tural Distance: Brigandage in Post-Unification Italy," *Journal of Economic Growth*, 2022, *27* (3), 453–483.

Levitt, Peggy, "Social Remittances: Migration-Driven Local-Level Forms of Cultural Diffusion," *International Migration Review*, 1998, *32* (4), 926–948.

_ **and Deepak Lamba-Nieves**, "Social Remittances Revisited," *Journal of Ethnic and Migration Studies*, 2011, *37* (1), 1–22.

Lindfors, Teppo, Jakko Meriläinen, and Matti Mitrunen, "Networks of Nation-Building: Evidence from the Fennoman Movement," *Working Paper*, 2025.

Livi-Bacci, Massimo, *L'immigrazione e L'asimilazione degli Italiani negli Stati Uniti: Secondo le Statistiche Demografiche Americane*, Giuffrè, 1961.

Lowe, Matt, "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration," *American Economic Review*, 2021, *111* (6), 1807–1844.

Luconi, Stefano, "Forging an Ethnic Identity: The Case of Italian Americans," *Revue française d'études américaines*, 2003, *96* (2), 89–101.

_ , "The Pitfalls of the "italian Diaspora"',," *Italian American Review*, 2011, *1* (2), 147–176.

Manacorda, M, J Ponticelli, and A Tesei, "The International Transmission of Democratic Values: Evidence from the African Diaspora," *Working Paper*, 2024.

Melki, Mickael, Hillel Rapoport, Enrico Spolaore, and Romain Wacziarg, "Cultural Remittances and Modern Fertility," *NBER Working Paper*, 2024, (32990).

Molfese, Franco, *Storia del Brigantaggio Dopo L'unitÁ*, Feltrinelli, 1964.

Mourlane, Stéphane, Céline Regnard, Manuela Martini, and Catherine Brice, *Italianness and Migration from the Risorgimento to the 1960s*, Springer Nature, 2022.

Mylonas, Harris and Maya Tudor, "Nationalism: What We Know and What We Still Need to Know," *Annual Review of Political Science*, 2021, *24* (1), 109–132.

Okunogbe, Oyebola, "Does Exposure to Other Ethnic Regions Promote National Integration? Evidence from Nigeria," *American Economic Journal: Applied Economics*, 2024, *16* (1), 157–192.

Orsi, Robert A, *The Madonna of 115th Street: Faith and Community in Italian Harlem, 1880-1950*, Yale University Press, 1985.

Oto-Peralías, Daniel, "What Do Street Names Tell Us? The 'city-Text' as Socio-Cultural Data," *Journal of Economic Geography*, 2018, *18* (1), 187–211.

Postel, Hannah M, "Record Linkage for Character-Based Surnames: Evidence from Chinese Exclusion," *Explorations in Economic History*, 2023, *87*, 101493.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley, "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set," *Explorations in Economic History*, 2021, *80*, 101391.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 2023, *90* (5), 2555–2591.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.

Riall, Lucy, *The Italian Risorgimento: State, Society and National Unification*, Routledge, 2002.

Rohner, Dominic and Ekaterina Zhuravskaya, "The Economics of Nation-Building: Methodological Tool Kit and Policy Lessons," *Annual Review of Economics*, 2024, *17*.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, Renae Rodgers, Jonathan Schroeder, , and Kari C.W. Williams, "IPUMS USA: Version 16.0," *Minneapolis, MN: [dataset]*, 2025.

Schindler, David and Mark Westcott, "Shocking Racial Attitudes: Black GIs in Europe," *The Review of Economic Studies*, 2021, *88* (1), 489–520.

Shayo, Moses, "A Model of Social Identity with an Application to Political Economy," *American Political Science Review*, 2009, *103* (2), 147–174.

Smith, Anthony D, *The Ethnic Origins of Nations*, Blackwell Publishers, 1986.

Spilimbergo, Antonio, "Democracy and Foreign Education," *American Economic Review*, 2009, *99* (1), 528–543.

Spitzer, Yannay and Ariell Zimran, "Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925," *Journal of Development Economics*, 2018, *134*, 226–247.

_ **and** _ , "Like an Ink Blot on Paper: Testing the Diffusion Hypothesis of Mass Migration, Italy 1876-1920," *NBER Working Paper*, 2023, (30847).

Unrau, Harlan D., *Ellis Island, Statue of Liberty National Monument*, U.S. Department of the Interior, National Park Service, 1984.

Vecoli, Rudolph J, "The Formation of Chicago's "little Italies"," *Journal of American Ethnic History*, 1983, *2* (2), 5–20.

_ , "The Search for an Italian American Identity: Continuity and Change," in Lydio T Tomasi, ed., *Italian Americans: New Perspectives in Itahan Immigration and Ethnicity*, Center for Migration Studies, 1985.

Viazzo, Pier Paolo, *Upland Communities: Environment, Population and Social Structure in the Alps Since the Sixteenth Century*, Vol. 8, Cambridge University Press, 1989.

Villari, Luigi, *Gli Italiani Negli Stati Uniti D'america E L'emigrazione Italiana*, Treves, 1912.

Weber, Eugen, *Peasants into Frenchmen: The Modernization of Rural France, 1870-1914*, Stanford University Press, 1976.

Wyman, Mark, *Round-Trip to America: The Immigrants Return to Europe, 1880-1930*, Cornell University Press, 1993.

TABLES

Table I. Descriptive Statistics of Selected Variables

	Mean	Std. Dev.	Median	Min.	Max.	Observations
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. United States						
Panel A.1 Neighborhood-Level Data						
Italian Diversity	1.911	1.222	1.906	0.000	5.955	39,477
Population	1,679.481	843.645	1,553.000	13.000	32,766.000	39,477
Immigrants	24.232	13.691	22.607	0.000	84.400	39,477
Italian Immigrants	6.644	9.156	3.122	0.000	76.915	39,477
Panel A.2 Individual-Level Data						
Local Name Index	4.219	3.905	3.241	0.014	99.827	335,569
Distance-Adjusted Local Name Index	9.593	0.495	9.743	0.000	10.002	335,569
Foreign Name Index	0.759	0.153	0.786	0.000	1.000	330,247
Married to Italian from Other Municipality	0.180	0.384	0.000	0.000	1.000	317,885
log(Occupational Income)	2.458	1.356	3.045	0.000	4.394	335,569
Employed in Manufacturing	0.244	0.430	0.000	0.000	1.000	335,569
Skilled Manufacturing Job	0.154	0.361	0.000	0.000	1.000	335,569
Unskilled Manufacturing Job	0.165	0.371	0.000	0.000	1.000	335,569
Panel B. Italy						
Population	5,052.860	28,469.863	2,231.000	0.000	2,839,638.000	104,299
Number of US Emigrants	129.247	433.492	22.000	1.000	19,225.000	5,846
Share US Emigrants in Diverse Areas	0.175	0.164	0.168	0.000	1.000	5,846
Local Name Index	0.009	0.017	0.004	0.000	0.679	68,568
Distance-Adjusted Local Name Index	9,910.692	194.009	9,964.916	3,174.663	10,001.901	68,568
Number of Streets	107.014	281.237	53.000	0.000	17,873.000	8,176
Number of Streets by Designation:						
National Identity	19.327	27.333	10.000	0.000	680.000	8,176
Local Identity	35.738	118.161	14.000	0.000	5,905.000	8,176
Arts	4.396	14.274	1.000	0.000	774.000	8,176
Literature	5.807	13.449	1.000	0.000	548.000	8,176
Religious Figure	10.791	27.654	5.000	0.000	1,527.000	8,176
Number of Monuments	0.301	6.591	0.000	0.000	1,288.000	113,400
Number of National Identity Monuments	0.123	5.707	0.000	0.000	1,282.000	113,400

Notes. This table reports descriptive statistics for selected variables for the United States (Panel A) and Italy (Panel B). Panel A.1 uses census enumeration districts observed at decennial frequency from 1900 to 1930. Panel A.2 is the baseline individual-level dataset of Italian immigrants, observed in the first census in which they appear and, when possible, linked across subsequent censuses. Panel B consists of municipality-level data for Italy: population, name-based indices, and monuments are observed decennially, while all other variables are measured once per municipality. Referenced on page: 15.

Table II. Exposure to Immigrant Heterogeneity and National Identity Formation

	Ordinary Least Squares						Reduced Form		Two-Stage Least Squares	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Local Name Index										
Italian Diversity	-0.170***	-0.164***	-0.154***	-0.162***	-0.121***	-0.126***			-2.210***	-4.939***
	(0.014)	(0.015)	(0.016)	(0.014)	(0.017)	(0.020)			(0.686)	(1.330)
Ship Diversity							-0.049***	-0.047***		
							(0.009)	(0.010)		
Mean Dep. Var.	3.990	3.989	3.989	3.989	3.989	3.989	3.988	3.988	3.988	3.988
Panel B: Distance-Adjusted Local Name Index										
Italian Diversity	0.020***	0.020***	0.019***	0.019***	0.016***	0.015***			0.251**	0.560**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)			(0.114)	(0.232)
Ship Diversity							0.006***	0.005***		
							(0.002)	(0.002)		
Mean Dep. Var.	9.593	9.593	9.593	9.593	9.593	9.593	9.593	9.593	9.593	9.593
Panel C: Foreign Name Index										
Italian Diversity	0.017***	0.016***	0.016***	0.015***	0.015***	0.014***			0.105***	0.215***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)			(0.024)	(0.055)
Ship Diversity							0.002***	0.002***		
							(0.000)	(0.000)		
Mean Dep. Var.	0.758	0.758	0.758	0.758	0.758	0.758	0.758	0.758	0.758	0.758
Municipality FE	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	–	–	–	–	–	–	–	–	–
Year FE	Yes	Yes	–	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hexagon FE	No	Yes	–	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Municipality-Year FE	No	No	Yes	No	No	No	No	No	No	No
Hexagon-Year FE	No	No	Yes	No	No	No	No	No	No	No
Individual Controls	No	No	No	Yes	No	No	No	No	No	No
Demographic Controls	No	No	No	No	Yes	Yes	No	Yes	No	Yes
Economic Controls	No	No	No	No	No	Yes	No	Yes	No	Yes
# Individuals	223,721	223,531	221,913	223,521	223,531	223,531	208,845	208,845	208,845	208,845
# Observations	334,668	334,303	330,615	334,283	334,303	334,303	311,848	311,848	311,848	311,848
K-P *F*-stat	–	–	–	–	–	–	–	–	24.327	30.750

Notes. This table reports the effect of exposure to diversity within the Italian community on name-based indices from specification (4). Panel A uses the local name index defined in (1) as the dependent variable; Panel B uses the distance-adjusted local name index defined in (2); Panel C uses the foreign name index of Abramitzky et al. (2020). The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. Columns (1–6) report ordinary least squares estimates; columns (7–8) report reduced-form instrumental-variable estimates using the ship-based instrument defined in (5); columns (9–10) report the corresponding two-stage least squares estimates. All specifications include municipality and census-year fixed effects. Column (1) includes county fixed effects; columns (2) and (4–10) replace county with more granular hexagon fixed effects. Column (3) replaces municipality fixed effects with municipality-year fixed effects and hexagon fixed effects with hexagon-year fixed effects. Demographic controls (included in columns 5, 8, and 10) are measured at the district level and include population, the share of women, African Americans, immigrants, and Italian immigrants. Economic controls (included in columns 6, 8, and 10) are measured at the district level and include employment shares by industry. Standard errors, clustered at the district level, are reported in parentheses. Columns (9–10) also report the Kleibergen–Paap *F*-statistics. Referenced on page: 20.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table III. Name Particularism in Italy by Exposure to Immigrant Diversity

	Exposure Panel			Difference-in-Differences		
	(1) LNI	(2) LNI	(3) D-LNI	(4) LNI	(5) LNI	(6) D-LNI
Diversity Exposure $\times$ $I(1880 < t \leq 1890)$	-0.002	0.086	0.014			
	(0.249)	(0.284)	(0.029)			
Diversity Exposure $\times$ $I(1890 < t \leq 1900)$	-0.017	0.054	0.011			
	(0.249)	(0.283)	(0.029)			
Diversity Exposure $\times$ $I(1900 < t \leq 1910)$	-0.268	-0.039	0.030			
	(0.265)	(0.302)	(0.031)			
Diversity Exposure $\times$ $I(1910 < t \leq 1920)$	-0.505**	-0.273	0.065**			
	(0.250)	(0.284)	(0.029)			
Diversity Exposure $\times$ $I(1920 < t \leq 1930)$	-0.798***	-0.596**	0.087***			
	(0.258)	(0.293)	(0.030)			
Diversity Exposure $\times$ Post Mass Migration				-0.464***	-0.405***	0.045***
				(0.091)	(0.097)	(0.010)
Post Mass Migration				0.010	0.055*	-0.001
				(0.025)	(0.029)	(0.003)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
US Emigrants $\times$ Year FE	No	Yes	No	No	Yes	No
# Municipalities	7,823	5,715	7,823	7,823	5,715	7,823
# Observations	200,774	155,781	200,774	200,774	155,781	200,774
Mean Dep. Var.	0.879	0.886	9.909	0.879	0.886	9.909

Notes. This table reports the effect of exposure to diversity within the Italian migrant community on name-based indices in Italian municipalities. The unit of observation is a municipality observed at a decennial frequency between 1880 and 1930. The dependent variables are the local name index defined in (1) in columns (1–2) and (4–5) and the distance-adjusted local name index defined in (2) in columns (3) and (6). The baseline treatment is defined as the share of a municipality's US emigrants who settled in districts in the top quintile of the diversity distribution (3). Columns (1–3) report event-study specifications of (7), where we interact diversity exposure with decade indicators. Columns (4–6) report difference-in-differences specifications of (8), where the post indicator equals one in years at or after the municipality-specific onset of mass migration (defined as the first year with at least one emigrant per 1,000 inhabitants); the table reports the coefficient of interaction terms between the diversity exposure measure and the post indicator. All regressions include municipality and year fixed effects. Columns (2) and (5) additionally include interactions between the number of US emigrants and year fixed effects. Standard errors are clustered at the municipality level and reported in parentheses. Referenced on pages: 26, 27.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table IV. Street Names and Exposure to Immigrant Diversity

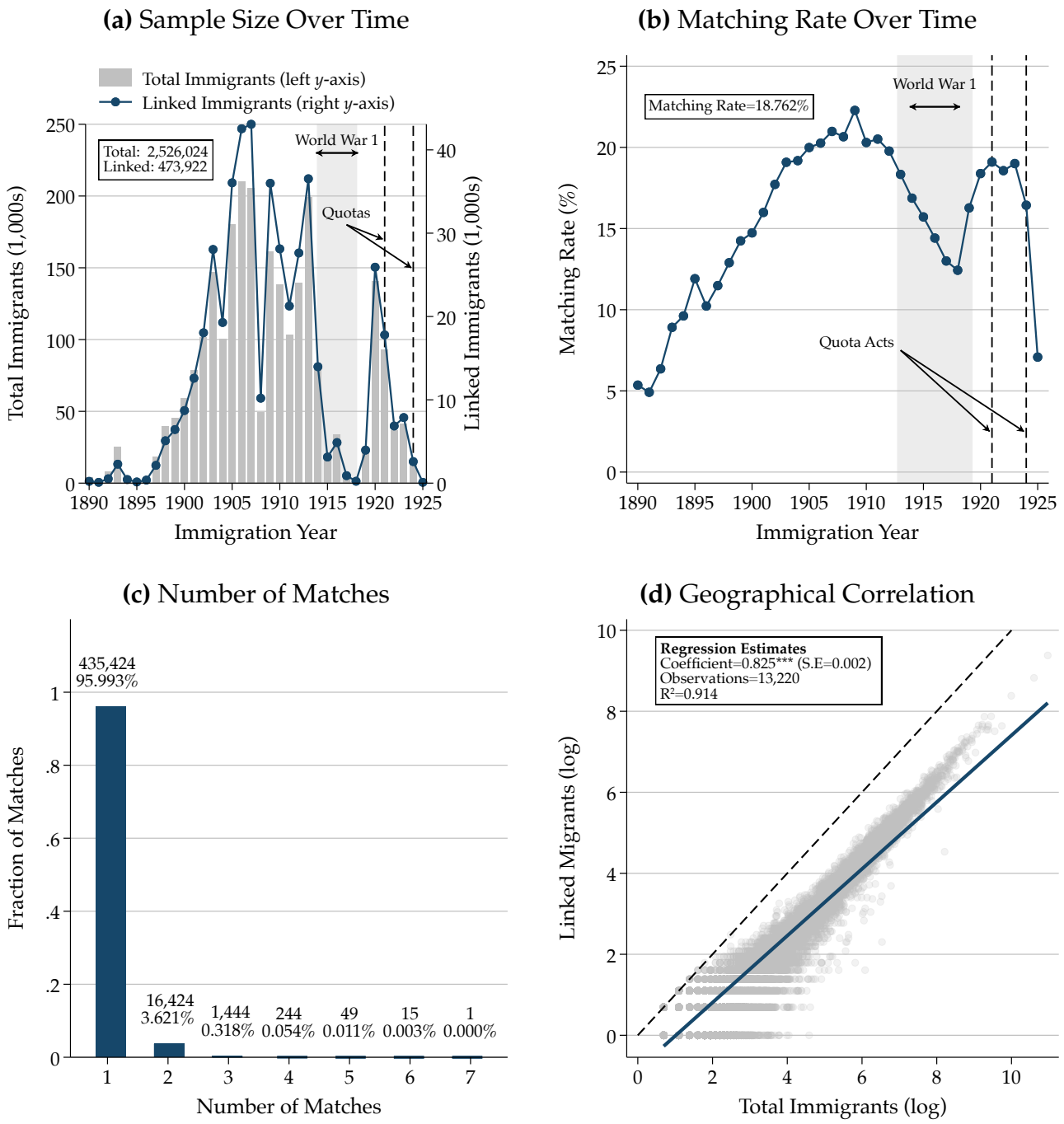
	National Identity Designation			Local Identity Designation			Placebo Designation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7) Arts	(8) Literature	(9) Religious
Number of Emigrants by Neighborhood Diversity:									
Diverse (β_1)	0.326***	0.191**	0.344***	-0.251**	-0.177	-0.317*	0.008	0.048	-0.024
	(0.079)	(0.084)	(0.095)	(0.115)	(0.122)	(0.192)	(0.027)	(0.034)	(0.053)
Homogeneous (β_2)	-0.108***	-0.074***	-0.105***	0.108***	0.092**	0.098**	-0.001	-0.014	0.010
	(0.024)	(0.025)	(0.025)	(0.036)	(0.039)	(0.049)	(0.007)	(0.009)	(0.014)
Region FE	Yes	–	Yes	Yes	–	Yes	–	–	–
Province FE	No	Yes	No	No	Yes	No	Yes	Yes	Yes
Controls FE	No	No	Yes	No	No	Yes	No	No	No
p-val: $\hat{\beta}_1 = \hat{\beta}_2$	0.000	0.002	0.000	0.004	0.006	0.001	0.543	0.079	0.266
# Municipalities	8,176	8,176	8,099	8,176	8,176	8,099	8,176	8,176	8,176
Mean Dep. Var.	0.192	0.192	0.194	0.298	0.298	0.299	0.027	0.041	0.105

Notes. This table reports cross-sectional estimates of specification (9) between emigration to diverse Italian communities in the United States and street-naming patterns in Italy. The unit of observation is a municipality. The dependent variables are the shares of streets whose eponyms reflect national identity (columns 1–3) and local identity (columns 4–6); and additional placebo categories (arts, literature, religious, in columns 7–9). The main explanatory variables are the number of emigrants to diverse neighborhoods and the number of emigrants to homogeneous neighborhoods. The outcomes are normalized by the total number of streets in the municipality. Columns (1–3) report national-identity eponyms; columns (4–6) refer to local-identity eponyms; columns (7–9) report placebo outcomes. Columns (1) and (4) include region fixed effects; columns (2) and (5) replace these with province fixed effects; columns (3) and (6) additionally include baseline municipality controls (population in 1881 and 1991, surface area, and elevation). Standard errors are clustered at the municipality level and reported in parentheses. The table also reports p-values from tests of equality between the coefficients on diverse and homogeneous emigration flows. Referenced on page: 28.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

FIGURES

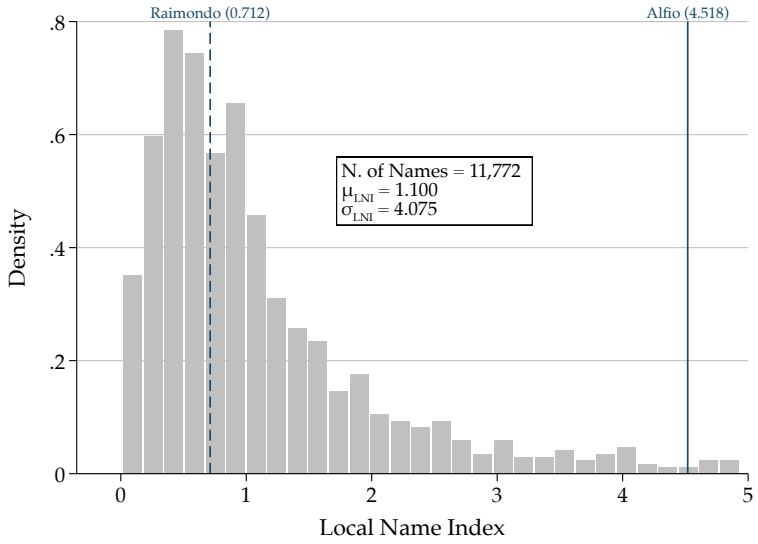
Figure I. Descriptive Statistics on Linked Migrant Sample



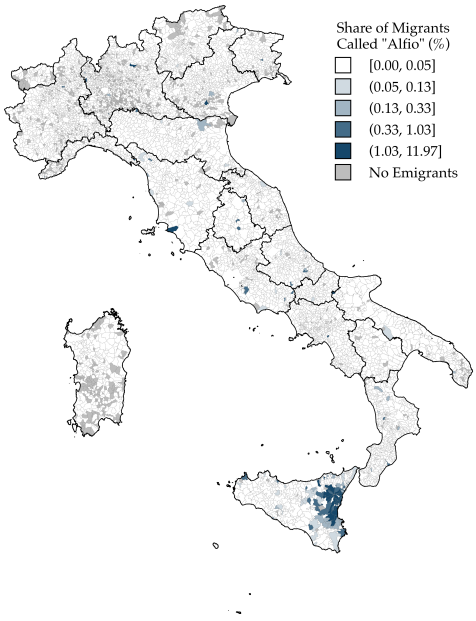
Notes. This figure reports descriptive statistics for the linked sample of U.S. census records of Italian migrants matched to Ellis Island ship records. Panel Ia shows the number of migrants in the Ellis Island data (gray bars) and the number successfully linked to U.S. census records (red bars). Panel Ib reports the share of U.S. census records that match at least one Ellis Island record. Panel Ic displays the distribution of the number of Ellis Island matches per U.S. census record, together with the corresponding shares relative to the total sample. Panel Id plots, at the municipality level, the number of migrants recorded in the Ellis Island dataset against the number appearing in the linked dataset, and includes a linear fit, the 45-degree line, and the estimated regression coefficient. Referenced on pages: 10, B16, B17.

Figure II. Distribution Local Name Index and Sample Names

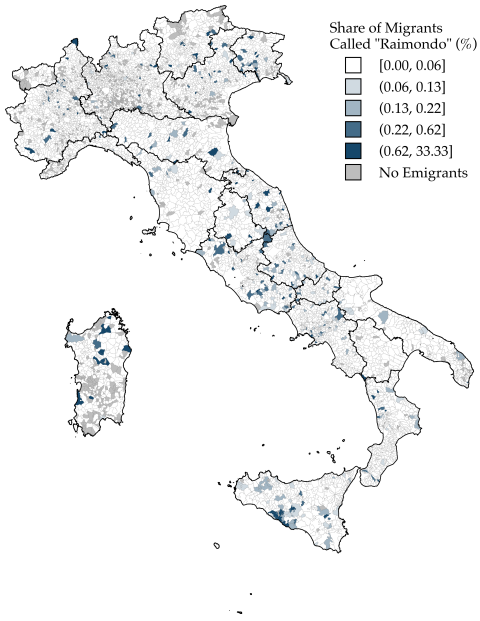
(a) Distribution of Local Name Index



(b) Distribution of Local Name

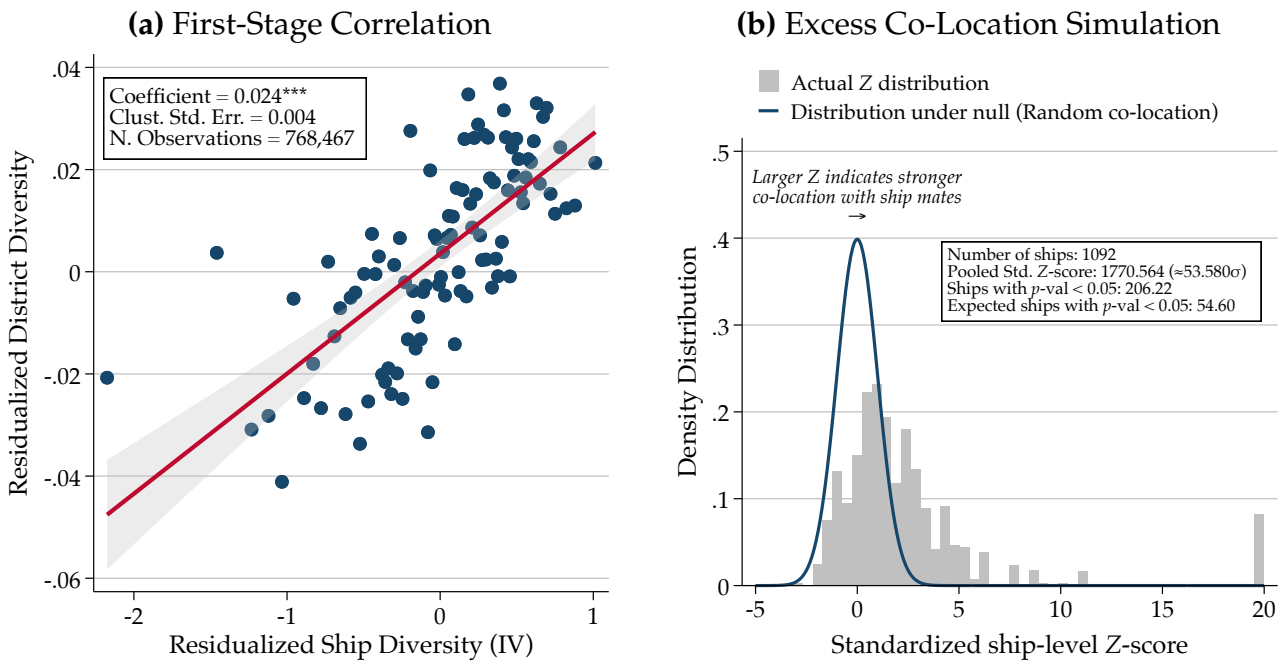


(c) Distribution of National Name



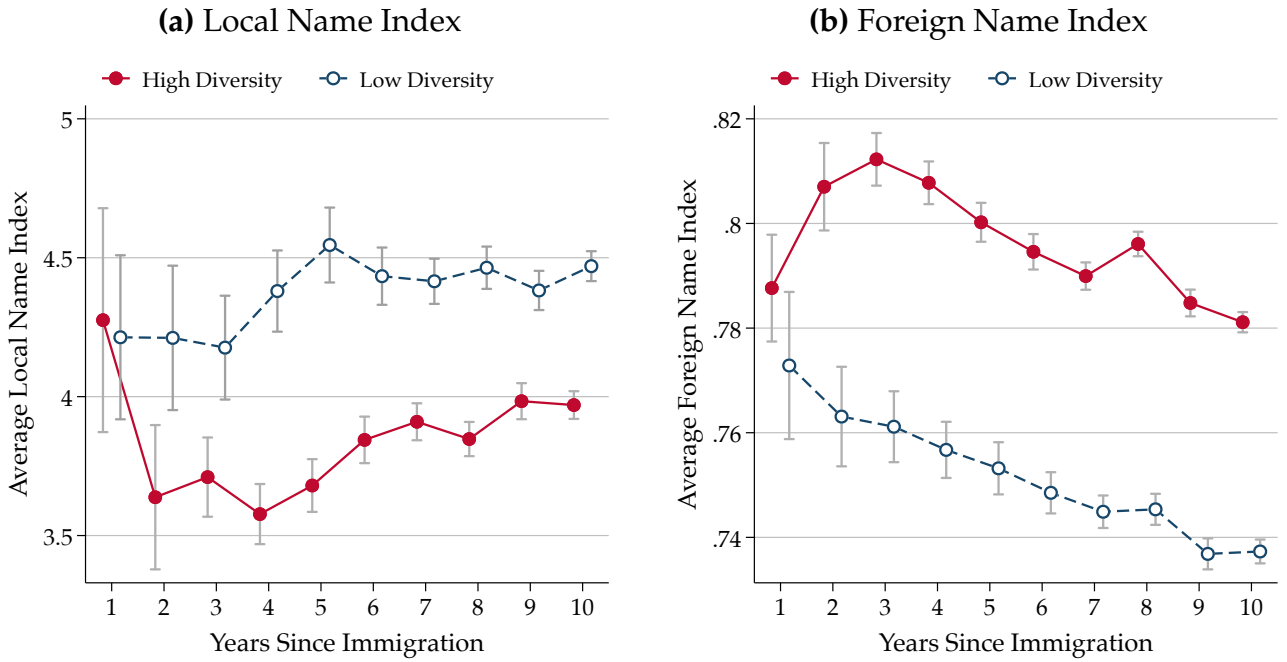
Notes. Panel IIa reports the distribution of the local name index defined in (1), together with the number of recorded names and the mean and standard deviation of the distribution. The panel also reports the LNI values for two names—"Raimondo" and "Alfio"—chosen because they appear with similar frequency. Panel IIb (respectively, IIc) maps the spatial distribution of individuals named "Alfio" (respectively, "Raimondo") across municipalities, expressed as the share of individuals with that name relative to the 1881 population, in percentage terms. The maps superimpose regional borders; municipality and district boundaries are shown using 1991 borders. Referenced on page: 11.

Figure III. Instrument Relevance Tests



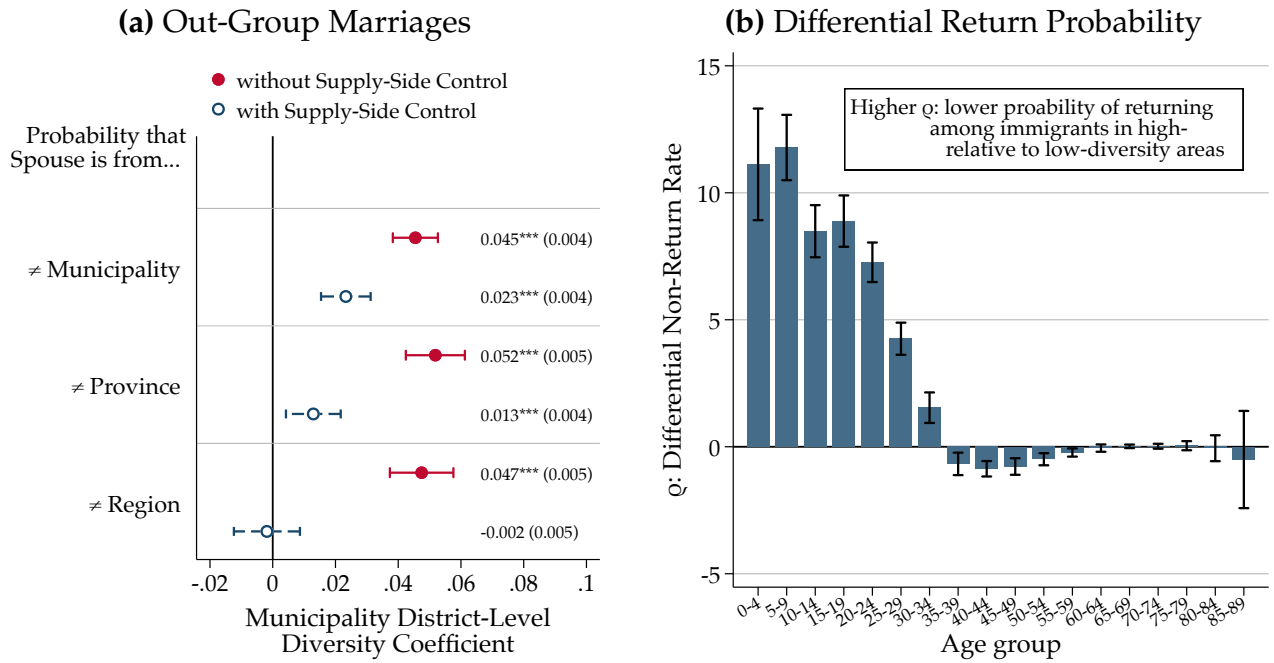
Notes. This figure reports evidence on the relevance of the ship-based instrument defined in (5). Panel IIIa shows the first-stage relationship between the instrument and observed diversity, as defined in (3). The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. To construct the binned scatterplot, both variables are residualized with respect to the baseline controls included in (7)—municipality, district, and year fixed effects. The red line displays a linear fit; the corresponding regression coefficient and standard error are reported in the panel. Panel IIIb reports the ship-level distribution of the z-statistic measuring excess co-location with shipmates relative to the benchmark case of random location with shipmates. Under the null hypothesis of no co-location, the z-score distribution would be approximately normal, shown by the blue line. Values to the right indicate excess co-location with shipmates. The panel also reports the number of ships with a p-value below 0.05, together with the number that would be expected under the null of no co-location. Referenced on pages: 18, A7.

Figure IV. Evolution of the Local and Foreign Name Indices



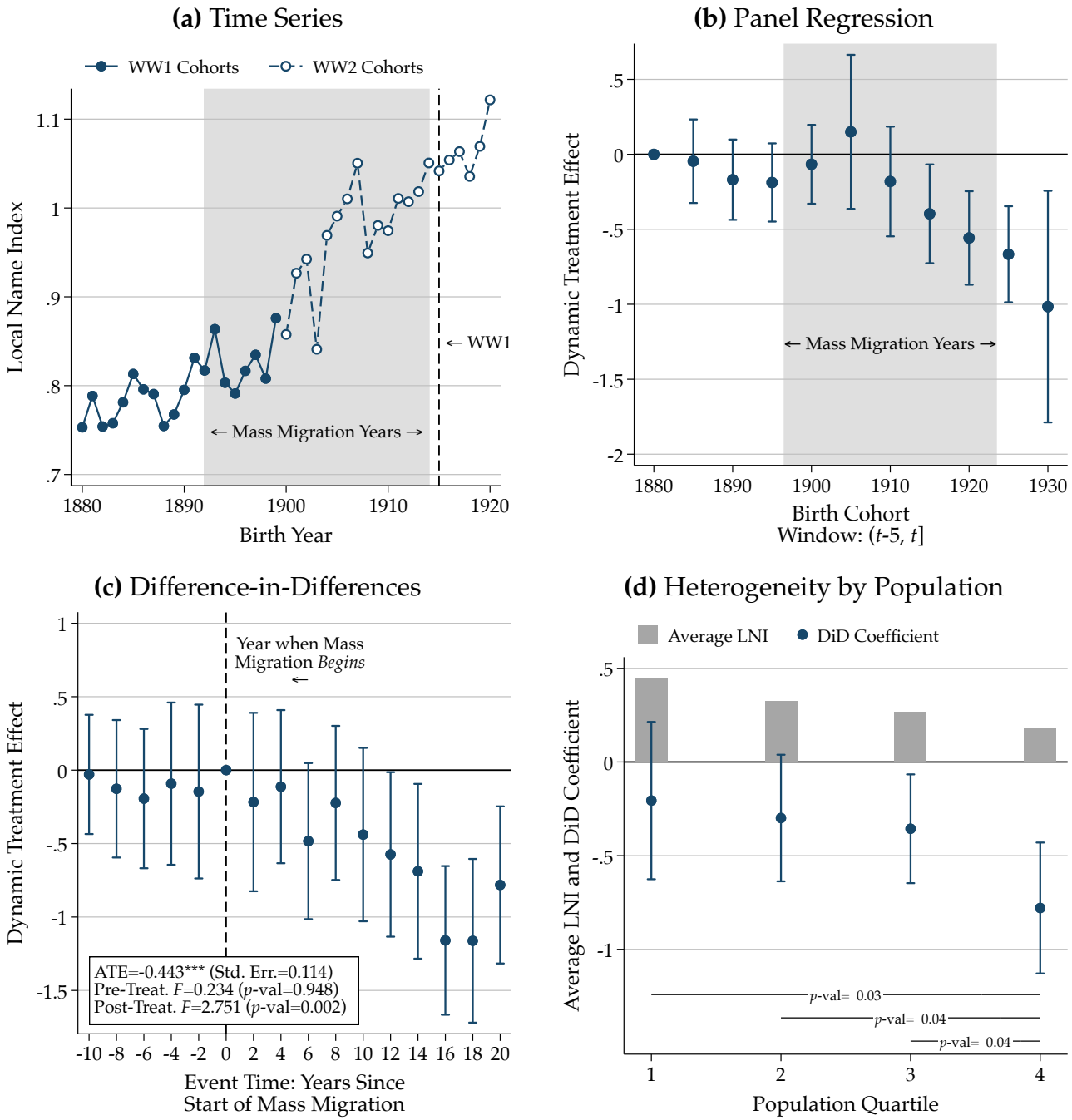
Notes. This figure reports the evolution of name-based indices by time since immigration, separately for immigrants settling in high- versus low-diversity Italian communities. The unit of observation is an individual Italian immigrant observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. Time since immigration is defined as the difference between year of immigration and census year, and individuals are split into above- and below-median groups based on the municipality-level diversity measure (3). Panel IVa plots the local name index by years since immigration. Panel IVb plots the foreign name index by years since immigration. Each point is the coefficient from a separate regression run at a given event time and estimated separately for the high- and low-diversity groups. All regressions include hexagon and census-year fixed effects. Bands report 90% confidence intervals. Referenced on page: 19.

Figure V. Additional Indicators of National Identity Formation



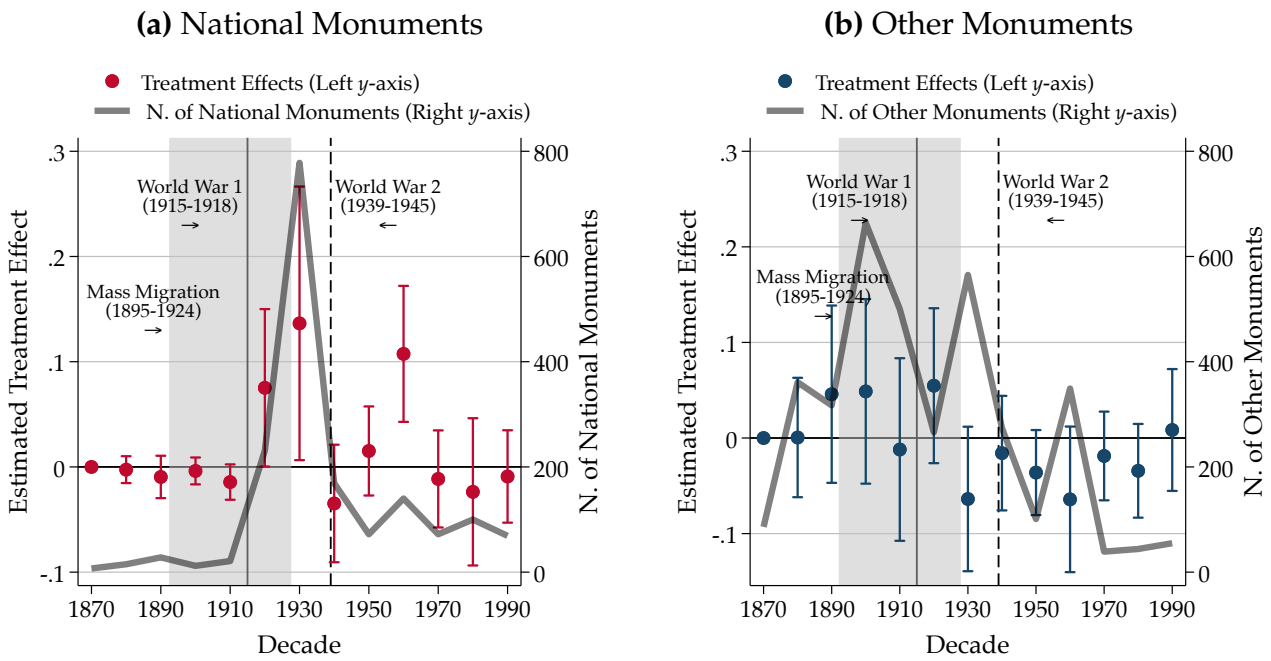
*Notes.*This figure reports the association between alternative measures of identity formation and exposure to Italian diversity. Panel Va shows regression estimates relating district-level Italian-community diversity to intermarriage outcomes. The unit of observation is an individual Italian immigrant observed between 1900 and 1930. All specifications include municipality, year, and hexagon fixed effects and control for demographic and economic district characteristics. Dots report the coefficient on Italian diversity as defined in (3). Dashed estimates additionally include supply-side controls capturing the predicted availability of same-origin spouses at the municipality, province, or region level, following Bazzi et al. (2019). Standard errors are clustered at the hexagon level and bands report 95% confidence intervals. Panel Vb plots estimates of the differential return statistic defined in (6) by age group at a 30-year horizon. High-diversity districts are those above the median of Italian diversity. For each age bin, the statistic compares the 30-year forward match rate of Italian immigrants with that of natives across high- and low-diversity districts. Standard errors are obtained via bootstrap, and bands report 95% confidence intervals. Referenced on pages: 22, 23.

Figure VI. Name Particularism in Italy by Exposure to Immigrant Diversity



Notes. This figure reports dynamic estimates of the effect of exposure to diversity within the Italian migrant community on the local name index in Italian municipalities between 1880 and 1930. Panel VIa reports the average LNI across municipalities over time. In the rest of the figures, the unit of observation is a municipality. Panel VIb reports an event-study specification (7) that interacts diversity exposure with five-year windows, controlling for municipality fixed effects, time fixed effects, and interactions between the number of US emigrants and time indicators. Panel VIc reports a difference-in-differences event-study specification (8) where the treatment time is defined as the first year with at least one emigrant per 1,000 inhabitants. In both panels, time dummies are interacted with a treatment defined as the share of a municipality's US emigrants who settled in districts in the top quintile of the diversity measure (3). Panel VId, the gray bars report the average LNI by municipality population quartiles; the blue dots report the coefficients of interaction terms between the baseline DiD coefficient (conflated in pre- and post-treatment periods) and population quartile dummies. Standard errors are clustered at the municipality level. Bands report confidence 95% confidence intervals. Referenced on pages: 26, 27.

Figure VII. Monuments by Designation in Italy



Notes. This figure reports dynamic estimates obtained from (10) of the effect of exposure to diversity within the Italian migrant community on monument construction in Southern Italy. The unit of observation is a municipality-decade for municipalities observed between 1861 and 2000. In panel VIIa, the dependent variable is an indicator equal to one if the municipality has at least one national-identity monument in a given decade. In panel VIIb, the dependent variable is an indicator equal to one if the municipality has at least one non-national-identity monument in a given decade. In both panels, each dot reports, on the left *y*-axis, the coefficient of a set of interaction terms between decade indicators and the share of US emigrants from each municipality who settled in districts in the top quintile of the diversity measure defined in (3). In both panels, the specification includes municipality fixed effects, decade fixed effects, and interactions between the number of US emigrants and decade indicators. The gray line reports, on the right *y*-axis, the total number of monuments in each category. Standard errors are clustered at the municipality level. Bands report 95% confidence intervals. Referenced on page: 29.

ONLINE APPENDIX

A Nation of Emigrants: Transatlantic Emigration and the Rise of Italian National Identity, 1890–1940

Davide M. Coluccia, Giampaolo Lecce, Laura Ogliari

March, 2026

TABLE OF CONTENTS

A DATA APPENDIX AND ROBUSTNESS CHECKS **A2**

A.I Measurement of Completed Fertility

A.II Street Data

A.III Monuments Data

A.IV Linking WW1 Death Recordswith the Ellis Island Database

A.V Instrumental Variable: First-Stage Simulation Test

A.VI Movers Design

A.VII Conley Standard Errors with High-Dimensional Fixed Effects

B LINKED SAMPLE OF ELLIS ISLAND AND POPULATION CENSUS RECORDS **B10**

B.I The Ellis Island Immigration Data

B.II Motivation and High-Level Description of the Approach

B.III Technical Description of the Algorithm Implementation

B.IV Internal Validation and Representativeness of the Linked Sample

B.V External Validation of the Linked Sample

C ADDITIONAL HISTORICAL BACKGROUND **C18**

C.I Geographic Concentration and Localism in Migrant Communities

C.II Cultural Fragmentation in Italy

C.III Cultural Convergence in Immigrant Communities

C.IV Return Migration and Feedback Effects to Italy

D A MODEL OF IMMIGRANT GROUP IDENTITY CHOICE **D23**

E ADDITIONAL TABLES **E24**

F ADDITIONAL FIGURES **F35**

A DATA APPENDIX AND ROBUSTNESS CHECKS

A.I Measurement of Completed Fertility

The decennial U.S. Census provides cross-sectional snapshots of households and therefore does not directly observe completed fertility. To recover completed fertility histories, we combine household structure information within each census year with longitudinal individual links across censuses to reconstruct parent-child relationships over time.

We begin by identifying Italian migrants and second-generation individuals in each census year (1900, 1910, 1920, and 1930), defined as individuals either born in Italy or with at least one parent born in Italy. Using household identifiers and relationship-to-head variables, we classify individuals as parents or children within each household and retain only nuclear-family relationships. For each household, we identify up to two parents (male and female) and link each child to the corresponding parent identifiers. This procedure yields a set of parent–child observations for each census year.

Because children may appear in different households or census years as families age, relying on a single census snapshot would undercount fertility. We therefore exploit longitudinal links across censuses to aggregate children observed in different years (Price et al., 2021; Buckles et al., 2025; Abramitzky et al., 2025). Specifically, for each census year taken as a reference, we employ the longitudinal crosswalks that link individual identifiers across census waves to map parents observed in other census years to the reference year. We then pool all parent–child links across census waves and collapse them at the parent level, removing duplicate children identified by year of birth and first name.

The resulting dataset assigns to each parent the full set of children ever observed co-residing with them in any census between 1900 and 1930. This approach recovers fertility histories that are substantially closer to completed fertility than those obtained from a single census snapshot, while relying solely on observed household relationships and longitudinal links.

A.II Street Data

We collect the universe of street names in Italy from the online repository `OpenStreetMaps`. Streets are grouped into municipal, provincial, regional, and state roads and highways. The national government is responsible for state roads and highways, while provinces and regions are responsible for provincial and regional roads, respectively. Since our analysis leverages local street naming decisions, we exclude non-municipal roads from the sample. We thus have information on 849,124 streets, which is approximately 93.7% of the entire sample of roads.

The dataset contains the name and coordinates of each road. We geo-reference roads to the municipality where they are located in 1991, i.e., the `GIS` file we employ throughout the analysis on Italian municipalities.

We employ a large language model (`GPT5`) to infer the motive for the designation of each road. We employ the following prompt to categorize designation motives into local or national identity. Importantly, we give the LLM both the eponym and the municipality of each road and instruct it to

identify local names using that information. Hence, names of local saints are identified as signals of local identity only where they are the subject of specific veneration.

Second, we ask the LLM to identify three "placebo motives" for designation, namely, artistic, literary, and religious figures. We employ this information as a falsification exercise in the analysis. Importantly, we do not constrain the model to assign road names to a single category. Hence, local saints, for example, are flagged as signifying both local identity and a religious motive. Similarly, we do not impose that each road must be classified in one category.

Overall, the algorithm identifies 33% of streets as signaling local identity, 18% as national identity, 4% as designated after (visual) artistic figures, 5% after literary figures, and 10% after religious figures. Approximately 40% of all roads are not classified under any such motive.

`gpt-5` Prompt for Street Classification

You are an expert on Italian cultural and political history and street toponymy.
Input:

- Street name: `{street}`
- Municipality: `{comune}`

Task: Interpret what this name commemorates in the context of Italian public toponymy and evaluate how strongly it expresses different collective identities.
Use historical and cultural knowledge about Italy to understand what the name symbolizes to Italians — for example, "Via Garibaldi" signals national identity because Garibaldi is recognized as a hero of unification.
Definitions of dimensions:

1. `national_identity` — reference to the Italian nation, its unification, or explicitly national heroes, institutions, or symbols.
   Examples: Garibaldi, Cavour, Mazzini, Vittorio Emanuele II, Italia, Unità d'Italia, XX Settembre, 25 Aprile.
2. `local_identity` — reference to regional or municipal belonging, local saints, local landmarks, or figures of local history.
   Examples: San Gennaro (in Naples), Sant'Ambrogio (in Milan).
3. `religious_identity` — reference to religion, spirituality, or ecclesiastical institutions.
   Examples: San Pietro, Santa Maria, Papa Giovanni XXIII, Madonna, Convento, Santuario.
4. `literature_identity` — reference to literary figures, authors, poets, or literary works.
   Examples: Dante, Petrarca, Manzoni, Leopardi, Pirandello.
5. `art_identity` — reference to artists, painters, sculptors, or art movements.
   Examples: Michelangelo, Leonardo, Raffaello, Caravaggio, Bernini.

Rules:

- Base scores on how the name would be understood by an informed Italian observer aware of national culture and history.
- Scores = 1 if the name clearly evokes the dimension; 0 if it does not.
- Multiple 1 for one street are possible (e.g., a saint who is also a local patron has `local_identity` = 1 and `religious_identity` = 1).
- If the meaning cannot be inferred confidently, set all scores = 0 and `named_entity = null`.
- If the commemorative meaning is unclear or generic, assign all zeros.

Return:

```
{
"street_name": "{street}",
"comune": "{comune}",
"named_entity": "<person, place, event, or concept commemorated>",
"national_identity": float,
"local_identity": float,
"religious_identity": float,
"literature_identity": float,
"art_identity": float
}
```

A.III Monuments Data

We collect buildings of significant cultural significance as identified by the Italian Ministry of Culture. This repository covers the universe of heritage sites across the Italian territory. Overall, the dataset covers 128,086 sites and includes, among other information, construction and designation years, addresses and coordinates, site names and typologies, a historical text documenting the monument's historical significance, and an artistic description of the site's aesthetics.

We exclude 24,369 sites with a missing construction year and 5,702 sites with missing coordinates. Moreover, we exclude sites erected before the Italian unification (1861) and after 2000. We exclude sites older than 1861 because regional coverage before unification is uneven, possibly reflecting the heterogeneous designation criteria among pre-unitary states. This leaves us with 32,369 sites. Among these, 44 are georeferenced in areas that are no longer part of Italy (Fiume, Dalmatia, and various islands that are now part of Greece). We discard these sites, yielding a sample of 32,325 monuments.

As with the street data, we assign monuments to municipalities at the 1991 borders to ensure geographical consistency.

We employ the following prompt to infer whether each monument is associated with a national identity motive. In most cases, national identity monuments commemorate soldiers from the munic-

ipality who died during World Wars 1 and 2.

gpt-5 Prompt for Monument Classification

You are an expert in Italian history and culture.
You will be provided with a description of a monument that commemorates a historical event, figure, or cultural aspect.
Your task is to determine whether the monument commemorates Italian national identity.
Information:

- Monument Category: `{category}`
- Historical Information: `{historical_information}`
- Description: `{description}`

Return a `JSON` object with the following structure:
`{`
`"national_identity"`: 1 if the monument commemorates Italian national identity, 0 otherwise
`}`

A.IV Linking WW1 Death Recordswith the Ellis Island Database

We link World War I death records to the matched Ellis Island–Census database described in II.A. Because we do not observe whether emigrants returned to Italy, we adopt a conservative approach that combines fuzzy string matching with information on individuals' municipalities of origin and years of birth.

Consider an individual i in the World War I death records, with first name n_i, surname s_i, year of birth b_i, and with origin municipality coordinates (x_i, y_i). Let the Ellis Island–Census database contain individuals j with standardized names (n_j, s_j), birth year b_j, and origin municipality coordinates (x_j, y_j). We begin by restricting attention to a set of candidate matches, retaining only individuals j in the Ellis Island–Census data whose year of birth falls within a two-year window of b_i and whose first and last names share the same initial letters. These restrictions are employed in the census-linking literature to allow for misreporting and reduce computational time (Abramitzky et al., 2021).

Within this candidate set, we implement fuzzy string matching to allow for transcription errors and minor spelling differences. For each pair (i, j), we compute the Jaro–Winkler similarity between first names and surnames, denoted $JW(n_i, n_j)$ and $JW(s_i, s_j)$. We classify a pair as a potential match if both similarities exceed 0.95.

We further refine the set of matches using spatial proximity. For each pair (i, j), we compute the great-circle distance d_{ij} between (x_i, y_i) and (x_j, y_j). Starting from a radius of one kilometer around municipality m_i, we iteratively expand the search radius until at least one candidate satisfies the distance criterion. Let $r^\star$ denote the smallest radius such that $d_{ij} \leq r^\star$ for at least one candidate. We retain all candidates with $d_{ij} \leq r^\star$ and discard those located farther away.

This procedure yields, for each World War I record i, a set of linked Ellis Island–Census records that are close in birth year, highly similar in both first and last name, and geographically consistent with the reported municipality of origin. In the final sample, we discard all matches with $r^* > 50$ Km, although results are robust to alternative thresholds between 20 and 100 kilometers. Furthermore, we retain only single matches, as multiple matches largely reflect spurious links due to common names and surnames.

The final sample comprises 16,847 U.S. emigrants who died while serving in the Italian army during World War I, corresponding to roughly 3% of all Italian military deaths. Approximately 2.2% of individuals who migrated to the United States died during the war, compared to 1.9% in the overall population. This comparison, however, partly reflects compositional differences, as the emigrant population was disproportionately young and male.

A.V Instrumental Variable: First-Stage Simulation Test

This appendix describes the logic and implementation of the ship-level co-location diagnostic used to assess the plausibility of the first-stage mechanism underlying the instrumental variable strategy.

The instrumental variable strategy exploits variation in ship-level origin composition to generate quasi-exogenous differences in post-arrival exposure to other Italian origins. A necessary condition for relevance is that interactions formed during the transatlantic voyage affected subsequent settlement decisions. Under this hypothesis, passengers traveling on the same ship should display excess residential co-location after arrival, relative to what would be expected based solely on aggregate destination sizes.

The diagnostic implemented here provides a direct test of this implication. It evaluates whether shipmates from different municipalities are disproportionately likely to settle in the same U.S. enumeration districts, after accounting for the distribution of district sizes and removing the mechanical contribution of the ship itself. This test applies the logic of Basse et al. (2024) to construct an exact finite-sample randomization test under a null of conditional exchangeability, in which post-arrival district assignments are independent across shipmates once aggregate destination capacities are held fixed.

Consider a ship-year s with n_s passengers indexed by $j = 1, \ldots, n_s$. Let $d(j)$ denote the enumeration district in which passenger j is observed in the census, and let $m(j)$ denote the municipality of origin. For each district d, define n_{sd} as the number of passengers from ship s residing in district d, and n_{smd} as the number of passengers from ship s residing in district d and originating from municipality m.

We define the observed cross-origin co-location statistic as

$$T_s^X = \sum_d \left[\binom{n_{sd}}{2} - \sum_m \binom{n_{smd}}{2} \right], \tag{A.1}$$

which counts the number of pairs of shipmates who settle in the same district but originate from different municipalities. By construction, this statistic excludes within-municipality pairs and therefore does not mechanically capture pre-existing origin-based networks.

To construct a benchmark under the null of no ship-based settlement ties, we simulate counterfactual district assignments for the passengers of ship s. Let N_{dt} denote the total number of individuals observed in district d in census year t, and let n_{sdt} denote the contribution of ship s to district d. We define a leave-one-out district capacity vector

$$N_{dt}^{(-s)} = N_{dt} - n_{sdt}\,, \tag{A.2}$$

which removes the mechanical contribution of ship s from the destination distribution.

Under the null hypothesis, the n_s passengers of ship s are assigned to districts without replacement, with probabilities proportional to $N_{dt}^{(-s)}$. Municipality labels $m(j)$ are held fixed across simulations. For each simulated assignment, we recompute the statistic T_s^X, yielding a simulated distribution $\{T_{s,r}^{X,MC}\}_{r=1}^{R}$.

For each ship-year s, we compute the upper-tail Monte Carlo p-value

$$p_s^{MC} = \Pr\left(T_s^{X,MC} \geq T_s^X\right)\,, \tag{A.3}$$

as well as the associated standardized statistic based on the simulated mean and variance. A systematic excess of low p_s^{MC} values across ship-years indicates that shipmates from different municipalities co-locate more than would be expected based on district sizes alone.

Figure IIIb provides strong empirical support for the maintained first-stage mechanism. The distribution of ship-level standardized co-location statistics is sharply shifted relative to the null distribution implied by random assignment, with a large mass at positive values. In the pooled sample, the aggregate standardized statistic is extremely large, reflecting pervasive excess co-location across ship-years, and the number of ships with p-values below conventional significance thresholds substantially exceeds what would be expected under the null. Importantly, these patterns arise from cross-municipality co-location and are computed using a leave-one-out assignment scheme, ruling out mechanical explanations based on same-origin networks or district size effects. Taken together, the evidence indicates that ship-level interactions shaped post-arrival settlement patterns, thereby validating the relevance of ship-based variation as a source of exogenous shifts in local exposure to Italian diversity.

A.VI Movers Design

We complement the analysis presented in Section III with a design along the lines of Chetty and Hendren (2018), which leverages individuals who move across neighborhoods ("movers") to identify the effect of exposure to diversity of the Italian community on national identity formation.

We employ the dataset described in Section II.E.2, in which each individual is observed whenever they have a child. Consider individual i, who has a child at times t and t'. A mover is anyone who moves between two hexagons between t and t'.[1] We denote the "diversity shock" i is exposed to upon moving as $D_i \equiv H_{d'(i)} - H_{d(i)}$, which is the difference in diversity between the new and the

[1] We use hexagons rather than districts because we cannot consistently follow districts over time. The diversity of the local community, however, is measured at the district level, as in the rest of the analysis.

old neighborhood. We binarize D_i into a categorical variable, *Diversity Shock*$_i$, equal to one if D_i is in the top 20% of the distribution and zero otherwise, and denote t_i^* as the first year when i moves between two hexagons.

We estimate variations of the following specification:

$$LNI_{idht} = \alpha_i + \alpha_{h\times t} + \sum_{\substack{k=-10 \\ k\neq 0}}^{+10} \beta_k \cdot I(t - t_i^* = k) \cdot \textit{Diversity Shock}_i + \varepsilon_{idht}\,, \tag{A.4}$$

where α_i denote individual fixed effects and α_{ht} denote hexagon-by-time fixed effects. The treatment effect coefficients $\{\beta_k\}$ capture differences in the local name index between movers and non-movers before and after movers change their neighborhood.

Figure F.8 presents the estimates from equation (A.4). Identification relies on a standard parallel trends assumption requiring that, conditional on fixed effects, treated and control individuals would not have experienced diverging LNI trajectories if the former had not moved to a different neighborhood. While this assumption is untestable, we find that pre-move coefficients $\{\beta_k\}_{k=-10}^{-1}$ are never statistically significantly different from zero, indicating that movers and non-movers were not systematically different before the former moved to a different neighborhood.

Over the decade following the move, we estimate that migrants who experienced a "diversity shock" gave their children names that were 10% less local compared to individuals who did not move or moved to neighborhoods with diversity comparable to their origin district. This effect builds over time and is largest 6 to 7 years after migrants move. Quantitatively, exposure to diverse neighborhoods upon moving decreases LNI by approximately 9% of a standard deviation.

In Table E.2, we report estimates from a mean-shift specification in which the event-time indicators are replaced by a categorical term $I(t \geq t_i^*)$ that equals one for all observations after the treatment period and zero otherwise. In the baseline specification, shown in column (1), we include individual, area, and event-time fixed effects. Individual fixed effects imply that identification relies on within-individual changes in LNI before and after treatment. In column (2), we add birth-year fixed effects to control for cohort-specific shocks that may affect trends in name choice and correlate with the local name index. In column (3), we replace area and cohort fixed effects with area-by-cohort fixed effects, absorbing neighborhood-level time-varying shocks such as changes in local labor market conditions or immigrant inflows. Column (4) replaces the binary diversity indicator with a continuous treatment measure and confirms the qualitative results.[2]

In column (5), we add an interaction between the age at which migrants moved and the baseline treatment to test whether the effect of exposure to diversity is age-specific. We do not find evidence in this direction. The estimated difference-in-differences coefficient remains negative, although it is less precisely estimated, while the interaction term between the treatment and age at move is not statistically significant.

Last, in column (6), we replicate the analysis using the distance-adjusted LNI metric, which con-

[2]Because non-movers do not experience a cross-sectional diversity shock, we apply a monotone increasing transformation to the continuous diversity measure (exponential) and impute zero exposure for non-movers.

firms the result obtained from the unweighted LNI indicator. Exposure to more diverse neighborhoods has a positive and statistically significant effect on the distance-adjusted local name index, indicating *less* name particularism, equal to approximately 15% of its standard deviation.

A.VII Conley Standard Errors with High-Dimensional Fixed Effects

As a robustness check, we estimate standard errors that allow for spatial autocorrelation following Conley (1999). Although Conley standard errors are widely used in applied work, existing implementations, particularly in `Stata`, become computationally prohibitive in large samples and in the presence of high-dimensional fixed effects (HDFE, Colella et al., 2023). In our application ($N \approx 300{,}000$), this constraint is binding. We have developed a fast `Python` implementation of OLS and exactly identified 2SLS with multi-way HDFE and Conley-type spatial correlation, using blocked neighbor searches and just-in-time compilation to ensure memory safety and computational feasibility, to tackle this software limitation. The code is available upon request at `davide.coluccia@bristol.ac.uk`.

B LINKED SAMPLE OF ELLIS ISLAND AND POPULATION CENSUS RECORDS

This section provides a detailed discussion of the primary data and the algorithm we developed to link the Ellis Island immigration directories and the federal census records. Then, we assess the linked sample's reliability and representativeness and present key descriptive statistics.

B.I The Ellis Island Immigration Data

The individual-level dataset on Italian immigrants in the United States is from Coluccia and Spadavecchia (2025). The dataset comprises 4,194,728 uniquely identified records of Italians who entered the United States through the Ellis Island immigration station between 1892 and 1924. The data contain information on the name and surname, age, immigration year, ship name, origin municipality, and literacy of each immigrant. We defer the interested reader to the original paper for a detailed methodological discussion of the data.

Origin municipalities are missing in 12% of cases. The remaining 88% are geo-coded using commercial software. Approximately 32% are geo-coded to coordinates outside Italy. These cases conflate reporting errors and migrants returning from temporary stays in Italy who had previously traveled and settled in the United States. Ultimately, our analysis sample thus comprises 2,526,024 individuals who are located at the coordinates of their municipality of origin in Italy.

B.II Motivation and High-Level Description of the Approach

Linking the population census and the Ellis Island individual records shares several challenges faced by standard intergenerational census-linking algorithms (e.g., see Abramitzky et al., 2021). As is common in (historical) record linkage, frequent misspellings and transcription errors do not allow exact matching. Instead, automated algorithms implement "fuzzy" string matching methods that allow the linking of strings with minor discrepancies.

Our task, however, presents additional, more specific obstacles. First, the Ellis Island records were collected by American officers. Hence, Italian names and surnames display frequent phonetic misspellings, translations, and abbreviations. Second, when completing their census enumeration, Italians commonly Americanized—i.e., translated—their names (e.g., see Biavaschi et al., 2017). Both circumstances are problematic for standard fuzzy string-matching algorithms because translations and phonetic errors frequently result in vastly different string records.

We design a linking algorithm to deal with these and related challenges. Our approach leverages recent advances in Natural Language Processing (NLP) and Large Language Models (LLM) to circumvent the algorithmic shortcomings of common string similarity metrics. More specifically, we capture the similarity between translated versions of the same name through a fine-tuned sentence bidirectional encoder representations from transformers (SBERT) model (Reimers and Gurevych, 2019). SBERT models produce string embeddings that encode the contextual representation of tokens. We complement the SBERT similarity metric with a standard Jaro-Winkler string similarity measure to identify spelling and phonetic mistakes.

We use the SBERT-Jaro-Winkler distances and the immigration year to pair each Ellis Island record

with a set of plausible links in the U.S. Census. We label the resulting sample as "*Sample 1*." We propose two methodologies to reduce the prevalence of "multiple matches," i.e., Ellis Island records paired with more than one Census entry. First, in dataset "*Sample 2*," we employ a state-of-the-art LLM to identify the most likely match(es) among those that appear in *Sample 1*. Second, in "*Sample 3*," we take the match(es) that maximize a harmonic average of the SBERT-Jaro-Winkler similarities within each match. Our analysis is based on *Sample 1*, but we show that all our results remain unchanged using the other samples. As is standard in the literature, we weigh multiple matches by the inverse of the number of matches to avoid multiple counting.

B.III Technical Description of the Algorithm Implementation

Figure F.1 provides a visual overview of the algorithm. We now describe each component in greater technical detail.

B.III.1 Construction of the Training Data

The starting point is the individual-level data contained in the Ellis Island ($N \approx 2,500,000$) and the Census ($N \approx 4,400,000$). We extract the universe of first names listed in these records and harness the flexibility of a state-of-the-art LLM (OpenAI's `gpt-4o-mini`) to assign each name to an existing Italian name. This first step produces a set of approximately 40,000 names that appear in the original data and that the LLM pairs with their correct Italian counterpart. We split the sample into two equal-sized sub-samples. The following panel replicates the prompt that the LLM receives to classify name pairs.

`gpt-4o-mini` Prompt for Name Cleaning

You are an expert in reconstructing Italian **names** that were altered by U.S. immigration officials.

Task:

Return ONLY a valid JSON object with the most likely original Italian **first name** and a confidence score (-100 and 200).

Instructions:

- You must return an existing Italian first name.
- If you cannot find an Italian name, return the original and set the confidence to -100.
- If the name is already an Italian **real** first name, return as is and set the confidence to 200.

Rules for the confidence level:

- 200 = Name is a perfect match with no discrepancies.
- 90 = Extremely confident match (name clearly matches, considering translation/misspelling).
- 80 = Likely match but with minor discrepancies (plausible misspelling).
- 70 = Possible but uncertain match; significant discrepancies exist.
- 60 = Unlikely match; substantial doubts remain.

- 50 = Most likely match, but still very uncertain.
- 0 = Uncertain match; no clear match found.
- -100 = No Italian name found; return the original name.

Output Format:

{{

original: <original recorded name>,

inferred italian: <most likely Italian name>,

confidence: <confidence score>

}}

We restrict the attention to names with a confidence score of at least 80 to ensure that the training set comprises high-quality matches.

B.III.2 Training of the NLP Embeddings

Using the first subsample, we fine-tune a multilingual pre-trained SBERT model from the `huggingface` repository (`paraphrase-multilingual-mpnet-base-v2`). The training data contains both translations as well as minor spelling and phonetic mistakes. The embeddings produced by the fine-tuned SBERT model thus replicate the high contextual similarity between recorded names that map onto the same Italian name.

B.III.3 Filtering Classification Rule

Consider two first names, a and b. Name a appears in the Ellis Island data, and b is recorded in the Census. Let $\mathbf{e}_a$ and $\mathbf{e}_b$ denote the embeddings produced by the fine-tuned SBERT model. We say that b is a candidate match for a if the cosine similarity between the two embeddings is above a threshold $\tau_1 \in [0, 1]$:

$$\sigma_{ab}^{SBERT} \equiv \frac{\mathbf{e}_a^\intercal \mathbf{e}_b}{||\mathbf{e}_a|| \cdot ||\mathbf{e}_b||}, \tag{B.1}$$

or if the Jaro-Winkler similarity between a and b is above a certain threshold $\tau_2 \in [0, 1]$. The Jaro-Winkler similarity is defined as one minus the Jaro-Winkler distance. Formally, it is

$$\sigma_{ab}^{JW} \equiv 1 - \delta_{ab}^{JW} = 1 - \begin{cases} 0 & \text{if } m_{ab} = 0 \\ \frac{1}{3}\left(\frac{m_{ab}}{||a||} + \frac{m_{ab}}{||b||} + \frac{m_{ab} - t_{ab}}{m_{ab}}\right) & \text{else} \end{cases}, \tag{B.2}$$

where m_{ab} is the number of matching characters in the two names, and t_{ab} is the number of transpositions. The number of transpositions is defined as a number of matching characters that are not in the right order divided by two. The Jaro-Winkler similarity is a widely adopted edit metric in record linkage algorithms because it effectively captures minor typographical errors, transpositions, and abbreviations in short strings, and its prefix-weighted structure prioritizes matches that share the same initial characters, which are typically more stable in historical records. The score is normalized such that zero means no similarity, and one denotes exact matches.

We employ the SBERT and Jaro-Winkler disjoint similarities because they capture different imperfect linking instances. A low cosine similarity between two SBERT embeddings reveals that the two names are likely translations of a single name. Hence, they should be considered as candidate matches. A low Jaro-Winkler similarity between two names indicates that the two names are imperfect transcriptions of the same name and, as such, should be considered candidate matches. Since these are different cases, we do not take an average "SBERT-Jaro-Winkler" similarity. Instead, we populate the set of possible matches as those b's that have *either* a "high" SBERT-cosine *or* a "high" Jaro-Winkler similarity with a

B.III.4 Estimation of the Classification Hyperparameters

To estimate the thresholds τ_1 and τ_2, we employ the remaining half of the names matched to their proper Italian counterpart. Specifically, we split the 20,000 samples into a training set composed of 14,000 pairs and a test set of 6,000 names, over which we evaluate the performance of the composite distance classifier. We augment the train and test sets with a random sample of false matches drawn to reflect the frequency distribution of names in the population. Let $\sim$ denote an equivalence class whereby $a \sim b$ implies that names a and b constitute a match. Given the classification rule

$$a \sim b \;\Leftrightarrow\; \sigma_{ab}^{SBERT} \geq \tau_1 \;\vee\; \sigma_{ab}^{JW} \geq \tau_2 \tag{B.3}$$

we search for the values $\hat{\tau}_1$ and $\hat{\tau}_2$ that maximize the $F1$-score over the training set.[3]. The $F1$-score is defined as the harmonic average between the precision and recall of the classifier. The precision is defined as the number of true positive results divided by the number of all samples predicted to be positive, including those not identified correctly. The recall is the number of true positive results divided by the number of all samples that should have been identified as positive. The $F1$-score is a standard measure to target when fine-tuning classifiers, as it balances the tradeoff between accuracy and coverage.

In our application, the optimization routine selects $\hat{\tau}_1 = .65$ and $\hat{\tau}_2 = .95$. We provide additional details on the performance of the fine-tuned classifier in the next section.

B.III.5 Construction of Sample 1

The first sample comprises all matches obtained using the standard classification rule that satisfies the optimal thresholds. Formally, let B denote the universe of records in the Census. For each Ellis Island record a, the set of matches in *Sample 1* is denoted by

$$m_a^1 = \left\{ b \in B \;\text{ such that }\; |t_b - t_a| \leq 1 \;\wedge\; a \sim b \;\wedge\; \tilde{\sigma}_{ab}^{JW} \geq \tau_3 \right\}, \tag{B.4}$$

where t_b and t_a denote the immigration year recorded in the census and in the Ellis Island registers, and the classification rule $\sim$ is computed using the couple optimal thresholds $(\hat{\tau}_1, \hat{\tau}_2)$ estimated in the previous step. The term $\tilde{\sigma}_{ab}^{JW}$ denotes the *surname* Jaro-Winkler similarity between records a and b. Since census enumerators do not translate surnames, it is not necessary to apply NLP tools on

[3]Specifically, we run a grid search over the $[0,1]^2$ space for τ_1 and τ_2 for evenly spaced values of the thresholds using a 0.05 pass.

surnames because the Jaro-Winkler similarity alone ensures that we include misspellings in the set of potential matches. We select $\tau_3 = \hat{\tau}_2$ to ensure high-quality surname matches.

B.III.6 Construction of Sample 2

In *Sample 2*, we refine the set of matches obtained in the previous step using a state-of-the-art LLM (OpenAI's `gpt-4o-mini`). We run the following prompt using the API.

`gpt-4o-mini` Prompt for Match Disambiguation

You're an expert historian matching Italian immigrant records to U.S. census data.

Original immigrant details:

- Name: `{original_name}`

Census candidates:

- Candidates: `{candidates}`

Task

Carefully assess if any candidate matches the immigrant, considering:

- Italian-to-English translations (e.g., Giovanni ↔ John, Giuseppe ↔ Joseph).
- Phonetic errors, abbreviations, and misspellings from historical American records.

Prefer translations over phonetic matches.

Provide your best candidate selection clearly reflecting your certainty.

If there is one match that is clearly the best, only return that match.

If there is no clear match, return the most likely match(es) and set confidence to 6.

Rules for assigning confidence:

- 1 = Name is a perfect match with no discrepancies.
- 2 = Extremely confident match (name clearly matches, considering translation/misspelling).
- 3 = Likely match but with minor discrepancies (plausible misspelling).
- 4 = Possible but uncertain match; significant discrepancies exist.
- 5 = Unlikely match; substantial doubts remain.
- 6 = Most likely match, but still very uncertain.

Output Format:

```
{{
best matches: <list of candidate names>,
confidence: <confidence score>
}}
```

In the prompt, the set of candidate names that `gpt-4o-mini` is asked to disambiguate is given by, for the general record a, the set m_a^1. Thus, if we let $\succsim$ denote the operator that maps potential matches into GPT-validated matches. The set of matches in *Sample 2* is thus given by

$$m_a^2 = \left\{ b \in m_a^1 \text{ such that } b \succsim b' \ \forall \ b' \in m_a^1 \right\} \subseteq m_a^1, \tag{B.5}$$

which simply implies that, in m_a^2, all matches must be within m_a^1, so, in this sense, *Sample 2* is a refinement of *Sample 1*, and they must be valid according to GPT.

B.III.7 Construction of Sample 3

The set *Sample 3* is an alternative refinement of *Sample 1*. Let $\hat{\sigma}_{ab}$ be the similarity metric that determines the name-level match. Thus, if $\sigma_{ab}^{SBERT} \geq \hat{\tau}_1$ and $\sigma_{ab}^{JW} < \hat{\tau}_2$, it shall be $\hat{\sigma}_{ab} = \sigma_{ab}^{SBERT}$, and the reverse if $\sigma_{ab}^{JW} \geq \hat{\tau}_2$ and $\sigma_{ab}^{SBERT} < \hat{\tau}_1$. In the residual case where both similarities are above the threshold, we pick $\hat{\sigma}_{ab} = \sigma_{ab}^{JW}$. Then, the set of matches included in *Sample 3* for the generic record a is

$$m_a^3 = \left\{ b \in m_a^1 \text{ such that } \alpha \cdot \hat{\sigma}_{ab} + (1-\alpha) \cdot \tilde{\sigma}_{ab} \geq \alpha \cdot \hat{\sigma}_{ab'} + (1-\alpha) \cdot \tilde{\sigma}_{ab'} \ \forall \ b' \in m_a^1 \right\} \subseteq m_a^1, \tag{B.6}$$

where $\alpha \in (0,1)$ is a weight that determines the importance assigned to the name relative to the surname. In this case, the refinement relative to *Sample 1* is given by taking the most similar match(es) in the set of plausible candidate matches m_a^1, according to a weighted average of the name and surname similarities $\hat{\sigma}_{ab}$ and $\tilde{\sigma}_{ab}$.

B.IV Internal Validation and Representativeness of the Linked Sample

This section complements Section II.A.

When constructing candidate matches for each Ellis Island record, we restrict the census population to individuals born within a $\pm$2-year window around the reported birth year and who immigrated within a $\pm$5-year window around the reported immigration year. Such buffers are standard in historical record linkage as they reflect reporting errors in birth—and plausibly immigration-years, partly due to low literacy and recall issues (e.g., Abramitzky et al., 2021).

Figure F.2 shows the difference in birth year (Figure F.2a) and immigration year (Figure F.2b) between linked census and Ellis Island records. Three patterns emerge. First, in both dimensions, the modal match occurs at zero gap, indicating that a relative majority of links involve identical reported years. Second, the distributions are symmetric around zero, ruling out systematic biases toward matching individuals reported as born or immigrated earlier or later in the census. Third, the linking probability declines with the absolute gap between reported years. This pattern is mechanically expected in a well-behaved linkage procedure: if matches were driven by noise or by the blocking windows alone, the distribution of gaps would be approximately uniform within the admissible range, or even exhibit excess mass in the tails. The observed monotonic decline, therefore, indicates that the linkage algorithm assigns greater weight to records that agree more closely on blocking variables, and that spurious matches at the boundaries of the admissible windows are relatively rare.

An additional indicator of match quality is the similarity between names and surnames in the Ellis Island and census records. For each accepted link, the name similarity must exceed either 0.95

under the Jaro–Winkler metric or 0.6 under SBERT. Figure F.3a reports the distribution of Jaro-Winkler similarities, whereas Figure F.3b reports the distribution of SBERT similarities. There are very few matches with SBERT similarity below 0.6, while approximately 30% of the sample has Jaro-Winkler similarity below 0.95. This pattern is expected. The SBERT similarity is designed to capture semantic equivalence between name translations, which often appear orthographically distant and therefore score poorly under Jaro-Winkler. By contrast, high Jaro-Winkler matches generally also exhibit high SBERT similarity, indicating that the two distances are mutually consistent in identifying plausible matches. Importantly, a substantial share of the linked sample (approximately 40%) is constituted by nearly-perfect matches, i.e., matches with a name-similarity above 0.95. This pattern suggests that a large fraction of links is by near-exact agreement on names, rather than relying on the more permissive translation-aware metric.

Figure F.3c displays the distribution of surname similarities. By construction, accepted matches must have a Jaro-Winkler surname similarity above 0.95. We do not apply the SBERT correction to surnames because it was extremely rare for Italians to translate their surname (Carnevale, 2010). Interestingly, the majority of surnames ($\approx 60\%$) exhibit a near-perfect surname match, i.e., a surname similarity of one. This may reflect the fact that surnames were less likely than given names to be altered or Anglicized upon arrival, and that census enumerators tended to record them more consistently across sources.

B.V External Validation of the Linked Sample

In Figure F.4, we compare the baseline linked migrant sample (Sample 1 in Appendix Section B.III) with two refinements. In Sample 2, we use GPT 4 to validate the matches produced by the baseline algorithm and retain only those deemed most plausible. In Sample 3, for each Ellis Island record with at least one accepted link, we retain only the match with the highest combined name and surname similarity. Both samples are refinements in the sense that they impose additional restrictions on Sample 1, and therefore are weakly smaller by construction.

In Figure F.4a, we plot the matching rate of Sample 2 (in red) and 3 (in blue) over time. The two series co-move and are qualitatively very similar to the pattern displayed in Figure Ia for Sample 1. The average matching rate of the baseline sample (18.762%) is also similar to Sample 2 (16.142%), indicating that GPT 4 deems 90% of the matches in Sample 1 as extremely plausible.

In Figure F.4b, we report the distribution of the number of matches in Sample 2 (red-contoured bars) and 3 (gray lines). In both cases, as in Sample 1 (Figure Ib), single matches account for the vast majority of the accepted links in the sample.

Finally, in Figure F.4c, we report the correlation between Sample 1 and Samples 2 (in red) and 3 (in blue). To construct this figure, we aggregate the matches at the municipality level; hence, each dot displays the municipality-level share of Ellis Island migrants from each municipality that are paired with at least one valid census record in Sample 1 (on the x-axis) and in Samples 2 and 3 (on the y-axis). By construction, both scatter plots lie below the 45-degree line because Samples 2 and 3 are refinements of Sample 1. In both cases, the correlation between the matching rates is large (0.87

for Samples 1–2 and 0.98 for Samples 1–3) and highly statistically significant, indicating that both refinements rescale the matching rate without modifying its spatial coverage.

Validating the linked dataset using external sources at the individual level is not feasible, as this is the first dataset to recover individual-level information on Italian immigrants' municipalities of origin. More aggregated bilateral data, such as municipality-to-county migration flows, are also unavailable. In Figure F.5, we compare out-migration flows constructed from the Ellis Island records with official figures compiled by the Italian Commissioner General for Emigration and published in the *Annuario Statistico della Emigrazione Italiana dal 1876 al 1925*. As described in Coluccia and Spadavecchia (2025), these statistics report the annual number of emigrants to the United States by region, based on passports issued for transatlantic travel.

Figure F.5a (resp. Figure F.5c) reports the number of emigrants by year compiled using data from the *Annuario* (solid red line, on the left y-axis) and the Ellis Island full (resp. linked) dataset (dashed blue line, right y-axis). The Ellis Island dataset closely replicates the outflow constructed from Official Statistics data throughout the sample period. The number of emigrants in the Ellis Island dataset is lower than in the official statistics because the municipality of origin contains missing or invalid values, but, as explained by Coluccia and Spadavecchia (2025), this is unlikely to generate systematic selection into the dataset.

In Figure F.5b, we report the correlation between the full Ellis Island dataset and data from the *Annuario*. To construct this figure, we collapsed the full Ellis Island dataset, i.e., irrespective of whether migrants are linked to the census, at the region-year level. Hence, each dot reports the number of U.S. emigrants in each year between 1890 and 1924 according to the official statistics (on the x-axis) and the Ellis Island dataset (on the y-axis). We find a positive and highly statistically significant correlation between the two datasets, as evidenced by the regression coefficient reported separately. In all but one case (Calabria, 1902), the difference between the Ellis Island and the official statistics is small relative to the number of emigrants

Figure F.5d replicates the exercise using the linked Ellis Island dataset. Since, as we documented in Figure Id, the linked sample preserved the spatial distribution of the full Ellis Island dataset, we expect to recover a positive correlation between the linked dataset and official statistics. The figure confirms this pattern: the regression coefficient remains highly statistically significant (though scaled by a factor approximately equal to the matching rate).

Together with the validation results documented in Section II.A, these exercises indicate that (i) the Ellis Island dataset replicates less granular data compiled from official sources, (ii) the linked Ellis Island dataset preserves the representativeness of the sample, and (iii) the reliability of the linking algorithm is confirmed by multiple internal validation refinements.

C ADDITIONAL HISTORICAL BACKGROUND

This appendix provides additional historical detail supporting the discussion in Section I. It elaborates on three aspects only summarized in the main text: (i) the geographic and social organization of Italian migrant communities in the United States; (ii) the depth of linguistic and regional fragmentation within Italy at the time of mass migration; and (iii) specific mechanisms through which cultural convergence emerged within immigrant communities and fed back to Italy.

C.I Geographic Concentration and Localism in Migrant Communities

Italian migration to the United States was shaped by strong local networks that generated highly localized settlement patterns. Migrants from specific towns or provinces frequently clustered together in destination cities, often following earlier migrants from the same locality (Vecoli, 1983). These patterns of chain migration produced sharp geographic concentration across and within U.S. cities.

Numerous examples illustrate this localization. Stonecutters from the marble-producing area of Massa settled in Barre, Vermont; migrants from Molfetta concentrated in Hoboken, New Jersey; and emigrants from Floridia, Sicily, in Hartford, Connecticut (Corradini, 1923). Residential concentration persisted even within neighborhoods. As Luigi Villari (1912, p. 216), the Italian consul in New York, observed,

> *"[…] we can find only Sicilians in a street, only people from Calabria in another street, and immigrants from Abruzzi in a third one. There are even streets where only individuals from a single town live: a colony from Sciacca here, a colony from San Giovanni in Fiore there, a colony from Cosenza somewhere else."*

In East Harlem, New York City, even though most Italian migrants originated from the South, segregation by place of origin remained widespread. Neapolitans clustered between 106th and 108th Streets, while migrants from the nearby region of Basilicata settled primarily between 108th and 115th Streets (Orsi, 1985). In Chicago, segregation was similarly stark, as Italians settled across sixteen distinct neighborhoods according to their origins (Vecoli, 1983). Such patterns of segregation within Italian communities reflected the deep fragmentation that characterized Italy itself. Over time, however, these localized identities gradually gave way to broader forms of shared identity, as Italian communities coalesced into a more unified "Italian American" group.

These patterns mirrored the strong localism that characterized Italy itself and initially reproduced regional divisions within migrant communities. Interaction across groups nevertheless increased over time as migrants shared workplaces, institutions, and public spaces within American cities.

C.II Cultural Fragmentation in Italy

C.II.1 Linguistic Fragmentation

At the time of unification, standard Italian was spoken by only a small minority of the population, estimated between 2.5 and 6.5 percent (De Mauro, 1963). Fewer than one percent of individuals outside Tuscany and Rome spoke Italian as a first language. Italian functioned largely as a written language

accessible to literate elites, and literacy rates were low nationwide (Ciccarelli and Weisdorf, 2019).[4] Four decades later, a public inquiry reported that roughly 50 percent of the population had some knowledge of Italian, yet only 20 percent could use it for everyday communication (Corradini, 1910). The *legge Casati* (1861) introduced compulsory elementary schooling up to age eight (increased to nine with the *legge Coppino*, in 1877) in an effort to promote the "national language," but implementation remained weak. In 1870, between 40 and 65 percent of school-age children did not attend school, and as late as 1906, half of children aged six to ten had received no formal schooling.

Despite state efforts at linguistic homogenization, local dialects remained widespread, even among educated elites. King Vittorio Emanuele II, for instance, conferred with his ministers in his native Piedmontese dialect, while Camillo Benso, Count of Cavour, the architect of unification, preferred French to Italian and thought Sicilians spoke Arabic (Chabod, 1965; Smith, 1985). Dialect use was common even in schools, where, in 1910, 50 percent of teachers reportedly conducted lessons in dialect rather than in Italian, despite formal requirements to the contrary (Corradini, 1910). On the eve of World War I, linguistic fragmentation therefore persisted, particularly in rural areas, from which the majority of Italian migrants to the United States originated.

C.II.2 Geography

Geography reinforced fragmentation. The Apennine mountain chain limited interaction across regions, and travel across the peninsula remained slow prior to the expansion of the railway network (Gilmour, 2011). Economic divergence between the Center–North and the South further deepened regional cleavages. Quantitative evidence indicates that the ratio of Southern to Northern GDP per capita declined markedly between unification and the interwar period (Felice, 2019). These linguistic, economic, and geographic barriers shaped the identities migrants carried abroad.

Economic divergence reinforced cultural distance between the two areas, as illustrated by the civil conflict that followed unification. This conflict, fought largely by the Southern peasantry and commonly referred to as *brigantaggio*, at its peak required the mobilization of two-fifths of the regular army (Molfese, 1964; Dickie, 1999). Linguistic and educational disparities further deepened these divisions: school attendance remained low in the South, where the use of local dialects persisted well into the postwar period (De Mauro, 1963).

C.III Cultural Convergence in Immigrant Communities

Historians of the Italian diaspora emphasize that Italy's cultural fragmentation produced highly localized immigrant communities in the United States (Gabaccia, 2013). Over time, these parochial identities blended into an increasingly homogeneous "Italian American" identity, particularly in response to World War I and the rise of Fascism (Vecoli, 1985; Luconi, 2003). Reflecting this shift, in 1934, the leading Italian-American newspaper, *Il Progresso Italo-Americano*, wrote:

[4] Fewer than 1 percent of the population outside Tuscany and Rome spoke "Italian" as a first language. Italian functioned largely as a written language, accessible primarily to literate individuals. In 1861, the literacy rate averaged around 20 percent nationwide (and was closer to 10 percent in the South), placing a strict upper bound on the share of Italian speakers (Ciccarelli and Weisdorf, 2019).

> *"[. . .] there cannot be differences between a Lombard and a Roman, between a Venetian and a Sicilian, between a Piedmontese and a Neapolitan. The increasing difficulties in our lives require solidarity, not divisions."*

To underscore the importance of migrant group identity and its influence on their country of origin, Choate (2008) refers to Italy as an "emigrant nation."

C.III.1 Associational Life

Fragmentation and subsequent homogenization are evident in the evolution of mutual aid societies, grassroots associations that provided financial and social support to immigrants. Initially, these associations were organized along narrow local and regional lines: only migrants from Abruzzo, for instance, could join the Fratellanza Abruzzese in Providence, while membership in the Ateleta Beneficial Association in Pittsburgh was restricted to migrants from the small village of Ateleta (Luconi, 2003). In East Harlem alone, Covello (1967) counted as many as sixty-four mutual aid societies with origin-based admission criteria. Beginning in the 1910s, however, these organizations increasingly federated across local origins (Gabaccia, 2013). In Milwaukee, for example, all associations merged into a single "Italian Societies of Milwaukee" confederation. The first national Italian association, the "Order of the Sons of Italy," was established in 1905, and, in the 1920s, its membership increased to 300,000 (Vecoli, 1985).

C.III.2 Linguistic Convergence

Language barriers among Italians from different parts of the country were substantial. In 1888, for example, the Irish archbishop of New York sought Italian clergy to minister to newly arrived immigrants, but the priests available to him, largely from Northern Italy, were unable to communicate with Southern Italians (Choate, 2008). As noted above, most Italian immigrants did not speak standard Italian upon arrival and instead developed a syncretic creole that blended regional dialect—predominantly Neapolitan—with English (Carnevale, 2010). This Italian-American creole became the primary medium of communication across regional lines, and its cultural significance was such that it survived even among returnees in Italy.

Linguistic homogenization was further reinforced by the Italian-American press, which predominantly published in standard Italian (Carnevale, 2010). Between 1900 and 1930, the Italian-language daily press ranked second only to German-American newspapers. As Deschamps (1996) argues, these newspapers promoted the use of Italian among migrants more than they did in Italy.[5]

C.III.3 Religious Practices

Although virtually all Italian immigrants were Catholic, religious practices varied markedly across regions, and devotion to local saints remained strong. In St. Louis, for instance, Sicilians boycotted the

[5]Newspapers were more effective in spreading Italian among migrants than among stayers primarily because migrants gained unprecedented access to basic education upon arriving in the U.S., thanks to public and denominational (Catholic) schools.

Church of St. Ambrose, which had been erected by Lombards to honor a saint traditionally associated with Milan (Luconi, 2003).

Over time, however, the Catholic Church emerged as a force of homogenization. The Missionaries of St. Charles Borromeo, commonly known as the "Scalabrinians," were founded to minister to Italian migrants abroad and explicitly sought to preserve and cultivate both Catholic practice and a shared Italian cultural identity (Choate, 2008).[6]

Convergence in religious practices also emerged endogenously within increasingly diverse immigrant communities. In his ethnography of East Harlem, Orsi (1985) documents that Italian immigrants initially venerated a wide array of local saints and patrons brought from their communities of origin. As boundaries between regional groups blurred, however, few survived. Among them, the Madonna of Mount Carmel was the most relevant one. Originally associated with immigrants from parts of Campania and Sicily, devotion to the Madonna gradually became a shared religious experience among community members, irrespective of their origins. This convergence did not result from clerical intervention but instead arose through repeated collective practice in the common urban space, especially during the annual *festa*.

C.III.4 Cuisine

A final dimension of cultural integration concerns cuisine (Chiaricati, 2020; Soffiati and Grandi, 2024). Although Italian cuisine is a defining component of Italian culture, recent work shows that many contemporary "Italian" foods emerged within immigrant communities in the United States. This process involved two steps: the diffusion of regional foods across immigrant groups and the invention of new culinary practices.

Pasta and pizza exemplify the first channel. Originally concentrated in Naples, where it was eaten by hand, pasta spread widely among Italian immigrants to the point that Fascist authorities considered it an American invention. Until the early 1900s, pizza remained a similarly marginal food confined to the slums of Naples and to Neapolitan emigrants. Its modern form emerged in the U.S., where immigrants introduced toppings and commercialized it through specialized shops. When American troops landed in Sicily in 1943, they were surprised by the absence of pizza shops.

The second channel involved invention. Tomato sauce entered Italian cookbooks only in the 1930s after becoming common among emigrants, while dishes such as meatballs and spaghetti Alfredo remain distinctive of Italian-American cuisine.

C.IV Return Migration and Feedback Effects to Italy

Did cultural convergence among Italian immigrants in the United States influence Italy? Qualitative historical evidence suggests that it did (Choate, 2008). Partly because of high return migration rates,

[6] As their founder, Giovanni Battista Scalabrini, put it, *"Religion and fatherland [...] become intertwined. [...] May Italy, sincerely reconciled with the Apostolic See, emulate its ancient glories and add a new, undying glory, setting even its faraway children on the shining paths of true civilization and progress"* (Scalabrini, 1983, p. 145–161). In doing so, Scalabrini explicitly advocated against the "melting pot" and immigrant assimilation: *"Not many years ago, in the United States, there were enormous efforts to Americanize [...] Let us also work so that all Italians abroad possess the same motto, the same steadfastness, the same courage: for Religion and for the Fatherland"* (Scalabrini, 1898, p. 39).

values and practices acquired abroad plausibly diffused back to migrants' communities of origin. Return migrants founded mutual aid associations that transcended village boundaries and often evolved into political organizations. In the Apennine village of Montefegatesi, in 1902, returnees financed the construction of a monument to Giuseppe Garibaldi, a symbol of national unity (Wyman, 1993). Cultural feedback also operated through consumption practices: dishes developed within Italian-American communities later entered Italian culinary traditions, although their diasporic origins were frequently obscured (Chiaricati, 2020). More generally, qualitative studies indicate that Italy's identity as an "emigrant nation" developed partly through interactions among local identities in emigrant communities and the influence that return migrants exerted over their origin communities.

D A MODEL OF IMMIGRANT GROUP IDENTITY CHOICE

We develop a parsimonious framework to rationalize how exposure to within-group diversity shapes identity formation. We derive two implications which we test in the empirical analysis.

A continuum of individuals resides in a local community. Each individual originates from a municipality m. The community is characterized by a measure of within-Italian diversity $D \in [0,1]$, which increases with the heterogeneity of origin municipalities.

Agents choose one identity to express, $j \in \{L, N, A\}$, where L, N, and A respectively denote local, Italian, and American identity. Identity choice affects cultural investment.

Following Akerlof and Kranton (2000), identity yields utility if it is socially recognized. The payoff from choosing identity j is $U_j = v_j \cdot s_j(D) - c_j$, where $v_j > 0$ captures the intrinsic utility of identity j, $s_j(D) \in [0,1]$ denotes the share of individuals in the community for whom identity j is meaningful, and $c_j \geq 0$ is the cost of adopting identity j.

For simplicity, we let $s_L(D) = 1 - D$, $s_N(D) = D$, and $s_A(D) = \bar{s}_A$.[7] The interpretation is that local identities are recognized only in homogeneous communities, while national identity becomes a useful coordination device as local origins fragment. Exposure to natives is orthogonal to within-Italian diversity, so the salience of American identity is not affected by within-Italian diversity D. This structure is analogous to coordination-based models in which the payoff to identity depends on group size (e.g., Shayo, 2009).

We assume costs satisfy $0 = c_L < c_N < c_A$, reflecting that local identity is inherited, national identity requires modest adaptation, and American identity entails higher linguistic and social effort.

Individuals choose the identity that maximizes their utility $\{U_j\}_{j \in \{L,N,A\}}$. National identity is chosen over local identity whenever $v_N D - c_N \geq v_L(1 - D)$, which defines a threshold level of diversity above which national identity strictly dominates local identity. This yields the following implication.

Proposition D.1. *(Local-to-National Identity) Higher within-Italian diversity reduces the prevalence of local identities and increases the prevalence of national Italian identity.*

National identity is chosen over American identity whenever $v_N D - c_N \geq v_A \bar{s}_A - c_A$. Because the left-hand side increases in D while the right-hand side does not, greater within-Italian diversity weakens incentives to adopt American identity, yielding the following testable implication.

Proposition D.2. *(Americanization) Higher within-Italian diversity decreases the adoption of American identity.*

The framework highlights identity as a coordination device whose value depends on community composition. High within-group diversity reduces the payoff to narrow local identities while increasing the appeal of broader national identities. Crucially, diversity reallocates identity *within* the immigrant group rather than toward the host society, generating national identity formation.

[7] Results remain unchanged for functional forms of $s_j(\cdot)$ that are increasing and monotone in D.

E ADDITIONAL TABLES

Table E.1. Comparison Between Linked and Non-Linked Migrants

	Linked	Non-Linked	Difference	
	(1)	(2)	(3)	(4)
Panel A. Comparison with Ellis Island Records				
Age at Migration	22.810	26.225	-3.416	(0.480)
Latitude of Origin Municipality	40.727	41.011	-0.284	(0.075)
Longitude of Origin Municipality	13.837	13.676	0.162	(0.046)
Area of Origin Municipality:				
North-West	0.082	0.090	-0.008	(0.004)
North-East	0.055	0.070	-0.015	(0.002)
Center	0.109	0.130	-0.021	(0.004)
South	0.484	0.467	0.018	(0.013)
Islands	0.270	0.243	0.027	(0.015)
Panel B. Comparison with Census Records				
Age at Migration	22.629	22.387	0.246	(0.351)
Is Male	0.801	0.596	0.205	(0.020)
Is Employed	0.513	0.449	0.063	(0.013)
log(Occupational Income)	3.184	3.157	0.026	(0.004)
Industry:				
Durable Manufacturing	0.119	0.115	0.004	(0.003)
Nondurable Manufacturing	0.179	0.161	0.018	(0.003)
Trade	0.184	0.169	0.014	(0.003)
Transportation	0.054	0.054	0.000	(0.002)
Agriculture	0.076	0.089	-0.013	(0.003)
Occupation:				
Professional	0.018	0.017	0.002	(0.001)
Managerial	0.073	0.065	0.008	(0.003)
Clerical	0.012	0.013	-0.001	(0.001)
Sales	0.034	0.035	-0.001	(0.001)
Skilled Manufacturing	0.281	0.248	0.033	(0.003)
Unskilled Manufacturing	0.372	0.398	-0.026	(0.006)
Census Region of Residence:				
North-East	0.593	0.585	0.008	(0.008)
Midwest	0.026	0.024	0.002	(0.002)
South	0.014	0.022	-0.008	(0.002)
West	0.061	0.057	0.004	(0.003)

Notes. This table compares migrants in the linked sample with those without a valid link. The unit of observation is an individual, observed once. Columns (1) and (2) report mean characteristics for linked and non-linked individuals. In Panel A, the non-linked group consists of all non-linked records in the Ellis Island dataset; in Panel B, it consists of all non-linked census records of Italian immigrants. Characteristics in Panel A are observed in the Ellis Island data, while those in Panel B are observed in the population census. Columns (3)–(4) report differences between linked and non-linked migrants, with standard errors in parentheses. Referenced on page: 10.

Table E.2. Movers Design and National Identity Formation

	Local Name Index (LNI)					Distance LNI
	(1)	(2)	(3)	(4)	(5)	(6)
Diversity Shock × Post-Move	-0.336**	-0.342**	-0.351**		-1.619*	0.076***
	(0.137)	(0.135)	(0.144)		(0.867)	(0.023)
Diversity Shock (Cont.) × Post-Move				-16.384***		
				(5.035)		
Age					-0.002	
					(0.013)	
Diversity Shock × Post-Move × Age					0.035	
					(0.023)	
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Hexagon FE	Yes	Yes	–	–	–	–
Time Since Move FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	–	–	–	–
Hexagon-Cohort FE	No	No	Yes	Yes	Yes	Yes
# Individuals	114,052	114,051	112,155	112,155	112,155	112,155
# Observations	410,343	410,340	402,640	402,640	402,640	402,640
Mean Dep. Var.	3.909	3.909	3.902	3.902	3.902	9.603

Notes. This table reports results from the movers design specification in (A.4). The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. The dependent variables are the local name index (columns 1–5) and the distance-adjusted local name index (column 6) defined in (1)–(2). The baseline treatment is the interaction between an indicator equal to one after an individual moves to a different hexagon (and zero otherwise) and an indicator equal to one if the difference in diversity between the Italian communities in the destination and origin districts lies in the top 20% of the distribution (and zero otherwise). Column (4) replaces this indicator with the continuous diversity difference measure. Column (5) additionally includes age and its interaction with the baseline treatment. All specifications include individual and event-time (time since move) fixed effects. Column (2) also includes cohort (year) and hexagon fixed effects, while columns (3–6) replace these with hexagon-by-cohort fixed effects. Standard errors, clustered at the district level, are reported in parentheses. Referenced on pages: 19, A8.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.3. Exposure to Immigrant Heterogeneity and National Identity Formation, Excluding Firstborn Children

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Local Name Index						
Italian Diversity	-0.240***	-0.213***	-0.164***	-0.224***	-0.092	-0.157**
	(0.054)	(0.052)	(0.052)	(0.053)	(0.060)	(0.072)
Mean Dep. Var.	10.697	10.697	10.710	10.698	10.697	10.697
Panel B: Distance-Adjusted Local Name Index						
Italian Diversity	0.528***	0.544***	0.614***	0.510***	0.464***	0.350***
	(0.090)	(0.085)	(0.080)	(0.087)	(0.096)	(0.096)
Mean Dep. Var.	24.163	24.167	24.206	24.167	24.167	24.167
Panel C: Foreign Name Index						
Italian Diversity	0.065***	0.066***	0.073***	0.060***	0.055***	0.043***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
Mean Dep. Var.	1.966	1.966	1.970	1.966	1.966	1.966
Municipality FE	Yes	Yes	–	Yes	Yes	Yes
County FE	Yes	–	–	–	–	–
Year FE	Yes	Yes	–	Yes	Yes	Yes
Hexagon FE	No	Yes	–	Yes	Yes	Yes
Municipality-Year FE	No	No	Yes	No	No	No
Hexagon-Year FE	No	No	Yes	No	No	No
Individual Controls	No	No	No	Yes	No	No
Demographic Controls	No	No	No	No	Yes	Yes
Economic Controls	No	No	No	No	No	Yes
# Individuals	207,673	207,499	205,832	207,484	207,499	207,499
# Observations	308,520	308,196	304,480	308,173	308,196	308,196

Notes. This table reports the effect of exposure to diversity within the Italian community on name-based indices from specification (4). Unlike in the main analysis, we exclude firstborn children when computing parents' name-based indices. Panel A uses the local name index defined in (1) as the dependent variable; Panel B uses the distance-adjusted local name index defined in (2); Panel C uses the foreign name index of Abramitzky et al. (2020). The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. All specifications include municipality and census-year fixed effects. Column (1) includes county fixed effects; columns (2) and (4–6) replace county with more granular hexagon fixed effects. Column (3) replaces municipality fixed effects with municipality-year fixed effects and hexagon fixed effects with hexagon-year fixed effects. Demographic controls (included in column 5) are measured at the district level and include population, the share of women, African Americans, immigrants, and Italian immigrants. Economic controls (included in column 6) are measured at the district level and include employment shares by industry. Standard errors, clustered at the district level, are reported in parentheses. Referenced on page: 23.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.4. Exposure to Immigrant Heterogeneity and National Identity Formation, Fractionalization Index

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Local Name Index						
Italian Diversity (Fractionalization)	-0.549***	-0.523***	-0.474***	-0.495***	-0.093*	-0.108*
	(0.071)	(0.078)	(0.085)	(0.073)	(0.055)	(0.061)
Mean Dep. Var.	3.990	3.989	3.989	3.989	3.989	3.989
Panel B: Distance-Adjusted Local Name Index						
Italian Diversity (Fractionalization)	0.073***	0.072***	0.064***	0.065***	0.024**	0.021*
	(0.011)	(0.012)	(0.013)	(0.011)	(0.010)	(0.011)
Mean Dep. Var.	9.593	9.593	9.593	9.593	9.593	9.593
Panel C: Foreign Name Index						
Italian Diversity (Fractionalization)	0.080***	0.078***	0.078***	0.068***	0.043***	0.038***
	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Mean Dep. Var.	0.758	0.758	0.758	0.758	0.758	0.758
Municipality FE	Yes	Yes	–	Yes	Yes	Yes
County FE	Yes	–	–	–	–	–
Year FE	Yes	Yes	–	Yes	Yes	Yes
Hexagon FE	No	Yes	–	Yes	Yes	Yes
Municipality-Year FE	No	No	Yes	No	No	No
Hexagon-Year FE	No	No	Yes	No	No	No
Individual Controls	No	No	No	Yes	No	No
Demographic Controls	No	No	No	No	Yes	Yes
Economic Controls	No	No	No	No	No	Yes
# Individuals	223,721	223,531	221,913	223,521	223,531	223,531
# Observations	334,668	334,303	330,615	334,283	334,303	334,303

Notes. This table reports the effect of exposure to diversity within the Italian community on name-based indices from specification (4). Unlike in the main analysis, we use the fractionalization index (= 1− Hirschman-Herfindahl Index) to measure the diversity of the Italian immigrant community. Panel A uses the local name index defined in (1) as the dependent variable; Panel B uses the distance-adjusted local name index defined in (2); Panel C uses the foreign name index of Abramitzky et al. (2020). The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. All specifications include municipality and census-year fixed effects. Column (1) includes county fixed effects; columns (2) and (4–6) replace county with more granular hexagon fixed effects. Column (3) replaces municipality fixed effects with municipality-year fixed effects and hexagon fixed effects with hexagon-year fixed effects. Demographic controls (included in column 5) are measured at the district level and include population, the share of women, African Americans, immigrants, and Italian immigrants. Economic controls (included in column 6) are measured at the district level and include employment shares by industry. Standard errors, clustered at the district level, are reported in parentheses. Referenced on page: 24.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.5. Exposure to Immigrant Heterogeneity and National Identity Formation, Variants of Local Name Index and Diversity Definitions

	Italian Diversity Measured by:					
	Municipality -		Province -		Region -	
	(1) District	(2) County	(3) District	(4) County	(5) District	(6) County
Panel A: Municipality-Level Local Name Index						
Italian Diversity	-0.164***	-0.126**	-0.192***	-0.043	-0.200***	0.116
	(0.015)	(0.051)	(0.022)	(0.110)	(0.029)	(0.152)
Mean Dep. Var.	3.989	3.989	3.989	3.989	3.989	3.989
Panel B: Province-Level Local Name Index						
Italian Diversity	-0.105***	-0.067	-0.129***	-0.064	-0.152***	-0.045
	(0.012)	(0.042)	(0.018)	(0.096)	(0.025)	(0.137)
Mean Dep. Var.	4.804	4.804	4.804	4.804	4.804	4.804
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Hexagon FE	Yes	Yes	Yes	Yes	Yes	Yes
# Individuals	223,531	223,531	223,531	223,531	223,531	223,531
# Observations	334,303	334,303	334,303	334,303	334,303	334,303

Notes. This table reports the effect of exposure to diversity within the Italian community on name-based indices from specification (4). Panel A uses the local name index defined in (1) as the dependent variable; Panel B uses a variant where the name distribution is taken at the province-level instead of the municipality-level. Each column reports estimates from a variant of equation (3) that differs only in the geographic level at which we construct H. Column (1) uses the baseline municipality–district definition. Column (2) aggregates districts to the county level (municipality–county). Columns (3–4) redefine origins at the province level (province-district and province–county, respectively). Columns (5–6) analogously use regions as origins (region–district and region–county, respectively). Provinces are coarser than municipalities; Regions are coarser than provinces. Counties are coarser than districts. The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. All specifications include municipality, census-year fixed effects, and hexagon fixed effects. Standard errors, clustered at the district level, are reported in parentheses. Referenced on pages: 23, 24.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.6. Movers Design and National Identity Formation, Excluding Firstborn Children

	Local Name Index (LNI)					Distance LNI
	(1)	(2)	(3)	(4)	(5)	(6)
Diversity Shock × Post-Move	-0.430*	-0.429*	-0.509**		-0.059	0.066**
	(0.238)	(0.238)	(0.242)		(1.307)	(0.029)
Diversity Shock (Cont.) × Post-Move				-15.172*		
				(7.807)		
Age					0.006	
					(0.023)	
Diversity Shock × Post-Move × Age					-0.013	
					(0.041)	
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Hexagon FE	Yes	Yes	–	–	–	–
Time Since Move FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	–	–	–	–
Hexagon-Cohort FE	No	No	Yes	Yes	Yes	Yes
# Individuals	72,361	72,361	70,793	70,793	70,793	70,793
# Observations	235,386	235,385	229,669	229,669	229,669	229,669
Mean Dep. Var.	3.913	3.913	3.905	3.905	3.905	9.603

Notes. This table reports results from the movers design specification in (A.4). Unlike in the main analysis, we exclude firstborn children from the estimation sample. The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. The dependent variables are the local name index (columns 1–5) and the distance-adjusted local name index (column 6) defined in (1)–(2). The baseline treatment is the interaction between an indicator equal to one after an individual moves to a different hexagon (and zero otherwise) and an indicator equal to one if the difference in diversity between the Italian communities in the destination and origin districts lies in the top 20% of the distribution (and zero otherwise). Column (4) replaces this indicator with the continuous diversity difference measure. Column (5) additionally includes age and its interaction with the baseline treatment. All specifications include individual and event-time (time since move) fixed effects. Column (2) also includes cohort (year) and hexagon fixed effects, while columns (3–6) replace these with hexagon-by-cohort fixed effects. Standard errors, clustered at the district level, are reported in parentheses. Referenced on page: 23.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.7. Movers Design and National Identity Formation, Fractionalization Index

	Local Name Index (LNI)					Distance LNI
	(1)	(2)	(3)	(4)	(5)	(6)
Diversity Shock × Post-Move	-0.497***	-0.505***	-0.678***		-2.322	0.098***
	(0.155)	(0.155)	(0.164)		(1.796)	(0.021)
Diversity Shock (Cont.) × Post-Move				-582.492**		
				(227.827)		
Age					0.005	
					(0.021)	
Diversity Shock × Post-Move × Age					0.049	
					(0.053)	
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Hexagon FE	Yes	Yes	–	–	–	–
Time Since Move FE	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	No	Yes	–	–	–	–
Hexagon-Cohort FE	No	No	Yes	Yes	Yes	Yes
# Individuals	103,079	103,079	101,214	101,214	101,214	101,214
# Observations	359,898	359,898	352,641	352,641	352,641	352,641
Mean Dep. Var.	3.905	3.905	3.898	3.898	3.898	9.602

Notes. This table reports results from the movers design specification in (A.4). Unlike in the main analysis, we use the fractionalization index (= 1− Hirschman-Herfindahl Index) to measure the diversity of the Italian immigrant community. The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. The dependent variables are the local name index (columns 1–5) and the distance-adjusted local name index (column 6) defined in (1)–(2). The baseline treatment is the interaction between an indicator equal to one after an individual moves to a different hexagon (and zero otherwise) and an indicator equal to one if the difference in diversity between the Italian communities in the destination and origin districts lies in the top 20% of the distribution (and zero otherwise). Column (4) replaces this indicator with the continuous diversity difference measure. Column (5) additionally includes age and its interaction with the baseline treatment. All specifications include individual and event-time (time since move) fixed effects. Column (2) also includes cohort (year) and hexagon fixed effects, while columns (3–6) replace these with hexagon-by-cohort fixed effects. Standard errors, clustered at the district level, are reported in parentheses. Referenced on page: 24.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.8. Movers Design and National Identity Formation, Variants of Local Name Index and Diversity Definitions

	Italian Diversity Measured by:					
	Municipality -		Province -		Region -	
	(1) District	(2) County	(3) District	(4) County	(5) District	(6) County
Panel A: Municipality-Level Local Name Index						
Diversity Shock × Post-Move	-0.335**	-0.234	-0.423***	-0.236	-0.430***	-0.237*
	(0.137)	(0.169)	(0.140)	(0.176)	(0.133)	(0.134)
Mean Dep. Var.	3.909	3.909	3.909	3.909	3.909	3.909
Panel B: Province-Level Local Name Index						
Diversity Shock × Post-Move	-0.348***	-0.123	-0.340***	-0.095	-0.348***	-0.104
	(0.124)	(0.159)	(0.123)	(0.167)	(0.122)	(0.124)
Mean Dep. Var.	4.678	4.678	4.678	4.678	4.678	4.678
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Time Since Move FE	Yes	Yes	Yes	Yes	Yes	Yes
# Individuals	114,052	114,052	114,052	114,052	114,052	114,052
# Observations	410,343	410,343	410,343	410,343	410,343	410,343

Notes. This table reports results from the movers design specification in (A.4). The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. Panel A uses the local name index defined in (1) as the dependent variable; Panel B uses a variant where the name distribution is taken at the province-level instead of the municipality-level. The treatment is the interaction between an indicator equal to one after an individual moves to a different hexagon (and zero otherwise) and an indicator equal to one if the difference in diversity between the Italian communities in the destination and origin districts lies in the top 20% of the distribution (and zero otherwise). Each column reports estimates from a variant of equation (3) that differs only in the geographic level at which we construct H. Column (1) uses the baseline municipality–district definition. Column (2) aggregates districts to the county level (municipality–county). Columns (3–4) redefine origins at the province level (province-district and province–county, respectively). Columns (5–6) analogously use regions as origins (region–district and region–county, respectively). Provinces are coarser than municipalities; Regions are coarser than provinces. Counties are coarser than districts. All specifications include individual and event-time (time since move) fixed effects. Standard errors, clustered at the district level, are reported in parentheses. Referenced on pages: 23, 24.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.9. Name Particularism in Italy by Exposure to Immigrant Diversity, Alternative Mass Migration Threshold

	Threshold of Emigration Rate for Treatment:					
	1 ‰		2‰		5‰	
	(1) LNI	(2) D-LNI	(3) LNI	(4) D-LNI	(5) LNI	(6) D-LNI
Diversity Exposure × Post Mass Migration	-0.464***	0.045***	-0.634***	0.062***	-0.824***	0.077***
	(0.091)	(0.010)	(0.125)	(0.014)	(0.177)	(0.020)
Post Mass Migration	0.010	-0.001	-0.008	0.000	-0.060	0.006
	(0.025)	(0.003)	(0.029)	(0.003)	(0.038)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
# Municipalities	7,823	7,823	7,823	7,823	7,823	7,823
# Observations	200,774	200,774	200,774	200,774	200,774	200,774
Mean Dep. Var.	0.879	9.909	0.879	9.909	0.879	9.909

Notes. This table reports the effect of exposure to diversity within the Italian migrant community on name-based indices in Italian municipalities from regression (8). The unit of observation is a municipality observed at a decennial frequency between 1880 and 1930. The dependent variables are the local name index defined in (1) in columns (1), (3), and (5), and the distance-adjusted local name index defined in (2) in columns (2), (4), and (6). The baseline treatment is defined as the share of a municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3) interacted with a post indicator equals one in years at or after the municipality-specific onset of mass migration (defined as the first year with at least one (columns 1–2), two (columns 3–4), or five (columns 5–6) emigrant per 1,000 inhabitants). The table reports the coefficient of interaction terms between the diversity exposure measure and the post indicator. All regressions include municipality and year fixed effects. Standard errors are clustered at the municipality level and reported in parentheses. Referenced on page: 25.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.10. Monuments by Designation in Italy

	National Monuments					Other Monuments				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Diversity Exposure × $I(1870 \leq t < 1880)$	-0.002	0.004	0.000	-0.003	-0.001	0.055	0.074*	0.066	0.046	0.036
	(0.007)	(0.004)	(0.005)	(0.007)	(0.008)	(0.034)	(0.041)	(0.041)	(0.035)	(0.033)
Diversity Exposure × $I(1880 \leq t < 1890)$	-0.005	0.005	0.002	-0.010	-0.011	0.040	0.092**	0.087*	0.047	0.034
	(0.010)	(0.007)	(0.008)	(0.010)	(0.011)	(0.042)	(0.046)	(0.046)	(0.042)	(0.043)
Diversity Exposure × $I(1890 \leq t < 1900)$	-0.004	0.002	-0.000	-0.004	-0.004	0.083*	0.108**	0.090*	0.083*	0.068
	(0.007)	(0.004)	(0.005)	(0.007)	(0.007)	(0.043)	(0.051)	(0.051)	(0.043)	(0.045)
Diversity Exposure × $I(1900 \leq t < 1910)$	-0.012	-0.016	-0.016	-0.014*	-0.014	0.019	0.090*	0.064	0.013	-0.041
	(0.009)	(0.010)	(0.011)	(0.009)	(0.009)	(0.045)	(0.053)	(0.053)	(0.046)	(0.042)
Diversity Exposure × $I(1910 \leq t < 1920)$	0.082**	0.019	0.011	0.075**	0.094**	0.053	0.112**	0.100**	0.048	0.037
	(0.038)	(0.045)	(0.047)	(0.038)	(0.045)	(0.040)	(0.047)	(0.047)	(0.041)	(0.043)
Diversity Exposure × $I(1920 \leq t < 1930)$	0.139**	0.111	0.076	0.136**	0.148**	-0.038	0.040	0.031	-0.040	-0.067
	(0.066)	(0.072)	(0.070)	(0.066)	(0.075)	(0.041)	(0.048)	(0.048)	(0.041)	(0.042)
Diversity Exposure × $I(1930 \leq t < 1940)$	-0.023	-0.034	-0.039	-0.035	-0.039	-0.032	0.015	0.006	-0.043	-0.048
	(0.028)	(0.028)	(0.028)	(0.028)	(0.033)	(0.027)	(0.028)	(0.028)	(0.028)	(0.031)
Diversity Exposure × $I(1940 \leq t < 1950)$	0.019	0.024	0.001	0.015	0.022	-0.035	-0.012	-0.016	-0.033	-0.042
	(0.021)	(0.028)	(0.028)	(0.022)	(0.024)	(0.024)	(0.027)	(0.026)	(0.024)	(0.029)
Diversity Exposure × $I(1950 \leq t < 1960)$	0.122***	0.109***	0.104***	0.107***	0.131***	-0.043	-0.031	-0.033	-0.037	-0.046
	(0.032)	(0.036)	(0.035)	(0.033)	(0.037)	(0.031)	(0.033)	(0.033)	(0.032)	(0.036)
Diversity Exposure × $I(1960 \leq t < 1970)$	-0.015	-0.004	-0.013	-0.011	-0.004	-0.002	0.004	0.010	0.013	0.013
	(0.023)	(0.025)	(0.025)	(0.024)	(0.026)	(0.023)	(0.027)	(0.027)	(0.023)	(0.027)
Diversity Exposure × $I(1970 \leq t < 1980)$	-0.027	0.000	-0.014	-0.024	-0.021	-0.028	-0.034	-0.028	-0.014	-0.019
	(0.036)	(0.039)	(0.038)	(0.036)	(0.033)	(0.021)	(0.021)	(0.021)	(0.022)	(0.023)
Diversity Exposure × $I(1980 \leq t < 1990)$	-0.010	-0.020	-0.030	-0.009	-0.006	0.023	-0.008	-0.004	0.036	0.023
	(0.022)	(0.024)	(0.024)	(0.022)	(0.025)	(0.024)	(0.025)	(0.024)	(0.025)	(0.028)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Decade FE	Yes	–	–	Yes	Yes	Yes	–	–	Yes	Yes
Region-Decade FE	No	Yes	–	No	No	No	Yes	–	No	No
Province-Decade FE	No	No	Yes	No	No	No	No	Yes	No	No
US Emigrants × Decade FE	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Controls × Decade FE	No	No	No	No	Yes	No	No	No	No	Yes
# Municipalities	2,549	2,549	2,549	2,549	2,399	2,549	2,549	2,549	2,549	2,399
# Observations	33,137	33,137	33,137	33,137	31,187	33,137	33,137	33,137	33,137	31,187
Mean Dep. Var.	0.043	0.043	0.043	0.043	0.043	0.041	0.041	0.041	0.041	0.042

Notes. This table reports dynamic estimates obtained from (10) of the effect of exposure to diversity within the Italian migrant community on monument construction in Southern Italy. The unit of observation is a municipality-decade for municipalities observed between 1861 and 2000. In columns (1–5), the dependent variable is an indicator equal to one if the municipality has at least one national-identity monument in a given decade. In columns (6–10), the dependent variable is an indicator equal to one if the municipality has at least one non-national-identity monument in a given decade. The baseline treatment is an interaction term between decade indicators and the share of U.S. emigrants from each municipality who settled in districts in the top quintile of the diversity measure defined in (3). In columns (1), (4), (5), (6), (9), and (10), the specification includes municipality fixed effects and decade fixed effects; in columns (2) and (7), we include region-decade fixed effects; in columns (3) and (8), we add province-decade fixed effects; in columns (4) and (9), we control for the number of U.S. emigrants interacted with decade dummies; in columns (5) and (10), we further control for the population in 1881 and in 1991, the number of WW1 deaths, and an indicator for municipalities on the coast, interacted with decade dummies. Standard errors are clustered at the municipality level and are reported in parentheses. Referenced on page: 29.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

Table E.11. U.S. Emigrants' Participation in WW1 and Diversity Exposure

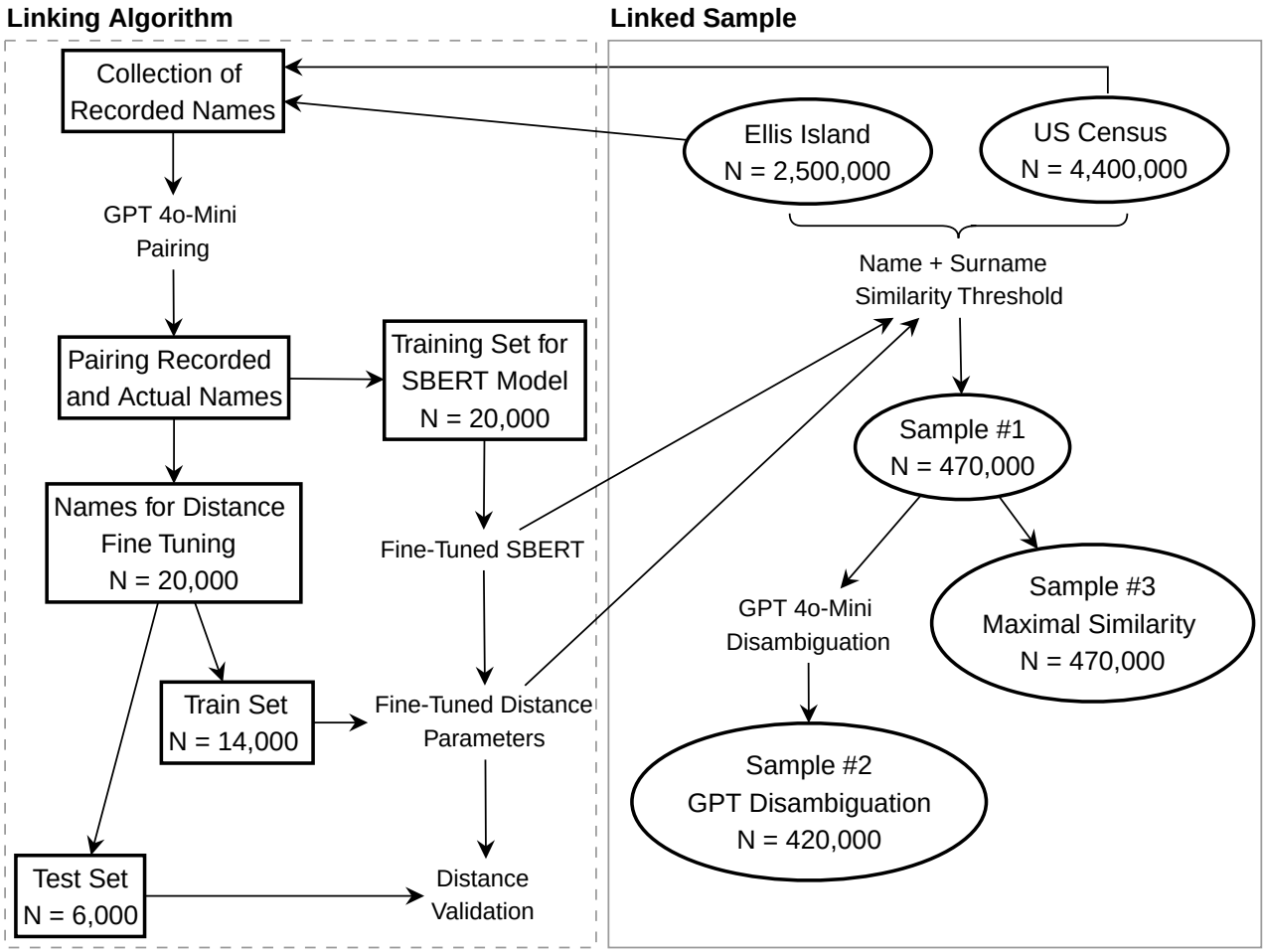
	Death Rate Among US Emigrants (%)				Overall Death Rate (%)
	(1)	(2)	(3)	(4)	(5)
Diversity Exposure	4.527***	4.046***	4.019***	3.979***	0.127
	(0.576)	(0.579)	(0.584)	(0.591)	(0.098)
US Emigrants (1,000s)			0.186	-0.287	-0.100
			(0.179)	(0.205)	(0.064)
Region FE	Yes	–	–	–	–
Province FE	No	Yes	Yes	Yes	Yes
Controls	No	No	No	Yes	Yes
# Municipalities	8,100	8,100	8,100	8,023	8,023
Mean Dep. Var.	1.926	1.926	1.926	1.940	1.979

Notes. This table reports the correlation between the death rate during WW1 of U.S. emigrants (columns 1–4), the overall WW1 death rate (column 5), and exposure to diversity, as in (11). The unit of observation is a municipality. Each municipality is observed once; the outcome variable is measured after WW1; the control variables are measured before 1914. The dependent variables are the share of U.S. emigrants who died during WW1 (columns 1–4), and the WW1 death rate, computed as the ratio between the number of deaths and population in 1881. The baseline treatment is defined as the share of a municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). In columns (3–5), we further control for the number of U.S. emigrants, in thousands. In column (1), we include region fixed effects; in columns (2–5), we include province fixed effects. Columns (3–5) further include municipality-level controls: population in 1881, area, and altitude. Standard errors are clustered at the municipality level and reported in parentheses. Referenced on page: 30.

***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.10$

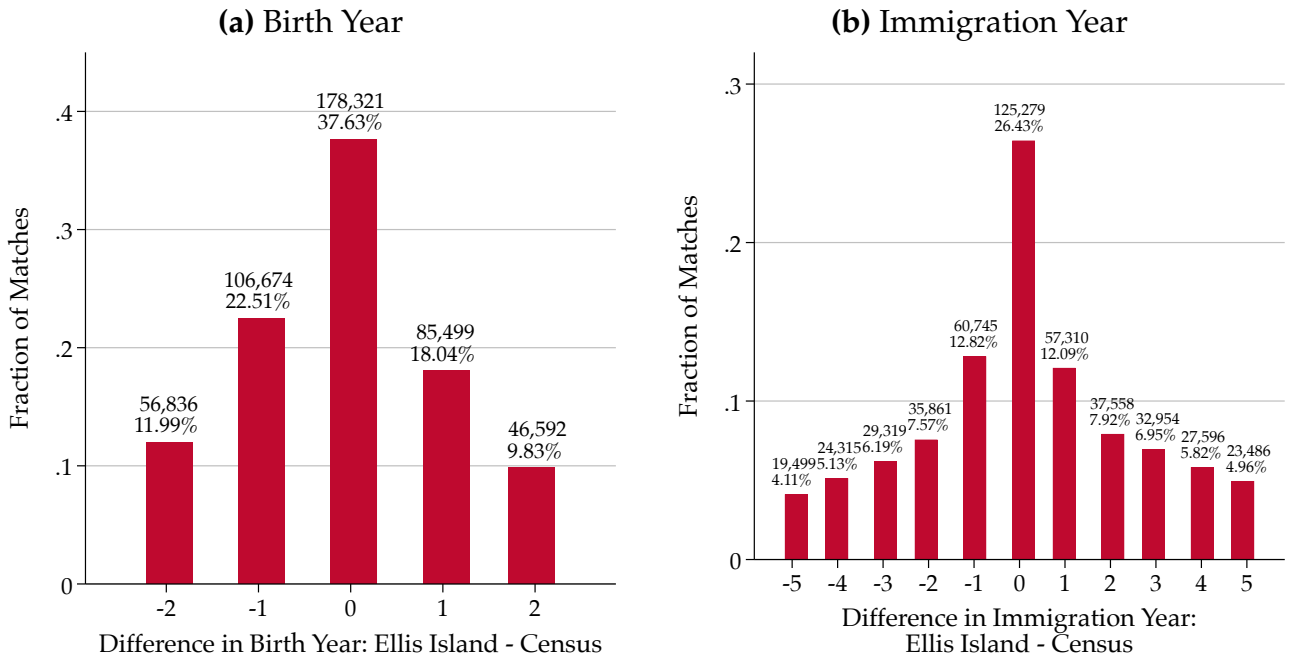
F ADDITIONAL FIGURES

Figure F.1. Visual Representation of the Linking Algorithm



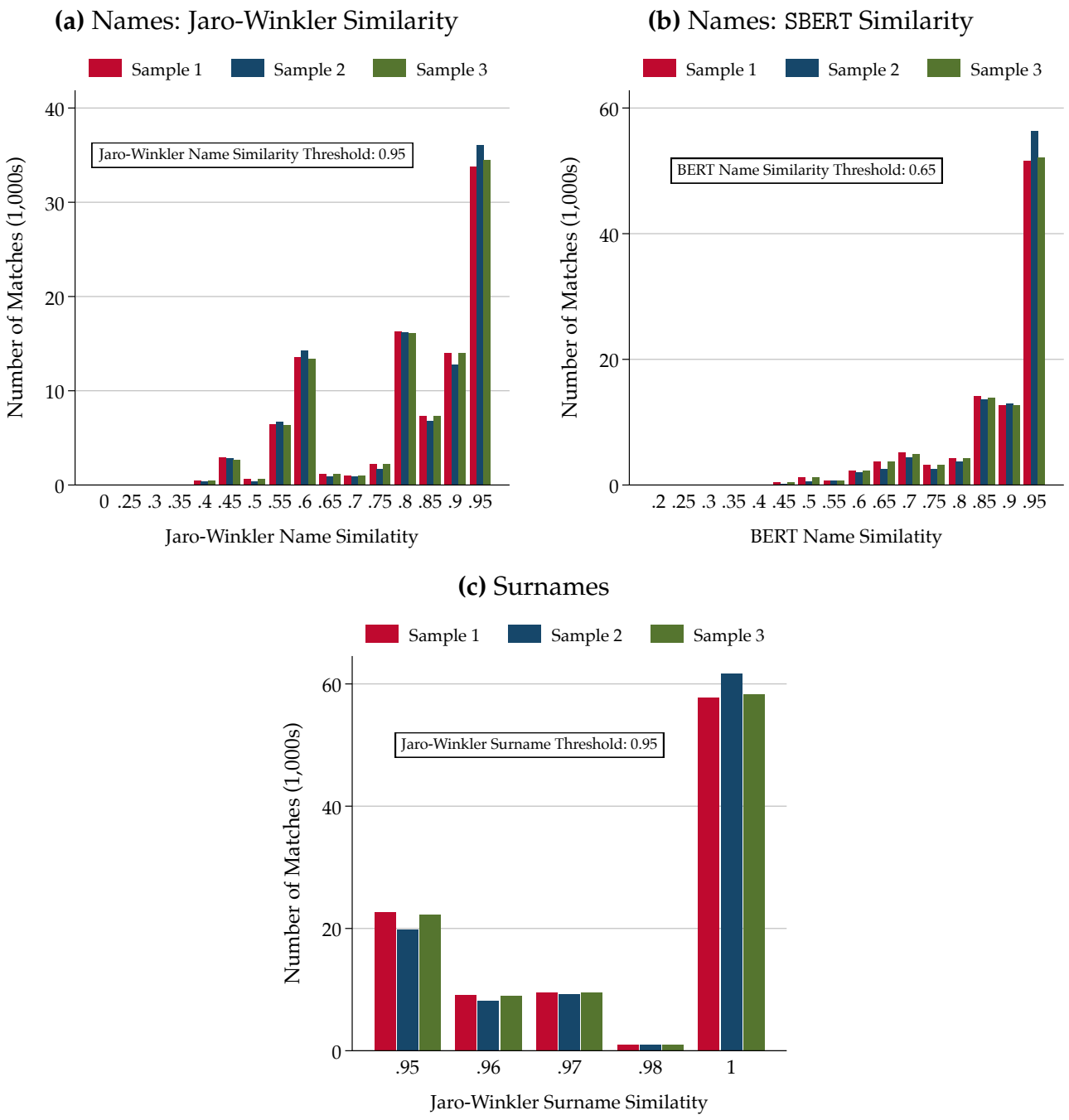
Notes. This Figure provides a graphical overview of the linking algorithm. The round boxes depict the starting data (Ellis Island and U.S. Census) and the final linked sample output. The square boxes denote the dataset we constructed from the original datasets to implement the training and validation of the SBERT and Jaro-Winkler name similarities. The text not embedded in squares or circles denotes logical operations applied to the source dataset to reach the target output. Sample sizes are approximate. Referenced on pages: 9, B11.

Figure F.2. Linked Sample: Blocking Variables Discrepancy



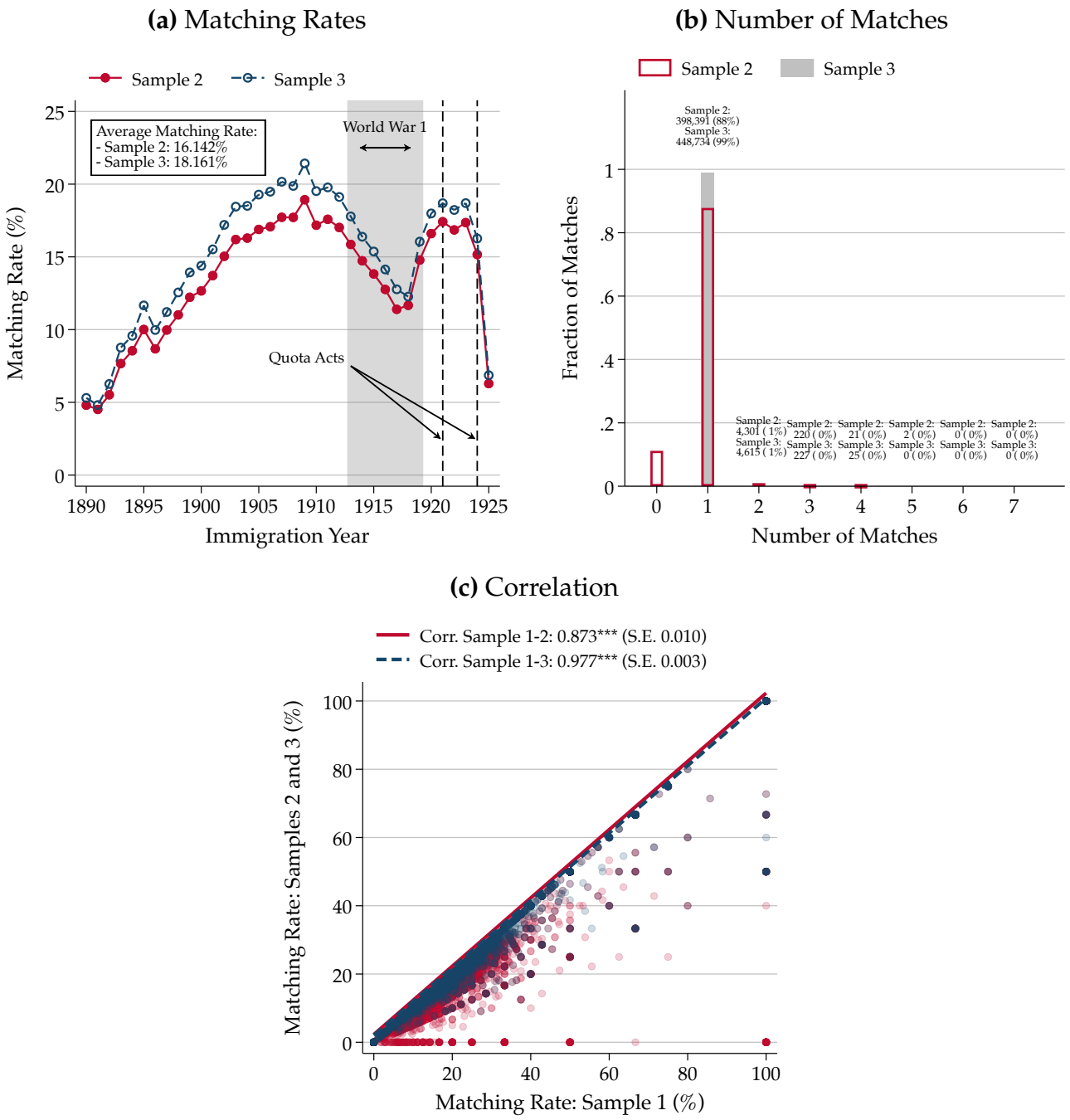
Notes. This figure displays differences between birth year (Panel F.2a) and immigration year (Panel F.2b) as recorded in the census and the Ellis Island datasets. To account for misreporting in the linkage procedure, matches are allowed to differ by up to two years in birth year and up to five years in immigration year. Each bar reports the number of matches in each difference bin and the corresponding share of the total sample. Referenced on page: B15.

Figure F.3. Linked Sample: Match Quality



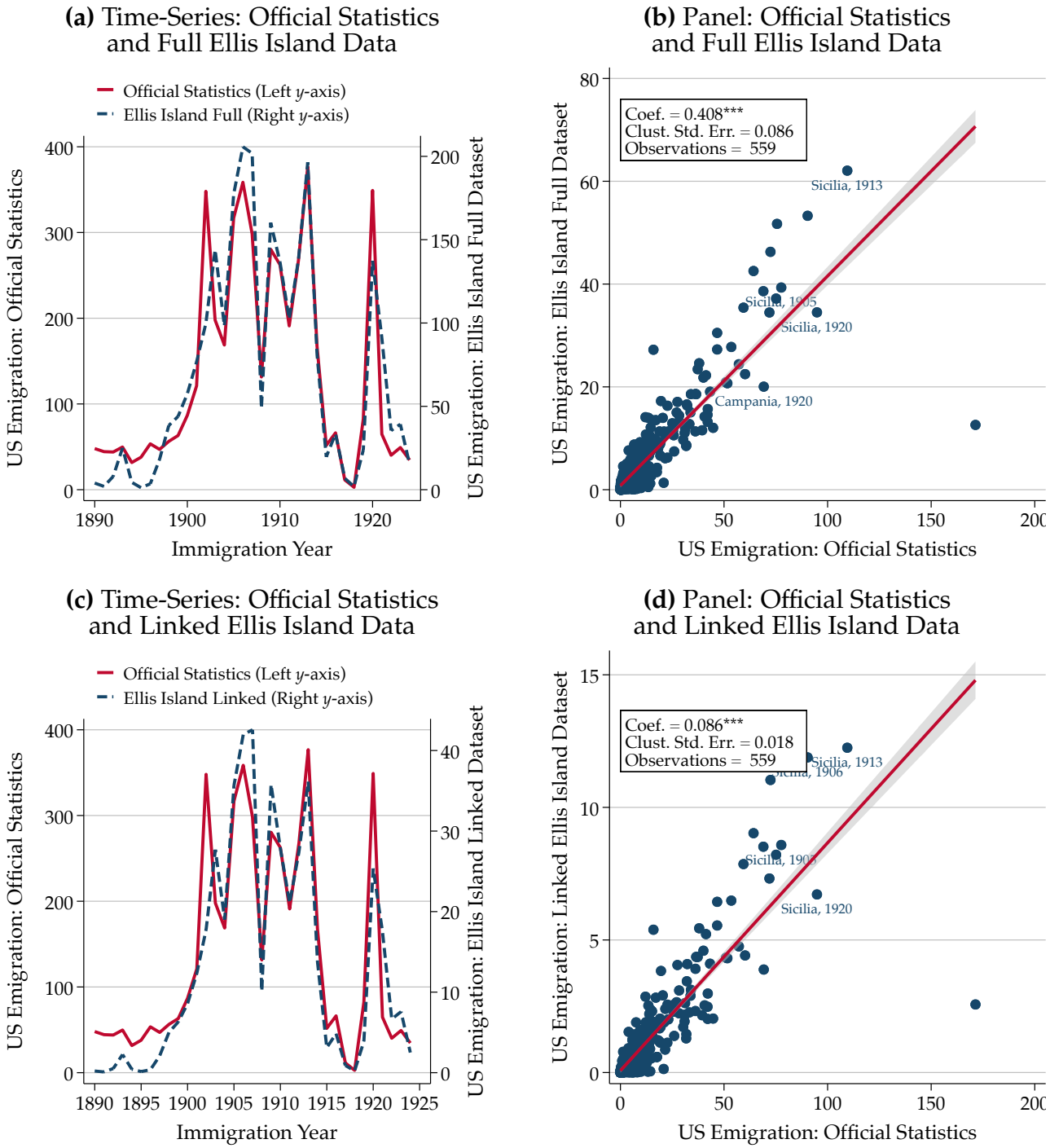
Notes. This figure reports the quality of matches produced by the linking algorithm. Red bars correspond to the baseline linked sample (Sample 1), while blue and green bars correspond to refinements 2 and 3. Panel F.3a shows the distribution of Jaro–Winkler similarities between first names recorded in the census and in the Ellis Island data; matches below the Jaro–Winkler threshold (0.95) satisfy the SBERT similarity criterion. Panel F.3b reports the distribution of SBERT similarities; matches below the SBERT threshold (0.65) satisfy the Jaro–Winkler rule. Panel F.3c reports the distribution of Jaro–Winkler similarities for last names; all matches exceed the Jaro–Winkler threshold (0.95). The number of matches is expressed in thousands. Referenced on page: B15.

Figure F.4. Linked Sample: Comparisons Between Samples



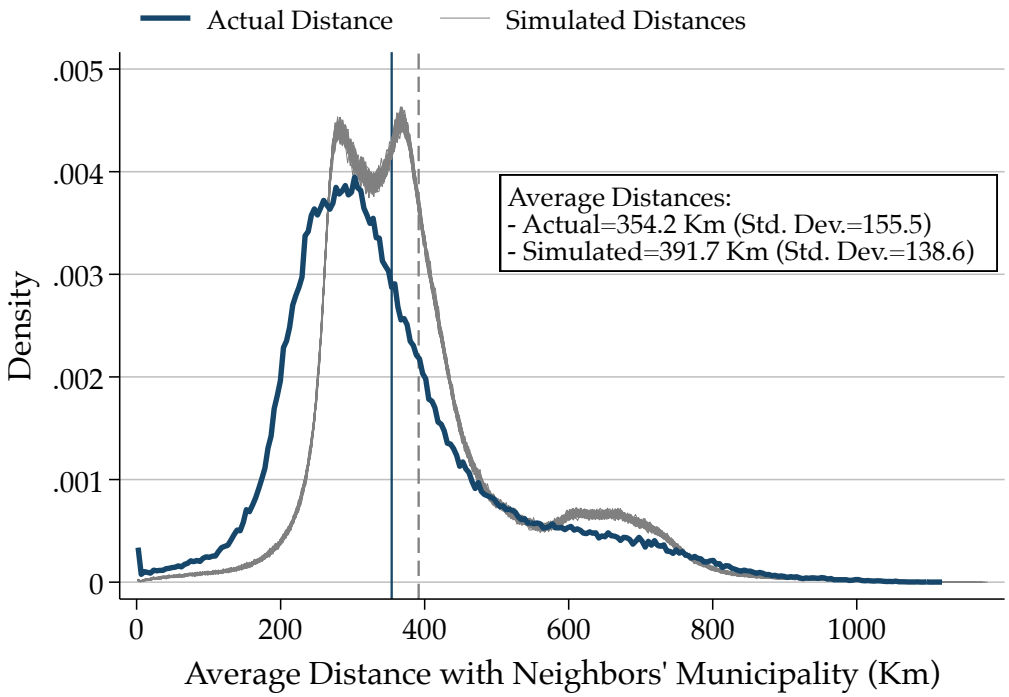
Notes. This figure compares the baseline linked sample with two refinements described in Appendix Section B.III. Red dots correspond to the refinement based on `GPT 4-mini` ("Sample 2"), and blue dots to the maximal-similarity refinement ("Sample 3"). Panel F.4a reports matching rates over time and on average. Panel F.4b shows the distribution of the number of census matches per Ellis Island record and the corresponding shares of the total sample. Panel F.4c plots, at the municipality level, the correlation between the (log) number of migrants in the baseline sample ("Sample 1," x-axis) and in Samples 2 and 3 (y-axis). Reported values give the correlation and its standard error. The solid red and dashed blue lines report the fitted values from a linear relationship between the two series in the two samples. Referenced on pages: 11, B16.

Figure F.5. Linked Sample: Validation with Official Statistics



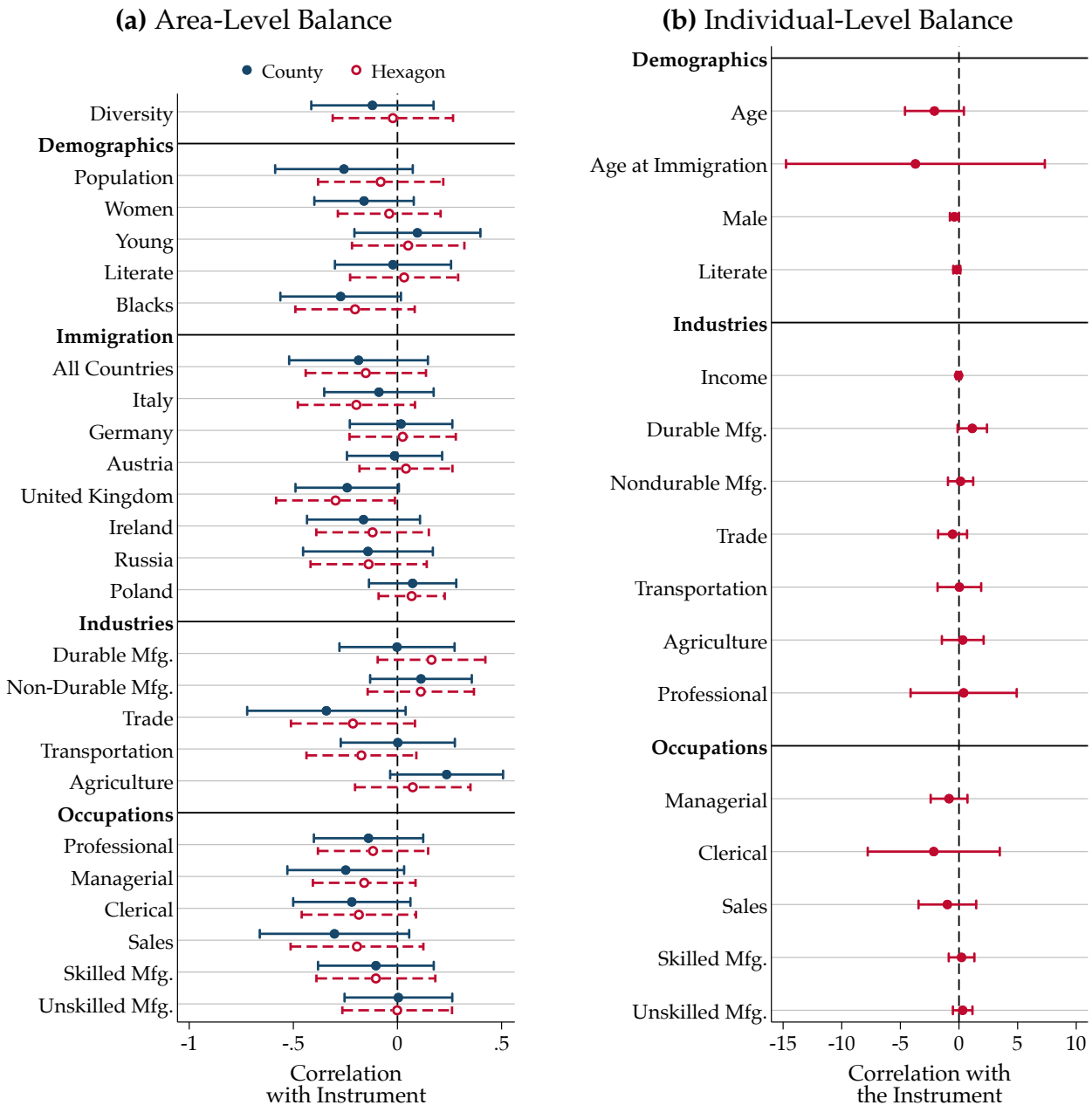
Notes. This figure reports the correlation between the Ellis Island dataset and data compiled from the *Annuario Statistico della Emigrazione Italiana dal 1876 al 1925* published by the General Commissioner for Emigration and containing regional data on U.S. emigration between 1876 and 1925. Panels F.5a and F.5c display the total number of U.S. emigrants recorded in the *Annuario* and the immigrants recorded in the full and linked Ellis Island dataset, respectively. Panels F.5b and F.5d display the correlation between the three datasets. In these figures, a dot is a region-year pair, and each panel separately reports the correlation between data in the *Annuario* and in the Ellis Island full and linked datasets, along with their standard errors clustered by region. The red lines overlay the fitted values from the linear regression and the gray bands display 95% confidence intervals. Referenced on page: B17.

Figure F.6. Distance from Neighbors' Municipality of Origin



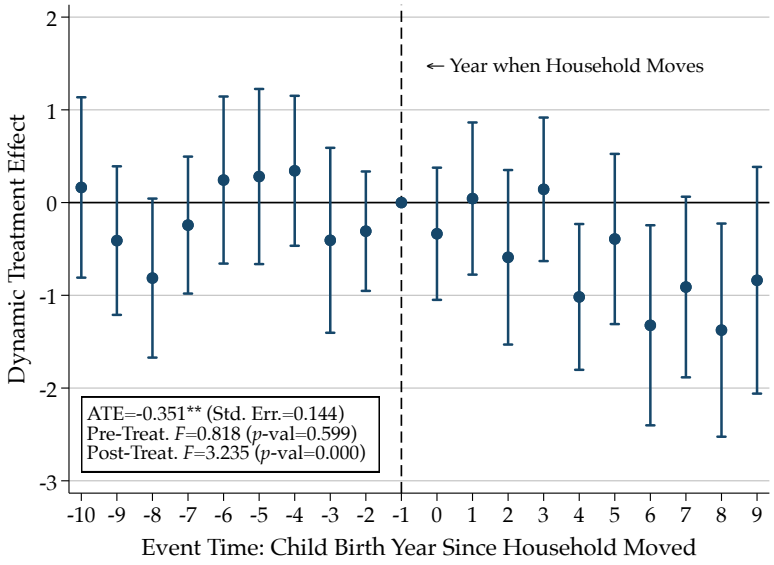
Notes. This figure displays the average distribution of distances between individuals' municipalities of origin within each enumeration district (blue). Gray lines report results from 1,000 simulations in which individuals are randomly allocated across districts; in each simulation, all members of the same household are excluded before reassignment. Excess mass of the blue distribution on the left of the random-matching counterfactual indicates that individuals disproportionately co-locate with immigrants originating from municipalities geographically close to their own. Referenced on page: 16.

Figure F.7. Tests of Instrument Balance



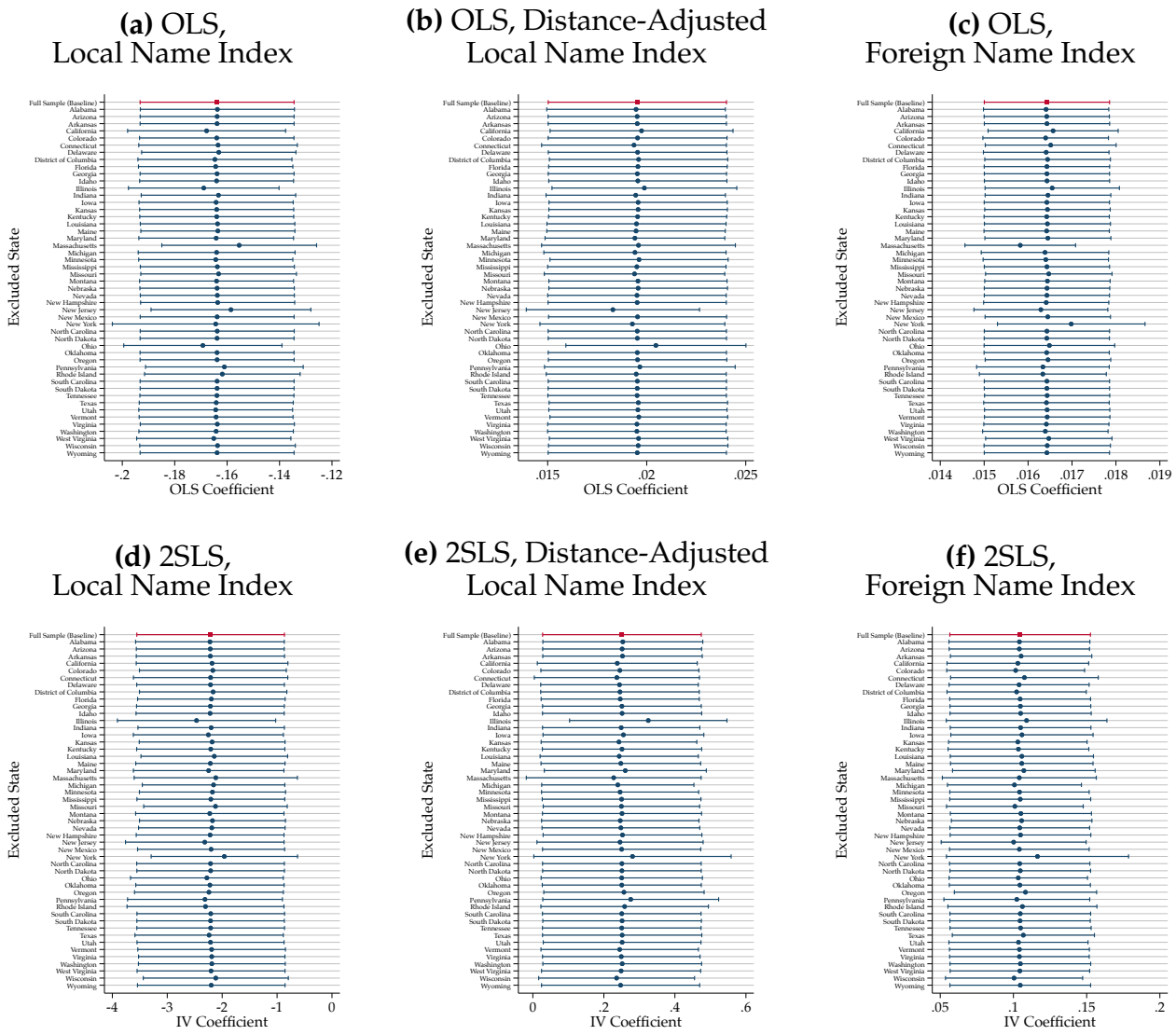
Notes. This figure reports correlations between the area-diversity instrument defined in (5) and area-level (Panel F.7a) and individual-level (Panel F.7b) characteristics. The unit of observation in both panels is an individual observed between 1900 (or the year of immigration, if later) and 1930. In Panel F.7a, each point reports the coefficient from a separate instrumental-variables regression of a predetermined county-level (blue) or hexagon-level (red) characteristic, lagged one census period, on Italian-community diversity instrumented by the ship-based instrument. In Panel F.7b, each point reports the coefficient from a separate instrumental-variables regression of a predetermined individual characteristic, forwarded one census period, on Italian-community diversity instrumented by the ship-based instrument. All specifications include year fixed effects, municipality-of-origin fixed effects, and individual controls (sex, birth year, and immigration year). Panel F.7a additionally includes county or hexagon fixed effects, while Panel F.7b includes hexagon fixed effects. Coefficients are standardized by the standard deviation (Panel F.7a) or the mean (Panel F.7b) of the dependent variable. Standard errors are clustered at the county level for county characteristics and at the hexagon level for hexagon characteristics; bars report 95% confidence intervals. Referenced on pages: 18, 19.

Figure F.8. Movers Design Event-Study Estimates



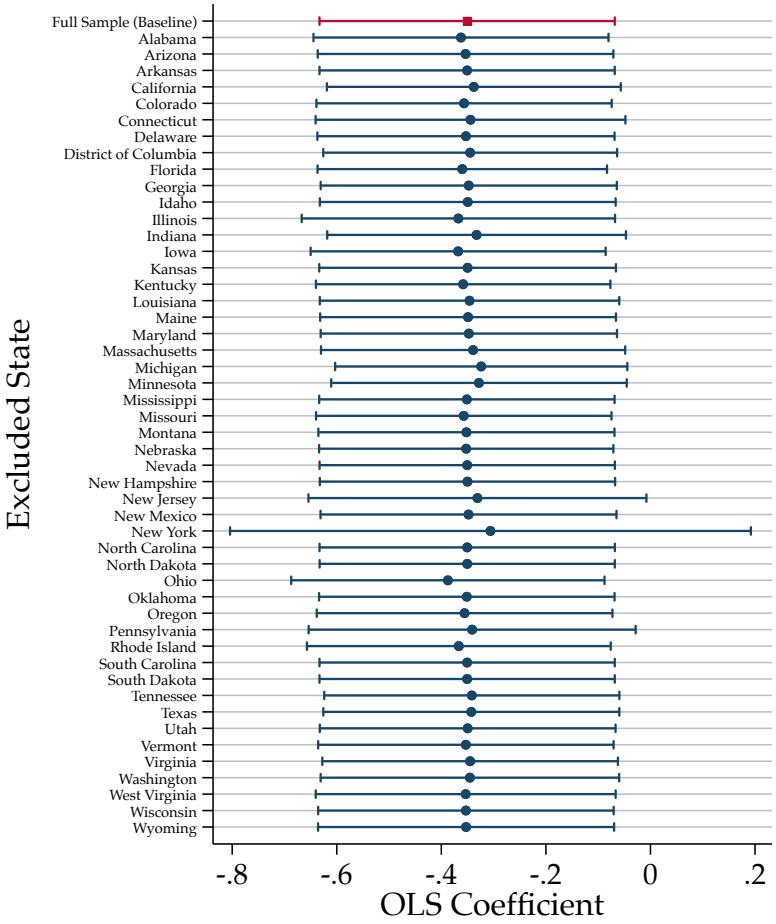
Notes. This figure reports results from the movers design specification in (A.4). The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. The dependent variable is the local name index defined in (1). The treatment is an interaction between an indicator equal to one for individuals experiencing a large diversity shock, defined as a move to a district where the change in Italian-community diversity between the destination and origin lies in the top quintile of the distribution, and a set of event-time indicators relative to the year of the move. The specification includes individual and event-time fixed effects, as well as hexagon-by-birth-year fixed effects. Standard errors are clustered at the hexagon level. Bands report 95% confidence intervals. We report the average treatment effect as well as joint F-tests for the joint significance of pre- and post-treatment coefficients. Referenced on pages: 22, 23.

Figure F.9. State Leave-Out: OLS and 2SLS Panel Regressions



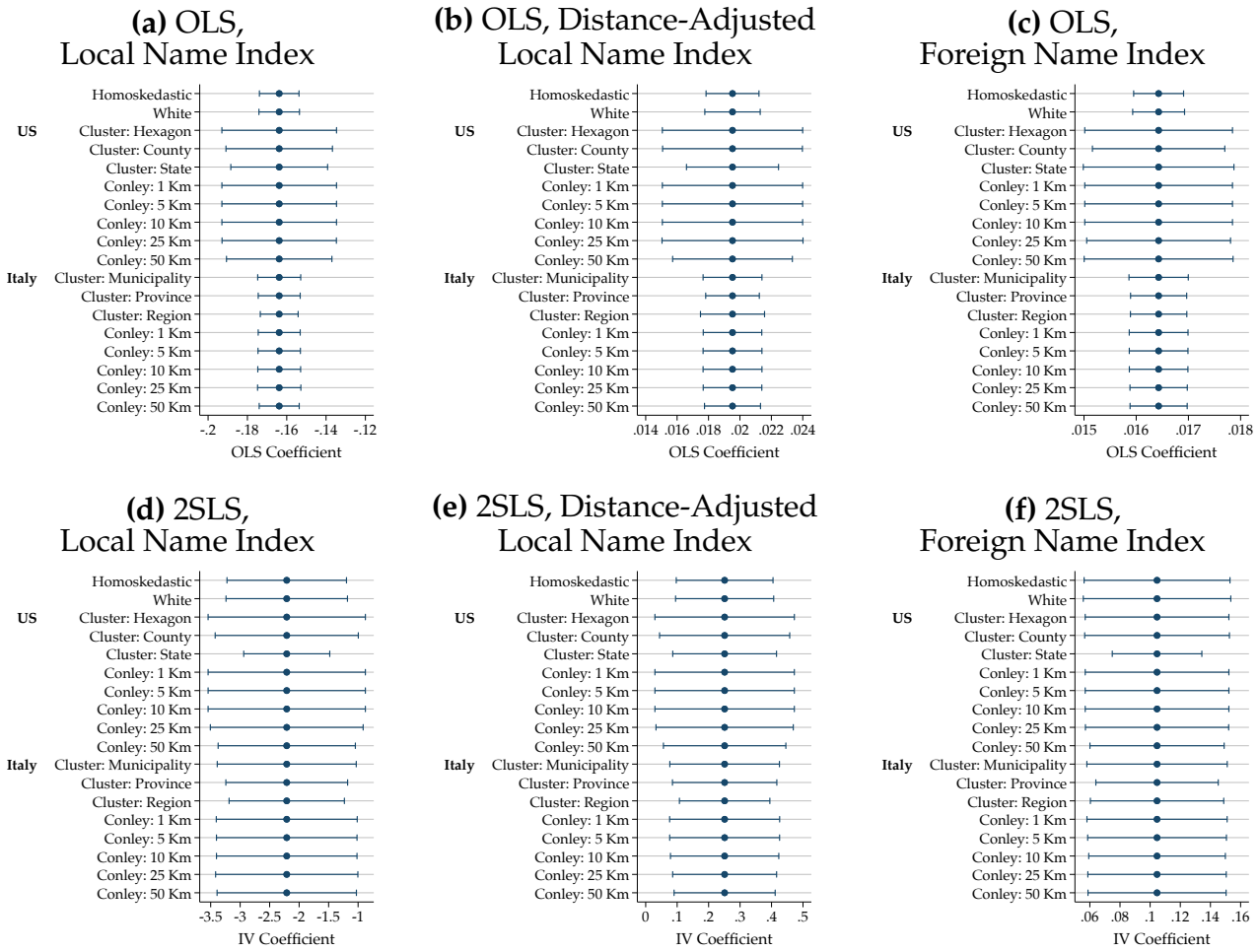
Notes. This figure reports the effect of exposure to diversity within the Italian community on name-based indices. Panels F.9a and F.9d use the local name index defined in (1) as the dependent variable; Panels F.9b and F.9e use the distance-adjusted local name index defined in (2); Panels F.9c and F.9f use the foreign name index of Abramitzky et al. (2020). The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. Panels F.9a, F.9b, and F.9c report ordinary least squares estimates; Panels F.9d, F.9e, and F.9f report two-stage least squares estimates. The first marker in each graph (in red) reports the baseline estimate; the subsequent markers report the estimates obtained when dropping one state at a time from the estimation sample. All specifications include municipality, census year, and hexagon fixed effects. Bands report 95% standard errors, clustered at the district level. Referenced on page: 24.

Figure F.10. State Leave-Out: Movers Regressions



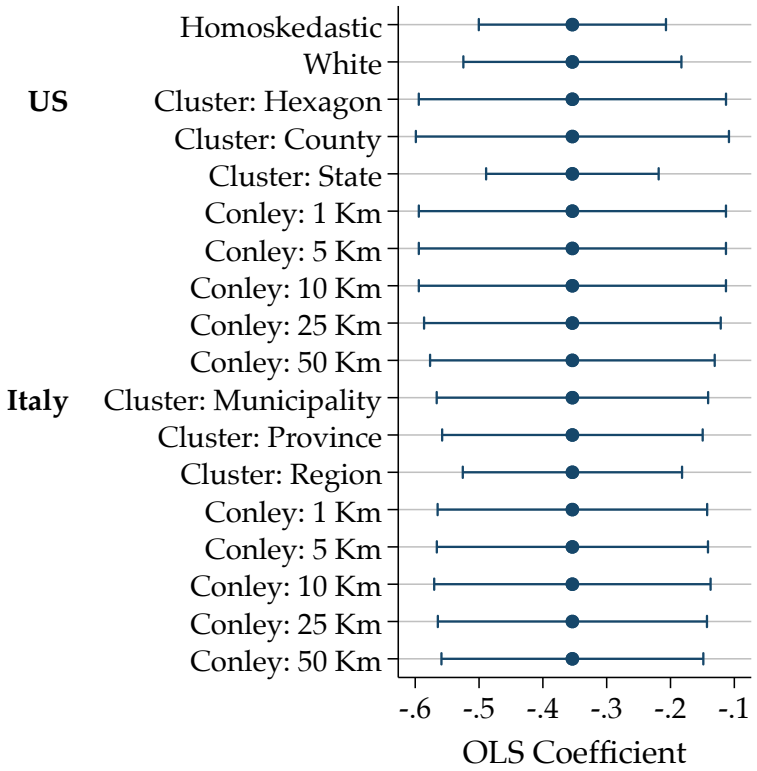
Notes. This figure reports results from the movers design specification in (A.4). The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. The dependent variable is the local name index defined in (1). The treatment is the interaction between an indicator equal to one after an individual moves to a different hexagon (and zero otherwise) and an indicator equal to one if the difference in diversity between the Italian communities in the destination and origin districts lies in the top 20% of the distribution (and zero otherwise). The first marker (in red) reports the baseline estimate; the subsequent markers report the estimates obtained when dropping one state at a time from the estimation sample. The specification includes individual, event-time (time since move), and hexagon-by-cohort fixed effects. Bands report 95% standard errors, clustered at the district level. Referenced on page: 24.

Figure F.11. Alternative Standard Errors: OLS and 2SLS Panel Regressions



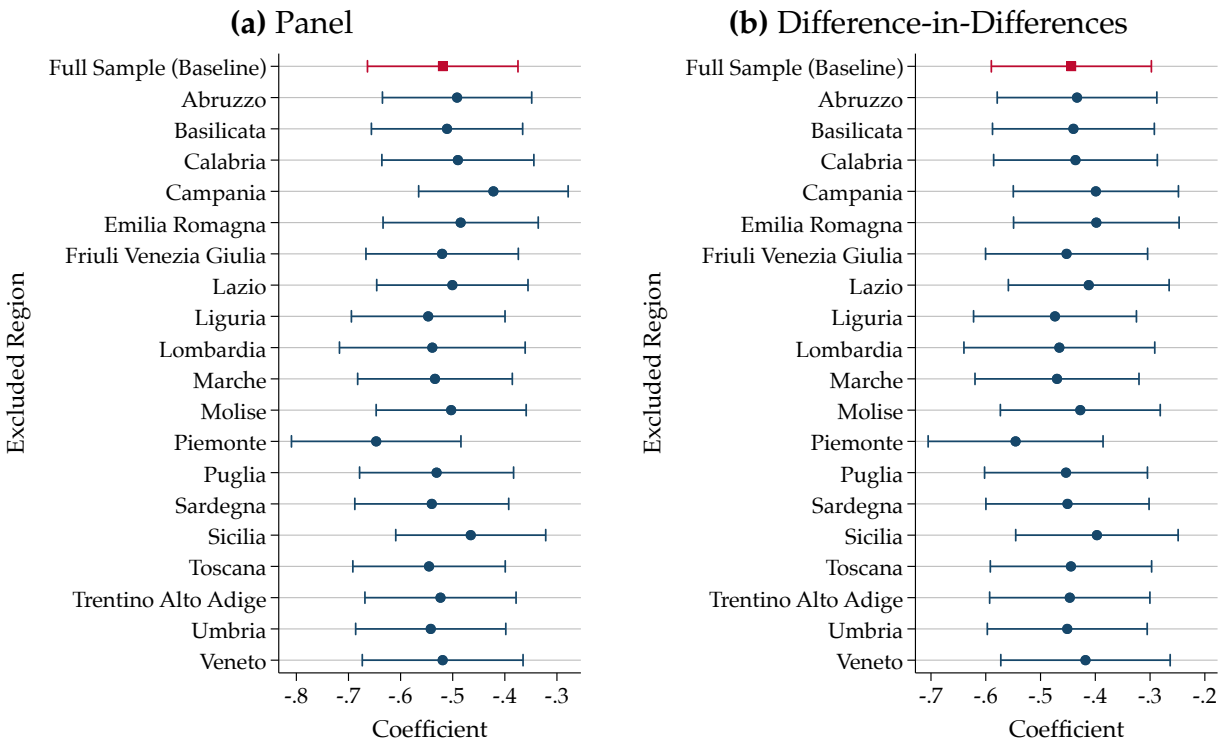
Notes. This figure reports the effect of exposure to diversity within the Italian community on name-based indices. Panels F.11a and F.11d use the local name index defined in (1) as the dependent variable; Panels F.11b and F.11e use the distance-adjusted local name index defined in (2); Panels F.11c and F.11f use the foreign name index of Abramitzky et al. (2020). The unit of observation is an individual observed in the first census in which they appear and, when possible, linked across subsequent censuses between 1900 and 1930. Panels F.11a, F.11b, and F.11c report ordinary least squares estimates; Panels F.11d, F.11e, and F.11f report two-stage least squares estimates. In each regression, we compute the standard errors using different estimators: unadjusted, White, clustered at different U.S. geographical layers, Conley-corrected using various thresholds centered on the hexagon's centroid, clustered at different Italy geographical layers, and Conley-corrected using various thresholds centered on the origin municipality's centroid. All specifications include municipality, census year, and hexagon fixed effects. Bands report 95% standard errors. Referenced on page: 24.

Figure F.12. Alternative Standard Errors: Movers Regressions



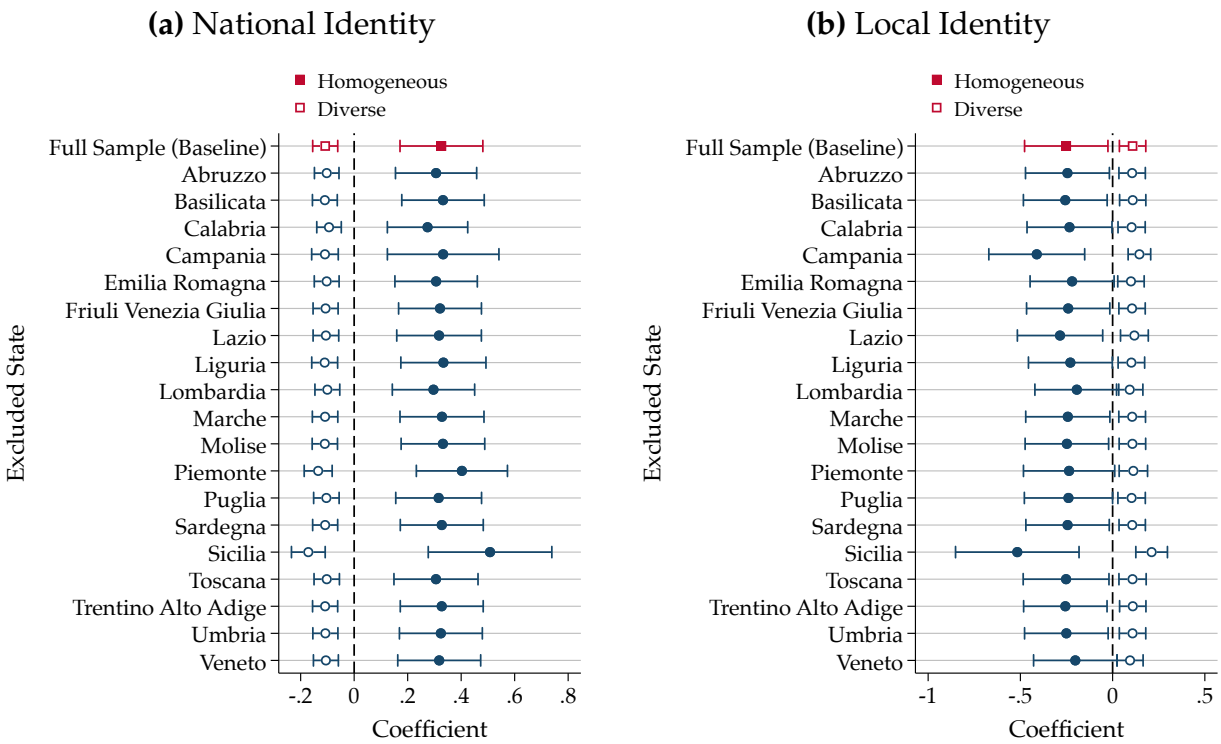
Notes. This figure reports results from the movers design specification in (A.4). The unit of observation is an individual migrant observed in each year in which they have a child born in the United States between 1900 and 1930. The dependent variable is the local name index defined in (1). The treatment is the interaction between an indicator equal to one after an individual moves to a different hexagon (and zero otherwise) and an indicator equal to one if the difference in diversity between the Italian communities in the destination and origin districts lies in the top 20% of the distribution (and zero otherwise). In each regression, we compute the standard errors using different estimators: unadjusted, White, clustered at different U.S. geographical layers, Conley-corrected using various thresholds centered on the hexagon's centroid, clustered at different Italy geographical layers, and Conley-corrected using various thresholds centered on the origin municipality's centroid. The specification includes individual, event-time (time since move), and hexagon-by-cohort fixed effects. Bands report 95% standard errors. Referenced on page: 24.

Figure F.13. Region Leave-Out: Name Particularism in Italy



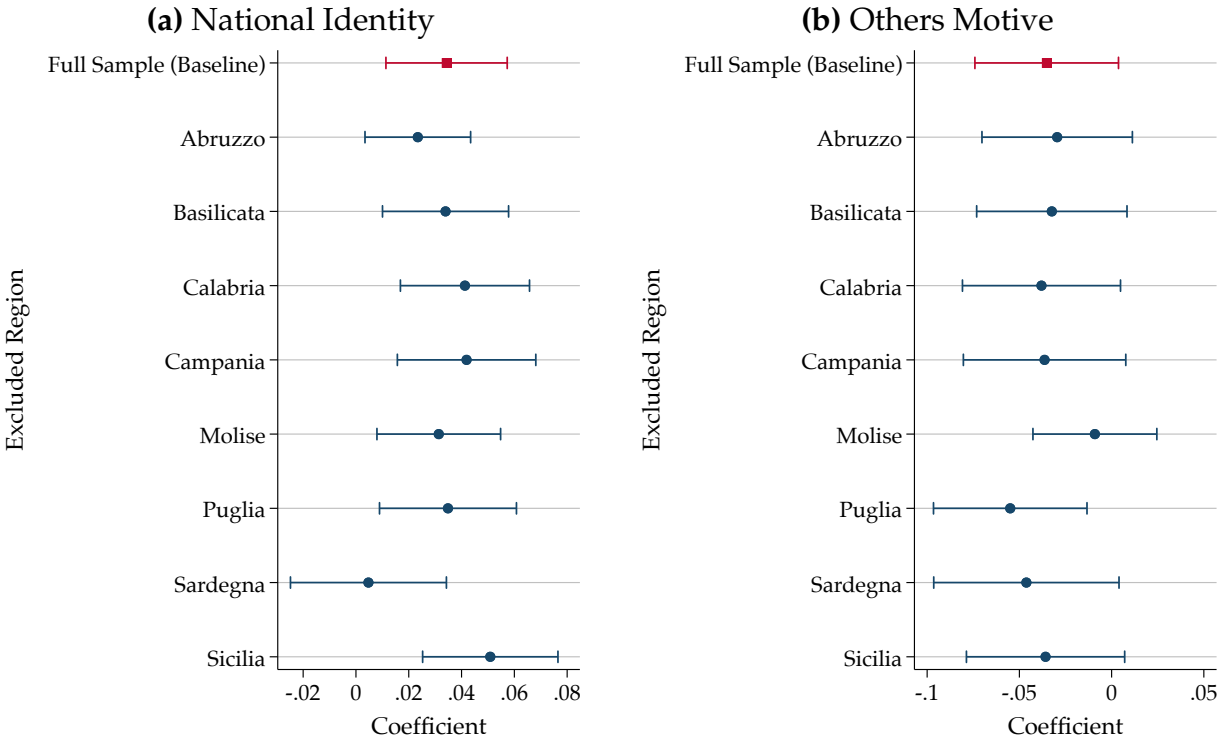
Notes. This figure reports the effect of exposure to diversity within the Italian migrant community on name-based indices in Italian municipalities from equations (7)–(8). The unit of observation is a municipality observed at a decennial frequency between 1880 and 1930. The dependent variable is the local name index defined in (1). In Panel F.13a, each dot reports the coefficient of a treatment term defined as the share of a municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). In Panel F.13b, each dot reports an interaction term between a post indicator equal to one in years at or after the municipality-specific onset of mass migration (defined as the first year with at least one emigrant per 1,000 inhabitants) and the share of the municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). The first marker (in red) reports the baseline estimate; the subsequent markers report the estimates obtained when dropping one region at a time from the estimation sample. All regressions include municipality and year fixed effects. Bands report 95% confidence intervals; Standard errors are clustered at the municipality level. Referenced on page: 30.

Figure F.14. Region Leave-Out: Street Naming in Italy



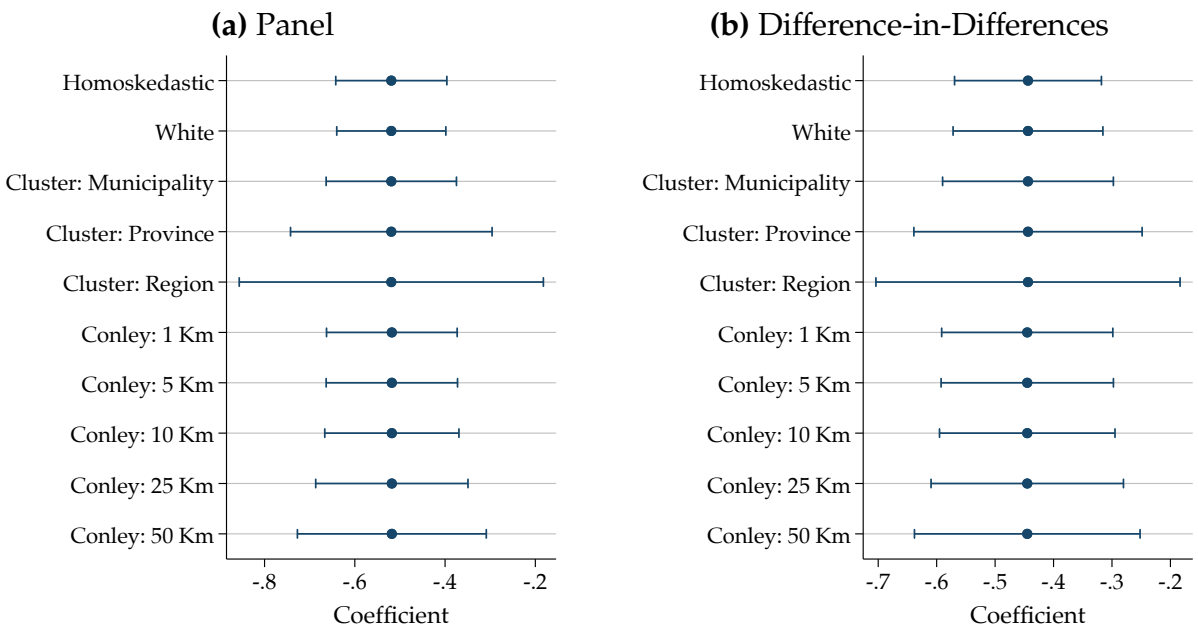
Notes. This figure reports cross-sectional estimates of specification (9) between emigration to diverse Italian communities in the United States and street-naming patterns in Italy. The unit of observation is a municipality. The dependent variables are the shares of streets whose eponyms reflect national identity (Panel F.14a) and local identity (Panel F.14b). The main explanatory variables are the number of emigrants to diverse neighborhoods (displayed by the hollow markers) and the number of emigrants to homogeneous neighborhoods (displayed by the colored markers). The first markers (in red) report the baseline estimates; the subsequent markers report the estimates obtained when dropping one region at a time from the estimation sample. The outcomes are normalized by the total number of streets in the municipality. All specifications include region fixed effects. Bands report 95% confidence intervals; Standard errors are clustered at the municipality level. Referenced on page: 30.

Figure F.15. Region Leave-Out: Monuments in Italy



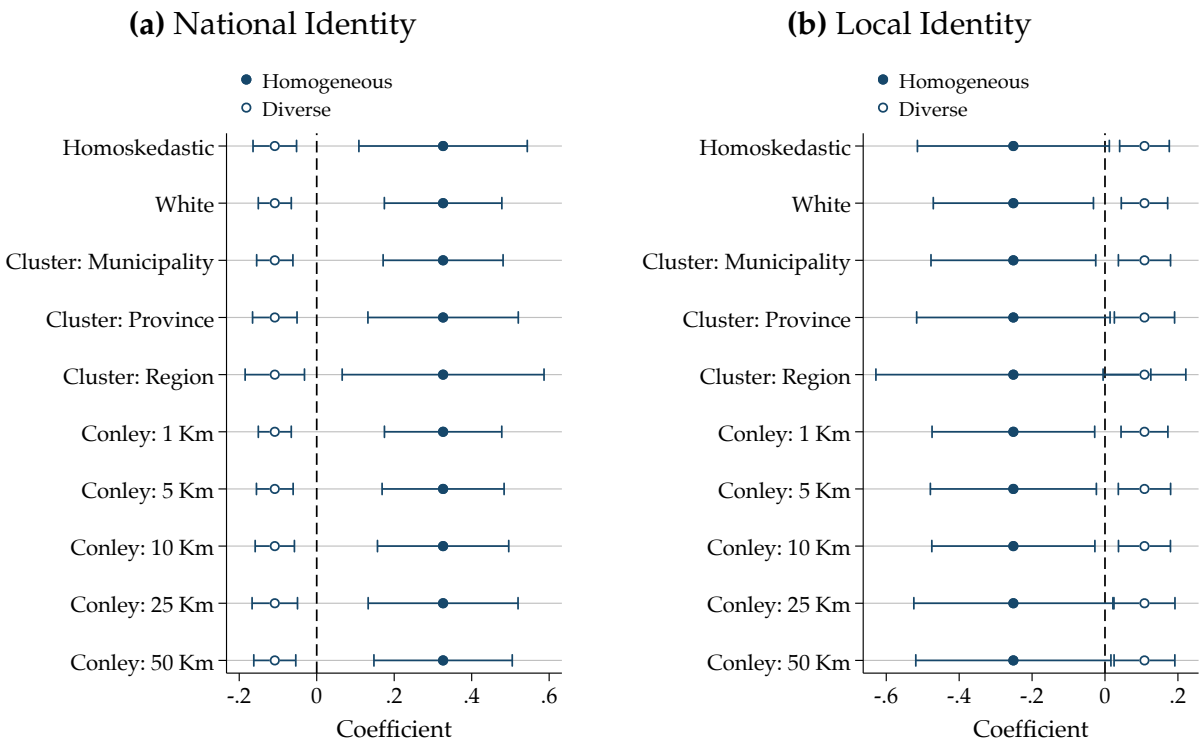
Notes. This figure reports estimates obtained from (10) of the effect of exposure to diversity within the Italian migrant community on monument construction in Southern Italy. The unit of observation is a municipality-decade for municipalities observed between 1861 and 2000. In Panel F.15a, the dependent variable is an indicator equal to one if the municipality has at least one national-identity monument in a given decade. In Panel F.15b, the dependent variable is an indicator equal to one if the municipality has at least one non-national-identity monument in a given decade. The baseline treatment is an interaction term between decade indicators and the share of U.S. emigrants from each municipality who settled in districts in the top quintile of the diversity measure defined in (3). The first markers (in red) report the baseline estimates; the subsequent markers report the estimates obtained when dropping one region at a time from the estimation sample. Each specification includes municipality fixed effects and decade fixed effects. Bands report 95% confidence intervals; Standard errors are clustered at the municipality level. Referenced on page: 30.

Figure F.16. Alternative Standard Errors: Name Particularism in Italy



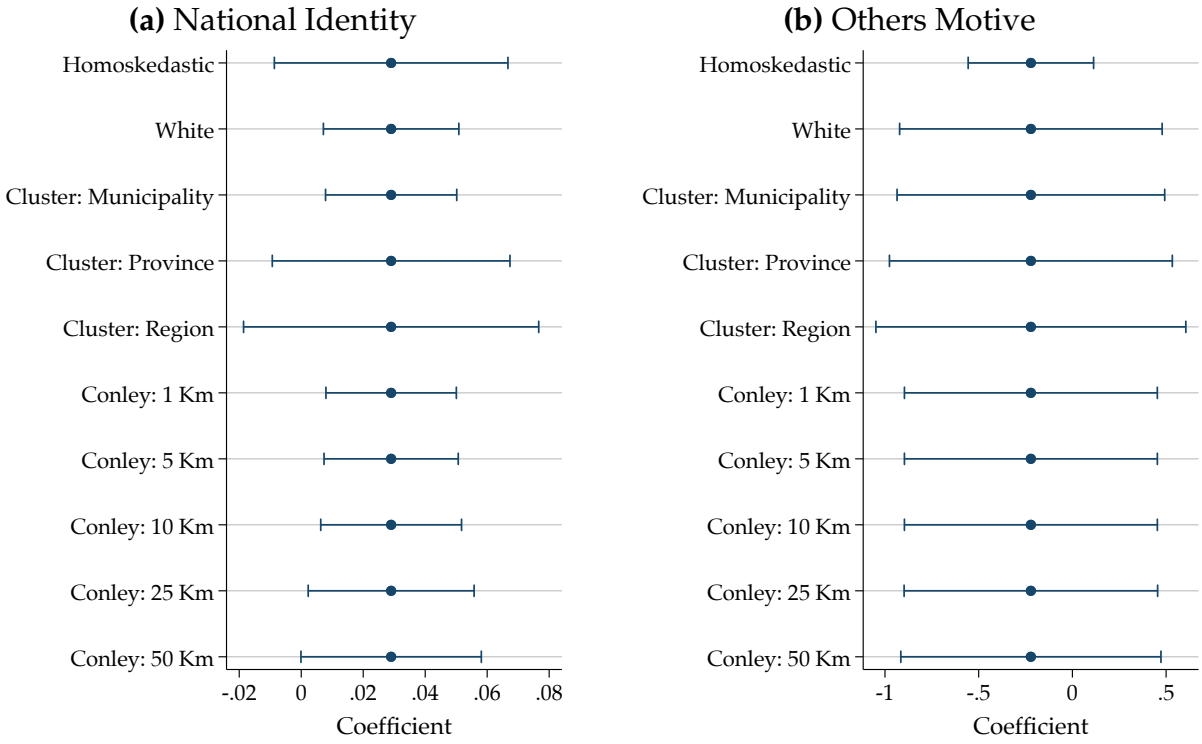
Notes. This figure reports the effect of exposure to diversity within the Italian migrant community on name-based indices in Italian municipalities from equations (7)–(8). The unit of observation is a municipality observed at a decennial frequency between 1880 and 1930. The dependent variable is the local name index defined in (1). In Panel F.16a, each dot reports the coefficient of a treatment term defined as the share of a municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). In Panel F.16b, each dot reports an interaction term between a post indicator equal to one in years at or after the municipality-specific onset of mass migration (defined as the first year with at least one emigrant per 1,000 inhabitants) and the share of the municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). In each regression, we compute the standard errors using different estimators: unadjusted, White, clustered at different geographical layers, and Conley-corrected using various thresholds centered on the municipality's centroid. All regressions include municipality and year fixed effects. Bands report 95% confidence intervals. Referenced on page: 30.

Figure F.17. Alternative Standard Errors: Street Naming in Italy



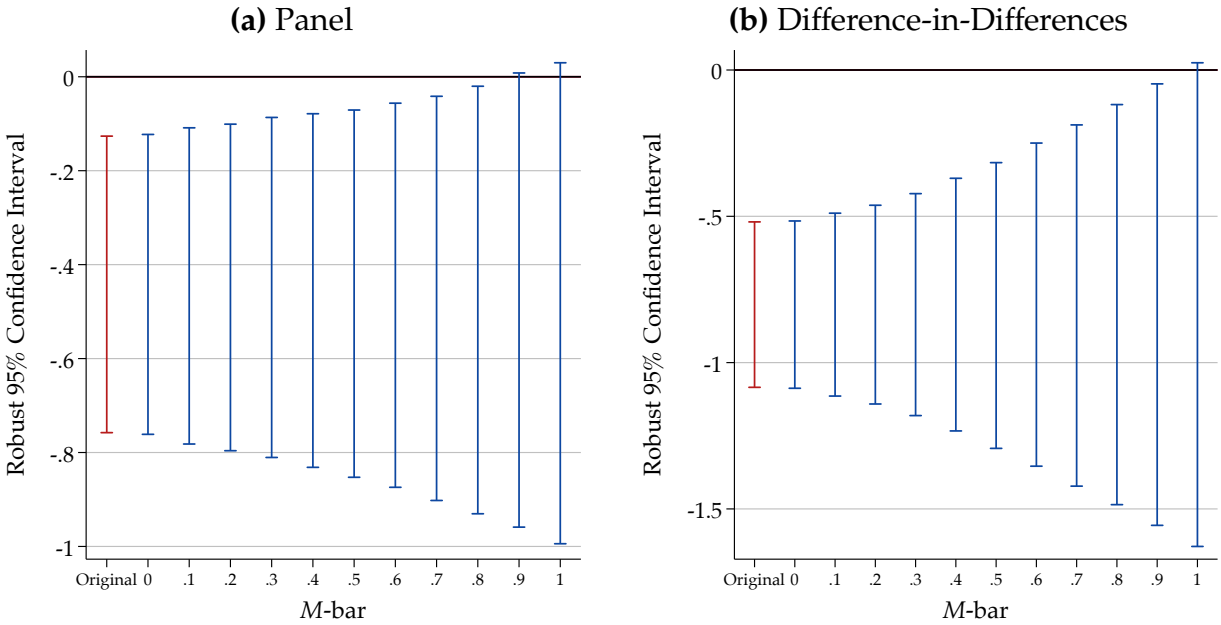
Notes. This figure reports cross-sectional estimates of specification (9) between emigration to diverse Italian communities in the United States and street-naming patterns in Italy. The unit of observation is a municipality. The dependent variables are the shares of streets whose eponyms reflect national identity (Panel F.17a) and local identity (Panel F.17b). The main explanatory variables are the number of emigrants to diverse neighborhoods (displayed by the hollow markers) and the number of emigrants to homogeneous neighborhoods (displayed by the colored markers). In each regression, we compute the standard errors using different estimators: unadjusted, White, clustered at different geographical layers, and Conley-corrected using various thresholds centered on the municipality's centroid. The outcomes are normalized by the total number of streets in the municipality. All specifications include region fixed effects. Bands report 95% confidence intervals. Referenced on page: 30.

Figure F.18. Alternative Standard Errors: Monuments in Italy



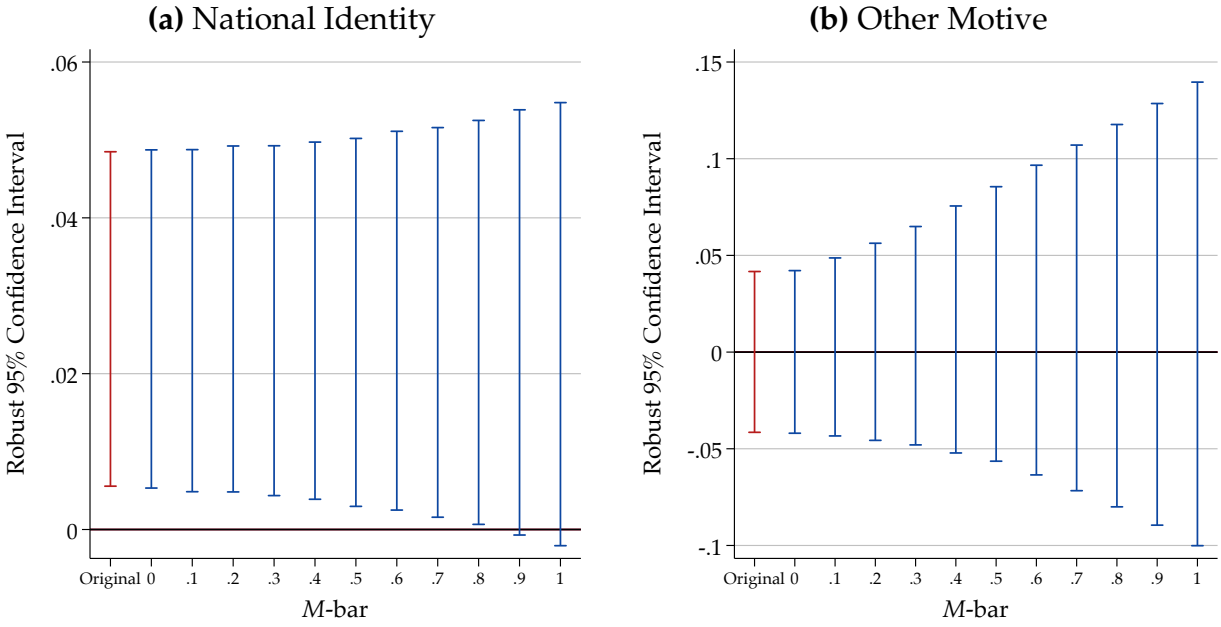
Notes. This figure reports estimates obtained from (10) of the effect of exposure to diversity within the Italian migrant community on monument construction in Southern Italy. The unit of observation is a municipality-decade for municipalities observed between 1861 and 2000. In Panel F.18a, the dependent variable is an indicator equal to one if the municipality has at least one national-identity monument in a given decade. In Panel F.18b, the dependent variable is an indicator equal to one if the municipality has at least one non-national-identity monument in a given decade. The baseline treatment is an interaction term between decade indicators and the share of U.S. emigrants from each municipality who settled in districts in the top quintile of the diversity measure defined in (3). In each regression, we compute the standard errors using different estimators: unadjusted, White, clustered at different geographical layers, and Conley-corrected using various thresholds centered on the municipality's centroid. Each specification includes municipality fixed effects and decade fixed effects. Bands report 95% confidence intervals. Referenced on page: 30.

Figure F.19. Robustness to Violations of Parallel Trends: Naming in Italy



Notes. This figure reports alternative confidence intervals for the effect of exposure to diversity within the Italian migrant community on name-based indices in Italian municipalities from equations (7)–(8). The bands report 95% robust confidence intervals constructed following the method developed by Rambachan and Roth (2023). The unit of observation is a municipality observed at a decennial frequency between 1880 and 1930. The dependent variable is the local name index defined in (1). In Panel F.19a, each dot reports the coefficient of a treatment term defined as the share of a municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). In Panel F.19b, each dot reports an interaction term between a post indicator equal to one in years at or after the municipality-specific onset of mass migration (defined as the first year with at least one emigrant per 1,000 inhabitants) and the share of the municipality's U.S. emigrants who settled in districts in the top quintile of the diversity distribution (3). All regressions include municipality and year fixed effects. Referenced on page: 31.

Figure F.20. Robustness to Violations of Parallel Trends: Naming in Italy



Notes. This figure reports alternative confidence intervals obtained from (10) on the effect of exposure to diversity within the Italian migrant community on monument construction in Southern Italy. The unit of observation is a municipality-decade for municipalities observed between 1861 and 2000. The bands report 95% robust confidence intervals constructed following the method developed by Rambachan and Roth (2023). In Panel F.20a, the dependent variable is an indicator equal to one if the municipality has at least one national-identity monument in a given decade. In Panel F.20b, the dependent variable is an indicator equal to one if the municipality has at least one non-national-identity monument in a given decade. The baseline treatment is an interaction term between decade indicators and the share of U.S. emigrants from each municipality who settled in districts in the top quintile of the diversity measure defined in (3). Referenced on page: 31.

APPENDIX REFERENCES

Abramitzky, Ran, Katherine Eriksson, Santiago Pérez, and Myera Rashid, "Census Linking Project: Version 3.0," *[dataset]*, 2025.

_ , Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez, "Automated Linking of Historical Data," *Journal of Economic Literature*, 2021, *59* (3), 865–918.

Basse, Guillaume, Peng Ding, Avi Feller, and Panos Toulis, "Randomization Tests for Peer Effects in Group Formation Experiments," *Econometrica*, 2024, *92* (2), 567–590.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique, "The Economic Payoff of Name Americanization," *Journal of Labor Economics*, 2017, *35* (4), 1089–1116.

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley EB Wilbert, "Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project," *Explorations in Economic History*, 2025, p. 101717.

Carnevale, Nancy C, *A New Language, a New World: Italian Immigrants in the United States, 1890-1945*, University of Illinois Press, 2010.

Chabod, Federico, *Storia della Politica Estera Italiana Dal 1870 Al 1896*, Laterza, 1965.

Chiaricati, Federico, "Nationalism and Nation-Building in the Dietary Consumption of Italian Migrants in the United States: A Transnational Perspective," *Modern Italy*, 2020, *25* (4), 403–419.

Choate, Mark I., *Emigrant Nation: The Making of Italy Abroad*, Harvard University Press, 2008.

Ciccarelli, Carlo and Jacob Weisdorf, "Pioneering into the Past: Regional Literacy Developments in Italy Before Italy," *European Review of Economic History*, 2019, *23* (3), 329–364.

Colella, Fabrizio, Rafael Lalive, Seyhun Orcan Sakalli, and Mathias Thoenig, "`acreg`: Arbitrary Correlation Regression," *The Stata Journal*, 2023, *23* (1), 119–147.

Coluccia, Davide M and Lorenzo Spadavecchia, "Emigration Restrictions and Economic Development: Evidence from the Italian Mass Migration to the United States," *Working Paper*, 2025.

Conley, Timothy G, "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Corradini, Camillo, "Relazione Presentata a S. E. Il Ministro della Pubblica Istruzione Dal Direttore Generale per L'istruzione Primaria E Popolare," in "L'istruzione primaria e popolare in Italia, con speciale riguardo all'anno scolastico 1907-1908, vol. I," Roma, 1910.

Corradini, Enrico, "L'emigrazione Italiana Nell'america del Sud (1909)," *Discorsi politici (1902–1923)*, 1923.

Covello, Leonard, *The Social Background of the Italian American School Child*, Leiden, 1967.

De Mauro, Tullio, *Storia Linguistica dell'italia Unita*, Gius. Laterza & Figli Spa, 1963.

Deschamps, Bénédicte, "De la Presse "Coloniale"' à la Presse Italo-Américaine, le Parcours de Six Périodiques Italiens aux Etats-Unis (1910-1935)." PhD dissertation, Université Paris 7 1996.

Dickie, John, *Darkest Italy: The Nation and Stereotypes of the Mezzogiorno, 1860–1900*, Springer, 1999.

Felice, Emanuele, "The Roots of a Dual Equilibrium: GDP, Productivity, and Structural Change in the Italian Regions in the Long Run (1871–2011)," *European Review of Economic History*, 2019, *23* (4), 499–528.

Gabaccia, Donna R, *Italy's Many Diasporas*, Routledge, 2013.

Gilmour, David, *The Pursuit of Italy: A History of a Land, Its Regions, and Their Peoples*, Macmillan, 2011.

Luconi, Stefano, "Forging an Ethnic Identity: The Case of Italian Americans," *Revue française d'études américaines*, 2003, *96* (2), 89–101.

Molfese, Franco, *Storia del Brigantaggio Dopo L'unitÁ*, Feltrinelli, 1964.

Orsi, Robert A, *The Madonna of 115th Street: Faith and Community in Italian Harlem, 1880-1950*, Yale University Press, 1985.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley, "Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set," *Explorations in Economic History*, 2021, *80*, 101391.

Rambachan, Ashesh and Jonathan Roth, "A More Credible Approach to Parallel Trends," *Review of Economic Studies*, 2023, *90* (5), 2555–2591.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-

Networks," *arXiv preprint arXiv:1908.10084*, 2019.

Scalabrini, Giovanni Battista, "Seconda Conferenza Sull'emigrazione," in "Gli Italiani all'estero," Tipografia Roux-Frassati e C., 1898.

_ , "L'emigrazione Italiana in America. Osservazioni," in Silvano Tomasi and Gianfausto Rosoli, eds., *Scalabrini e le migrazioni moderne*, Società Editrice Internazionale, 1983.

Smith, Denis Mack, *Cavour and Garibaldi 1860: A Study in Political Conflict*, Cambridge University Press, 1985.

Soffiati, Daniele and Alberto Grandi, *La Cucina Italiana Non Esiste*, Mondadori, 2024.

Vecoli, Rudolph J, "The Formation of Chicago's "little Italies"," *Journal of American Ethnic History*, 1983, *2* (2), 5–20.

_ , "The Search for an Italian American Identity: Continuity and Change," in Lydio T Tomasi, ed., *Italian Americans: New Perspectives in Itahan Immigration and Ethnicity*, Center for Migration Studies, 1985.

Wyman, Mark, *Round-Trip to America: The Immigrants Return to Europe, 1880-1930*, Cornell University Press, 1993.