Unity in Diversity?
How Intergroup Contact Can Foster Nation Building[*]

Samuel Bazzi[†]
Boston University
NBER and CEPR

Arya Gaduh[‡]
University of Arkansas

Alexander D. Rothenberg[§]
Syracuse University

Maisy Wong[¶]
University of Pennsylvania

March 2019

Abstract

We use a population resettlement program in Indonesia to identify long-run effects of intergroup contact on national integration. In the 1980s, the government relocated two million ethnically diverse migrants into hundreds of new communities. We find greater integration in *fractionalized* communities with many small groups, as measured by national language use at home, intermarriage, and children's name choices. However, in *polarized* communities with a few large groups, ethnic attachment increases and integration declines. Residential segregation dampens these effects. Social capital, public goods, and ethnic conflict follow similar patterns. Overall, our findings highlight the importance of localized contact in shaping identity.

JEL Classifications: D02, D71, J15, O15, R23

Keywords: Culture, Diversity, Identity, Language, Migration, Nation Building

[*] We thank the following for helpful feedback: Alberto Alesina, Oriana Bandiera, Toman Barsbai, Giorgio Chiovelli, Raquel Fernandez, Paola Giuliano, Dilip Mookherjee, Nathan Nunn, Daniele Paserman, Ben Olken, Imran Rasul, Duncan Thomas, and audiences at Boston University, Duke, Georgetown, Harvard, the Kiel Institute, McGill, MIT, Notre Dame, Syracuse, University of Arkansas, University of Colorado Denver, UNC-Chapel Hill, University of Southern California, University of Toronto, Wellesley College, the Barcelona Graduate School Summer Forum, the 2017 DIAL Development Conference, the 2017 Midwest International Development Conference, the NBER Political Economy meeting, the Ninth International Migration and Development Conference, the 2016 Annual European Conference of ASREC, the Yale Interdisciplinary Migration Conference, and the Northeast Universities Development Consortium 2016 Conference. Bazzi is grateful for financial support from the Global Development Policy Center at Boston University. Wong is grateful for financial support from the Research Sponsors Program of the Zell/Lurie Real Estate Center and the Wharton Social Impact program. A previous version of this paper circulated under the title "Unity in Diversity? Ethnicity, Migration, and Nation Building in Indonesia." Richard Jin, Gedeon Lim, JoonYup Park, Ying Pei, Xuequan Peng, and Hanna Schwank provided excellent research assistance. All errors remain ours.

[†] Department of Economics. 270 Bay State Rd., Boston, MA 02215. Email: sbazzi@bu.edu.

[‡] Sam M. Walton College of Business. Department of Economics. Business Building 402, Fayetteville, AR 72701-1201. Email: agaduh@walton.uark.edu.

[§] Syracuse University. 426 Eggers Hall. Syracuse, NY 13244-1020. Email: adrothen@maxwell.syr.edu.

[¶] Wharton Real Estate. 3620 Locust Walk, 1464 SHDH, Philadelphia, PA 19104-6302. Email: maisy@wharton.upenn.edu.

[The] central challenge of modern, diversifying societies is to create a new, broader sense of 'we'.
—Robert Putnam, *The 2006 Johan Skytte Prize Lecture*

1 Introduction

Uniting people from diverse cultures is a founding principle of many nation states.[1] Throughout history, leaders have introduced policies to foster a national identity that would sustain an "imagined political community" in which citizens remain connected by shared history and values, despite never meeting one another (Anderson, 1983). However, rising geographic mobility has stoked concerns that growing local diversity may undermine this nation-building objective.[2]

This paper asks how intergroup contact affects the intergenerational process of nation building. Competing views abound in the social sciences. Some argue that exposure to new cultures provokes backlash and conflict (Blumer, 1958; Huntington, 2004). Others posit that, under certain conditions, negative sentiments may dissipate as intergroup relationships develop over time with greater contact (Allport, 1954). Alternatively, diversity may engender social anomie or isolation, which limits integration (Putnam, 2007). Whether intergroup contact is conducive to integration or to conflict may also depend on the relative size of different groups (Esteban and Ray, 2008, 2011). Empirically, it is difficult to distinguish a causal effect of diversity from the influence of endogenous sorting and location-specific amenities.

We address these challenges using a large-scale policy experiment that created hundreds of diverse communities across Indonesia. One of history's largest resettlement efforts, the Transmigration program provides a unique opportunity to understand how intergroup contact affects nation building. After independence, the government urgently needed to forge an Indonesian identity that would forestall secessionist tendencies. Policymakers viewed resettlement as an important part of efforts to unite more than 700 ethnolinguistic groups across the archipelago. From 1979 to 1988, the government relocated two million voluntary migrants (hereafter, transmigrants) from the Inner Islands of Java and Bali to newly created agricultural villages in the Outer Islands.[3]

The Transmigration program offers plausibly exogenous variation in ethnic diversity and segregation. Given the rapid scale-up and haphazard implementation, planners had little ability to systematically assign transmigrants. Nor could transmigrants choose their destinations. They queued for a short time at Inner-Island transit camps waiting for settlements to open in the Outer Islands. The coincidental timing determined the ethnic mix of Inner Islanders in the new villages. The government further mandated quotas for Outer-Island natives that varied over time and across regions. Reassuringly, we find that initially assigned diversity does not systematically differ in more hospitable locations. Moreover, upon arrival, all settlers received houses and farms by lottery. Full ownership rights were transferred after 5–10 years, and imperfect land markets effectively tied migrants to their initial plot.

[1] For example, "United in Diversity" is the motto in the European Union, *E pluribus unum* in the United States, and "Unity in Diversity" in South Africa. History abounds with efforts "to make" national citizens (see, e.g., Duggan, 2007, on Italy).

[2] See Putnam (2007). Alesina et al. (2017) and Miller (2012) discuss challenges of forging a shared identity within the European Union. More generally, migration pressures are growing among minorities within rich countries (see Frey, 2014, on the United States) and in newer migration corridors from poor to rich countries (Hanson and McIntosh, 2016).

[3] The program had three goals: population redistribution, agricultural development, and nation building. Bazzi et al. (2016) investigate the agricultural productivity effects. While unique in some respects, Transmigration has parallels with resettlement schemes in other developing countries and also with state-sponsored efforts to settle frontier areas in developed countries. Appendix C elaborates on these as well as possible implications for modern refugee resettlement (see Bansak et al., 2018).

We use the complete-count 2010 Census Population data to study diversity and the nation-building process. The data comprise more than two million individuals in 817 Transmigration villages. We observe self-reported ethnicity and three revealed-preference measures of identity and integration: language use at home, intermarriage, and children's name choices. The data also provide granular details on residential location, allowing us to examine hyper-local intergroup contact. We combine the Census with survey and administrative data sources capturing other integration outcomes.

Strikingly, even after three decades, Transmigration villages exhibit significantly greater ethnic diversity and less within-village segregation than other organically settled villages in the Outer Islands. The persistence of mixed communities suggests that tipping (à la Schelling, 1971) did not neutralize the initial policy assignment. Long-run diversity in these villages is unrelated to predetermined amenities associated with national integration (e.g., proximity to roads). Comparing across Transmigration villages, we can thus identify the effects of intergroup contact that are not due to endogenous sorting.

We develop a model of identity choice to understand how local diversity influences the nation building process. Individuals choose whether to retain their ethnic identity or to adopt a national identity (as revealed by language choice). Interactions in a diverse community benefit from a common culture. With many small groups, a neutral national identity can help solve coordination problems and maximize the gains from market and non-market interactions (Lazear, 1999). However, with a few large groups, intergroup antagonism grows in importance (Esteban and Ray, 1994). Diversity may accelerate or slow down the diffusion of the national identity. We embed these insights in a framework that generalizes the Darity Jr. et al. (2006) model on the evolution of identity.

The model predicts how a community's initial ethnic composition determines the long-run prevalence of the national identity. Under certain assumptions, we derive a closed-form expression that includes two widely-used measures of diversity: fractionalization (F) and polarization (P). In high-F villages (with many small groups), the national identity is more pervasive given the benefits of coordination. In high-P villages (with a few large groups), ethnic attachment is more likely as it provides protection from intergroup antagonism. Both of these forces are more muted in segregated communities where, holding F and P fixed, intergroup contact is more limited.

We test these novel implications using several proxies for national integration. Our primary measure is the choice of language used at home in 2010. Globally, language is seen as one of the most critical components of national identity (Pew Research Center, 2017). Policymakers view the national language, *Bahasa Indonesia* or Indonesian, as synonymous with the Indonesian identity, widely promoting its use across economic and social domains. Indonesian is rooted in a minority ethnic language (Malay) spoken by only 5 percent of the country when it was chosen as the national language in 1928. Today, nearly everyone can speak Indonesian. Yet, less than 20 percent choose it as the main language at home; most prefer their native ethnic language. In survey data, those speaking Indonesian at home (*homeIndo*) report significantly stronger national than ethnic identity. This suggests that *homeIndo* may contribute to Indonesian identity formation and advance nation building. Using auxiliary panel data, we describe this intergenerational process, linking *homeIndo* as a child to weaker ethnic attachment later in life as adults.

Our main results support the opposing effects of fractionalization and polarization suggested by the model. A one standard deviation (s.d.) increase in fractionalization leads to 12.9 percentage points (p.p.) greater *homeIndo*, consistent with the benefits of intergroup contact in settings with many small groups. A one s.d. increase of P leads to 8 p.p. lower *homeIndo*, consistent with the costs of intergroup antagonism in settings with a few large groups. These are large effects relative to the village-level mean of 14.4 percent for *homeIndo*.

Several additional findings point to an identity-based interpretation. If individuals speak Indonesian at home solely to improve language skills, we would see different effects across education levels or employment sectors, but we do not. It is particularly telling to find sizable effects of F and P on ethnic Malays whose native language forms the base of Indonesian. For Malays to choose *homeIndo* rather than their mother tongue, they must feel relatively more invested in the national identity. In fact, we find stable effects of F and P across major ethnic groups. Together, these results suggest that *homeIndo* likely captures something deeper than latent fluency or effort to improve skills thereof. Moreover, these findings are not likely due to endogenous sorting. We address compositional differences through an array of fixed effects (e.g., ethnicity, birthplace, age) and show that the small subset of residents that may have sorted endogenously cannot overturn our findings.

Importantly, we identify similar effects of diversity on two additional proxies for ethnic attachment. First, interethnic marriage rates, a leading indicator of integration (Gordon, 1964), are positively related to F and negatively related to P. Second, we study the identity content of names given to children born after resettlement. Name choices are the first act of intergenerational cultural transmission, conveying information about parental preferences and expectations about the value of different identities. Using several indices akin to the "black name index" of Fryer and Levitt (2004), we find that parents give their children less ethnically distinctive names in villages with greater F and lower P.

Furthermore, polarization exhibits adverse effects on social capital. At the individual level, we use survey data to measure intergroup tolerance and trust, community engagement, and preferences for redistribution. These subjective responses line up with village-level outcomes: P reduces growth-enhancing public goods provision by local governments, increases the likelihood of ethnic conflict, and ultimately hinders economics development. Meanwhile, F works in the opposite direction, indicating possible downstream benefits of integration. Together, these other outcomes bolster our revealed preference interpretation of *homeIndo* as reflecting broader investments in the national identity, weakening ethnic attachments, and integration with other groups.

To better understand why ethnic divisions matter, we identify three potentially important mechanisms, focusing on our core outcome of *homeIndo*. First and foremost, residential segregation determines the scope for intergroup contact to change behavior in diverse communities. Exploiting the lottery assignment of housing units, we identify granular effects of diversity, both at the level of neighborhoods within villages and among immediate next-door neighbors, reminiscent of the neighborhood effects in Bayer et al. (2008) and Chetty and Hendren (2018). Moreover, segregation dampens the effects of both F and P by limiting day-to-day contact with other groups. Second, as in Lowe (2018), the type of contact matters: F has weaker positive effects in settings with greater interethnic inequality in economic resources, proxied by location-specific human capital endowments. Third, deep-rooted linguistic differences between ethnic groups amplify both the benefits of F and the costs of P. In a final exercise, we show how coordination on *homeIndo* can reduce the effective degree of polarization by bringing otherwise culturally distant groups closer together.

This paper sheds light on how intergroup contact influences nation building. Many studies document adverse consequences of diversity (see Alesina and LaFerrara, 2005 and Esteban and Ray, 2017 for reviews). Relatively few examine how to mitigate ethnic divisions through nation building (Alesina and Reich, 2015; Miguel, 2004).[4] A survey by Paluck et al. (2018) notes that prior work on intergroup contact has tended to focus on short-run lab or field experiments and self-reported preferences. Within economics, a few studies find that contact fosters short-run increases in tolerance and out-group friendships (Boisjoly et al., 2006; Lowe, 2018; Rao, forthcoming).

We make three contributions to these literatures. First, we use a large-scale policy to examine long-run effects of intergroup contact on both self-reported preferences and behavioral measures of integration. Nation building is a slow process, and endogenous sorting makes it difficult to identify these effects in most settings. Second, our model complements the Esteban and Ray (2011) theory relating F and P to conflict. We show that F hastens and P hinders the diffusion of the national identity. There are positive externalities to adopting a common national identity, and as intergroup contact speeds up this process (through F), there may be increasingly less scope for intergroup antagonism to fuel conflict (through P). Third, our findings on segregation contribute to a small but growing literature highlighting the important role of physical proximity in mediating the aggregate effects of diversity.[5] Algan et al. (2016) also explore sharply local effects of diversity in public apartment blocks in France, arguing that diverse buildings tend to foster social anomie. We differ by focusing on identity and integration, by disentangling F and P, and by clarifying the distinct effect of segregation.

Our study offers insight on how diversity affects the formation of a new shared identity. This process differs from minority immigrant assimilation to the native majority explored in prior work (e.g., Abramitzky et al., 2018; Advani and Reich, 2015; Bleakley and Chin, 2010). Our findings suggest an important role for the national language. This novel focus matters for understanding nation-building processes in historical Europe as well as post-colonial developing countries.[6] It also provides a window into contemporary debates about national identity in rapidly diversifying developed countries. Convergence towards a "broader sense of 'we"' may be easier in some settings (high F) than others (high P). We are among the first to bring these two dimensions of diversity into a single framework for studying integration, which may be useful in other settings.

The paper proceeds in seven sections. Section 2 provides background on nation building in Indonesia and the Transmigration program. Section 3 develops a model for understanding how intergroup contact affects national integration. Section 4 describes our main data sources. Section 5 develops our empirical strategy, including details on the transmigrant allocation process. Section 6 presents our core empirical results. Section 7 explores mechanisms and other outcomes. Finally, Section 8 revisits the controversial legacy of the Transmigration program and offers concluding thoughts.

[4] Other recent work on nation building examines how public media (Blouin and Mukand, 2016), bureaucrat assignments (Okunogbe, 2015), schooling (Bandiera et al., forthcoming), shared religious experience (Clingingsmith et al., 2009), and external enemies (Dell and Querubin, forthcoming) influence intergroup tolerance and national identity.

[5] This interplay between local and aggregate diversity features in cross-country studies by Alesina and Zhuravskaya (2011) and Desmet et al. (2016) and is an emerging theme in the political science literature covered by Enos (2017).

[6] There is comparatively little empirical work on either setting. There are interesting case studies on France (Weber, 1976) and several African countries (Laitin, 2007), and various national language policies are discussed in books referenced in footnote 8. There are a few empirical studies looking at the effects of banning ethnic languages (Clots-Figueras and Masella, 2013; Fouka, 2016) and the determinants of national language choice by the government (Laitin and Ramachandran, 2015; Liu, 2015). Yet, a recent survey of the economics literature on language reveals no work on the national language and its implications for nation building in diverse countries (Ginsburgh and Weber, 2018). This is precisely where our study innovates, and our context should be of broad interest given Indonesia's remarkable diversity and relative success in promoting a national identity.

2 Background: Diversity, Language, and Nation Building in Indonesia

Indonesia is one of the world's largest and most diverse countries, with more than 1,200 self-identified ethnic groups living on roughly 6,000 islands. According to 2010 Population Census data, Indonesia has an ethnic fractionalization index F—the probability that any two residents belong to different ethnicities—of around 0.81. Despite its national diversity, most Indonesians live in segregated communities: of more than 60,000 urban and rural villages, the median village has an F of 0.04.[7]

For most of its history, several independent kingdoms governed the peoples of the Indonesian archipelago. Absent a common ruler, many different cultures and languages persisted throughout the region. The Dutch colonists pursued a divide-and-conquer strategy that pitted kingdoms against each other, ensuring that by the end of the nineteenth century "... a common Indonesian identity or [set] of common goals simply did not yet exist" (Ricklefs, 2008, p.189). After independence in 1945, many in the Outer Islands saw the consolidation of power as favoring the Javanese, Indonesia's largest ethnic group with 40 percent of the population (Bertrand, 2004). This fueled anti-Javanese sentiments and recurring secessionist threats from the Outer Islands (Thornton, 1972).

Not surprisingly, given Indonesia's vast diversity and disparate groups with little shared history, its founding leaders prioritized national unity. Anderson (1983, pp. 6–7) defines a nation as "an imagined political community" whose members are often strangers but think of each other as part of a "communion." To build a nation is to promote a shared national identity, with shared values and preferences that are strong enough to glue its citizens together (Alesina and Reich, 2015). In Indonesia, "national unity" became one of the state ideology's Five Key Principles (*Pancasila*), and "*Bhinneka Tunggal Ika*" (Unity in Diversity) is the state motto. The national language and the Transmigration program were two central policies, among many, designed to advance this objective.

2.1 National Language

Policymakers viewed the national language as a key vehicle to socialize Indonesia's national identity.[8] In 1928, nationalists at the Second All-Indonesian Youth Congress drafted a statement of unity opposing the Dutch. They pledged allegiance to Indonesia as "*satu nusa, satu bangsa, satu bahasa*" (one fatherland, one nation, one language). They aimed to create a nation "unified by ties of common language, common outlook, and common political participation, a people enthusiastically severing its outworn ties to local traditions and loyalties" and instead rooted in an "all-Indonesian culture" (Feith, 1962, pp. 34-35).

The national language, *Bahasa Indonesia* (or Indonesian), is a modified version of Malay, a trading language used in the region for centuries. Before its recognition as the national language in 1928, Malay was the mother tongue of 5 percent of Indonesia's colonial population. Choosing a lingua franca instead of Javanese, the language of the largest ethnic group, was critical. According to Liu (2015, pp. 4-5), this choice played "a decisive role in counteracting the potentially negative economic effects of the country's heterogeneity", minimizing intergroup tensions and cultivating the image of a pan-ethnic state.

[7] Villages (*desa* or *kelurahan*) comprise the lowest level of governance in Indonesia with an average population of over 2,000 (7,000) in rural (urban) areas in the early 2000s. They are the main administrative unit of analysis in our study.

[8] This view is pervasive globally. The role of language policy in shaping national identity is a key theme of several books covering countries across Europe (Barbour and Carmichael, 2000), Asia (Simpson, 2007), and Africa (Simpson, 2008).

Subsequent policies, including requiring its use in schools and official communications, promoted Indonesian as "a symbol of national unity and identification" (Sneddon, 2003). Policymakers leading this effort, like Alisjahbana (1962), believed that as people "... learned to express themselves in Indonesian, the more conscious they became of the ties which linked them." Many outsiders view Indonesia's language policy as exemplary.[9] According to the 2010 Census, nearly all Indonesians are able to speak the national language. Yet, less than 20 percent use it as their main language at home.

2.2 Transmigration

Some policymakers also saw the nation-building potential of resettlement. The Transmigration program aimed to relieve population pressures in Java/Bali and stimulate development in the Outer Islands. Policymakers believed that the program could also foster national integration by expanding the possibilities for intergroup contact. For instance, in 1985, the Minister of Transmigration stated "By way of transmigration, we will try to ... integrate all the ethnic groups into one nation, the Indonesian nation. The different ethnic groups will in the long run disappear because of integration and there will be one kind of man, Indonesian" (Hoey, 2003).

Transmigrants volunteered for the program. Only nuclear families were eligible, and couples had to be legally married, with the household head between 20 and 40 years old. In practice, most participants were poor, landless agricultural laborers, with few assets, and limited schooling (see Kebschull, 1986, for a pre-departure survey of transmigrants). Their education levels are more comparable to rural non-migrants than to voluntary migrants from their home districts.[10]

The Transmigration program provided free transport to the newly created settlements, housing, two-hectare farm plots, and supplies for the first few growing seasons. According to the 1978 Transmigration Manual, planners were keen to ensure that each settlement could produce enough food to overcome subsistence. Officials worked with agricultural experts to map arable land availability, elevation, vegetation, soil types, hydrology, climate, and market access (see Bazzi et al., 2016). They used these measures to determine the maximum potential population of each settlement.

The National Ministry of Transmigration (MOT) created and oversaw the new villages, endowing each with the same initial institutions. The MOT provided public goods, including health clinics and schools, where children were taught in Indonesian and would mix with students from different backgrounds. Moreover, upon arrival, farm plots and housing were assigned by lottery to newly-arriving settlers. This served, *ex ante*, to limit residential segregation and inequality in land (quality) across ethnic groups. After 5–10 years, households received ownership of housing and land, formerly under MOT authority, though this was not perfectly enforced in practice. As formerly landless, the delayed prospect of ownership may have tied them to their new land in the critical first few years after resettlement.

The program fostered ethnic diversity in the new villages in two ways. First, given the rushed implementation to meet lofty resettlement targets,[11] the assignment of transmigrants was neither rigorous nor

[9] According to Paauw (2009), "[No] other post-colonial nation has been able to develop and implement a national language with the speed and degree of acceptance which Indonesia has. No other national language ... is used in as wide a range of domains as Indonesian, a feat made more impressive by the size and ethnic, linguistic and cultural diversity of Indonesia".

[10] In the 2000 Population Census, transmigrants had around 0.7 fewer years of schooling than non-migrants from their origin district in Java/Bali and 3.5 fewer years of schooling than those who migrated independently to other Outer-Island districts.

[11] A total of 1.2 million people were resettled from 1979–1983, and an additional 3.75 million people were planned to be resettled

systematic. Instead, the allocation across villages was determined by the coincidental timing of transmigrants' arrival to transit camps in Java/Bali and the opening of settlements in the Outer Islands. If the transmigrants queuing in camp C happened to be diverse at the time village V was cleared, then V received a diverse mix of Inner-Island settlers. Second, to encourage contact between Inner and Outer Islanders, planners allocated quotas for native Outer Islanders in each settlement. These quotas, known as *Alokasi Pemukiman bagi Penduduk Daerah Transmigrasi* (APPDT), varied across time and space. Together, the haphazard assignment of transmigrants and the APPDT quotas induced variation both in the ethnic mix among Inner Islanders and the relative shares of Inner and Outer Islanders. We elaborate on both sources of variation when developing our empirical strategy in Section 5.1.

3 Model: Intergroup Contact and Identity Choice

This section presents a framework for understanding how intergroup contact influences the nation-building process. Building on Darity Jr. et al. (2006), we model the choice between maintaining one's own ethnic identity or adopting the common national identity. There are tradeoffs between the benefits of productive intergroup relationships (Lazear, 1999) and the costs of intergroup antagonism (Esteban and Ray, 1994). Coordinating on a national identity increases the returns to social interactions across groups, and when someone adopts this common identity, it spurs others to do the same. However, intergroup antagonism can hinder this process, particularly when large groups assert cultural dominance.

Our model uses an evolutionary game theory framework to study how contact between members of different ethnic groups slowly transforms identity choices for the community as a whole. We show how, under certain conditions, ethnic fractionalization (many small groups) hastens nation building while polarization (a few large groups) hinders it. The model is stylized in order to develop intuition for our empirical results. We address extensions in Appendix B.

3.1 Setup

Consider a community with multiple ethnic groups, indexed by $j = 1, ..., J$. Each individual is endowed with a fixed ethnicity, exogenously given at birth. For simplicity, we assume that individuals live forever, and, over the course of their lives, they decide whether to retain their own *ethnic identity* or to adopt the *national identity*. We assume infinite lives for simplicity; similar results would hold with finitely-lived individuals who transmit their identity to the next generation (see Montgomery, 2010). As a baseline, we model contact as a random matching process: the probability an individual is matched to someone from ethnic group j is equal to that ethnic group's population share, p_j. In Appendix B, we allow matching to be influenced by segregation between ethnic groups within the village. By limiting intergroup contact, segregation dampens the effects of diversity that we derive below.

Identity choices are persistent. Individuals match each period but are only able to revise their choices infrequently. More precisely, each individual must maintain their identity choice for T periods, where T is an independent draw from an exponential distribution with rate R. When revision opportunities arise, an individual is myopic and compares her current payoff with that from a random sample of strategies

from 1984–1988. A large and unexpected drop in oil revenue in the mid-1980s led to a significant shortfall in meeting the planned targets during this latter period (see Bazzi et al., 2016).

played by those around her.[12] She adopts the strategy with the higher payoff. The probability that she switches her identity is proportional to the difference in payoffs. This infrequent process of identity switching leads to inertia and makes convergence to an evolutionarily stable equilibrium relatively slow.

Payoff Structure. Table 1 shows how the payoff to group j from each match depends on identity choices (described in the last two columns) and the types of interactions (rows). There are three key parameters governing payoffs from interactions: (i) θ, which captures the market and non-market benefits from productive interactions; (ii) γ, which captures the costs of investing in an ethnic identity (γ^E) or a national identity ($\gamma^N > \gamma^E$), including the costs of learning a language or the costs of maintaining cultural traditions; (iii) and D, which captures disutility from intergroup antagonism.

Table 1: Payoffs of Identity Choices for Group j

		Identity Choices	
	Matched with	National	Ethnic
Own-group	National j	$\theta - \gamma^N$	$\theta - \gamma^E$
	Ethnic j	$\theta - \gamma^N$	$\theta - \gamma^E$
Intergroup	National k	$\theta - \gamma^N$	$-\gamma^E$
	Ethnic k	$-D_k^N - \gamma^N$	$-D_k^E - \gamma^E$

An individual adopting the national identity (an N-chooser) obtains a payoff of $\theta - \gamma^N$ from own-group interactions (the top two rows). For intergroup contact with a fellow N-chooser, the payoff is also $\theta - \gamma^N$ since they share a common national identity. However, intergroup contact with E-choosers gives rise to antagonism and a lower payoff of $-D_k^N - \gamma^N$.

For own-group interactions, the payoff from choosing the ethnic identity (E-chooser) is $\theta - \gamma^E$. However, for intergroup contact with an N-chooser, the payoff is only $-\gamma^E$: there is no benefit unless they share a common national identity. The payoff from intergroup contact with an E-chooser is $-D_k^E - \gamma^E$.

Individuals can only choose one identity, and relative group sizes (p's) affect the likelihood of own-group and intergroup interactions. Individuals enjoy benefits (θ) from sharing a common national identity. By contrast, remaining an E-chooser confers benefits from own-group interactions and protection from intergroup antagonism. We assume intergroup antagonism is costlier for N-choosers than E-choosers who enjoy protection from their own ethnic network ($D_k^E < D_k^N$). This is akin to a club-good benefit for ethnic loyalists that is excludable from others, including N-choosers from the ethnic group (see Cornes and Sandler, 1996; Iannaccone, 1992).

Expected Payoffs. Let w_j^s denote the expected payoff for group j from choosing the National ($s = N$) or the Ethnic ($s = E$) identity. Given the matching process, average payoffs can be written as a function of individual match payoffs, exogenous ethnic shares (p_j), and the endogenous share of N-choosers (π_j):

$$\text{Nationalist:}\quad w_j^N = p_j\Big[\pi_j\left(\theta - \gamma_N\right) + (1 - \pi_j)\left(\theta - \gamma_N\right)\Big] + \sum_{k \neq j} p_k\Big[\pi_k\left(\theta - \gamma_N\right) + (1 - \pi_k)\left(-\gamma_N - D_k^N\right)\Big]$$

[12] This revision protocol is based on a textbook formulation widely used in evolutionary game theory models (Sandholm, 2010). Formally, $T \sim \exp(R)$, so that $\mathbb{P}(T \leq t) = 1 - e^{-Rt}$. This means that the number of identity revisions that can occur during the time interval $[0, t]$ follows a Poisson distribution, with mean Rt. Sandholm (2015) details the interaction and updating process, which is akin to the imitation mechanism put forward in Young (2015). See Appendix B.2 for further details.

$$= \underbrace{\theta\left(p_j + \sum_{k \neq j} p_k \pi_k\right)}_{(i)} - \underbrace{\gamma_N}_{(ii)} - \underbrace{\sum_{k \neq j}(1-\pi_k)p_k D_k^N}_{(iii)} \tag{1}$$

$$\text{Ethnic loyal:} \quad w_j^E = \theta p_j - \gamma_E - \sum_{k \neq j}(1-\pi_k)p_k D_k^E. \tag{2}$$

For example, for strategy N, the first term in brackets corresponds to payoffs for own-group interactions, and the second term corresponds to payoffs from intergroup interactions. The latter imply social externalities to identity choices (i.e., π_k influences choices of group j).

The payoffs in equation (1) depend on three factors: *(i)* the gains from productive interactions, *(ii)* the cost of adopting the identity, and *(iii)* the cost of intergroup antagonism. Intuitively, small ethnic groups enjoy greater benefits from coordinating on a common national identity. Larger ethnic groups may not benefit as much. Instead, they may prefer to remain ethnic loyalists given the greater costs of adopting a national identity ($\gamma^N > \gamma^E$) and the club-good benefit of protection against intergroup antagonism ($D_k^E < D_k^N$). In Appendix Figure A.1, we use data from Transmigration villages (described below) to relate ethnic shares (p_j) to language choices. Indeed, small groups who benefit from coordination are more likely to speak the national language (left panel), and large groups are more likely to speak their own native ethnic language (right panel).

3.2 Diversity and Growth of the National Identity

The model reveals how intergroup contact can influence the evolution of nation building, captured by the growth rate of adoption of the national identity. We characterize this growth process here.

Nation-Building Process. We define the aggregate growth rate in the adoption of the national identity, $\dot{G}^N$, as the community's population average of ethnic-group-specific growth:

$$\dot{G}^N = \sum_j p_j \dot{g}_j^N = \sum_j p_j \frac{d\pi_j}{dt}. \tag{3}$$

In Appendix B.2, we describe how the revision protocol and the matching process lead to a so-called replicator dynamic, which characterizes group-level identity growth. Intuitively, the strategy with the higher expected payoff propagates faster, and the dominated strategy is progressively eliminated. The growth of π_j is given by

$$\dot{g}_j^N = \frac{d\pi_j}{dt} = \pi_j\left(w_j^N - w_j\right), \tag{4}$$

where $w_j = \pi_j w_j^N + (1-\pi_j)\, w_j^E$ measures group j's average payoff across both choices. Note that w_j^N and w_j depend on π_k. As more people from group k choose N, π_k increases, raising w_j^N above w_j, encouraging further adoption of N next period. These social externalities accelerate the diffusion of the national identity. Equation (3) aggregates this growth equation to the community level to understand how initial ethnic diversity affects the rate of diffusion. We focus on two widely-used measures of diversity to summarize ethnic-group size distributions: *fractionalization* (F) and *polarization* (P).

Measuring Diversity: Fractionalization and Polarization. Fractionalization in a village corresponds to

$F = 1 - \sum_{j=1}^{J} p_j^2$. This measures the probability that two individuals, randomly selected from the village population, belong to different groups. With many small groups, F increases. Following Montalvo and Reynal-Querol (2005), we define polarization as $P = 4\sum_{j=1}^{J} p_j^2 (1 - p_j)$. P is maximized when a village's ethnic group shares approach a symmetric bimodal distribution (i.e., two groups equal in size).[14]

Aggregate Growth of the National Identity. Our model reveals how F and P shape identity choices. To begin, we show in Appendix B.3 that the growth in the national identity can be rewritten as follows:

$$\dot{G}^N = \sum_j p_j \dot{g}_j^N = \Phi\theta\Big(1 - \sum_j \phi_j p_j^2\Big) - \sum_j \sum_{k \neq j} p_j p_k T_{jk} - \bar{A}\gamma, \tag{5}$$

where $A_j = \pi_j(1-\pi_j)$, $\bar{A} = \sum_j p_j A_j$, $\overline{\pi} = \sum_k p_k \pi_k$ is the (ethnic-share-)weighted average of π_j's, $\Phi = \bar{A}\bar{\pi}$, $\phi_j = (A_j \pi_j)/\Phi$, and $T_{jk} = A_j(1 - \pi_k)D_k$. The first term in parentheses captures the notion that high-F communities with many small groups encourage the adoption of the national identity. The overall benefit (θ) from coordinating on a common national identity is larger with many small groups since $(1 - \sum_j \phi_j p_j^2)$ decreases with p_j. Meanwhile, the second term captures effective interethnic antagonism at the community level. If T_{jk} is a function of p_k (through the antagonism cost D_k), this term is akin to the total polarization formula first introduced in Esteban and Ray (1994, equation 1). The effects of relative group sizes (p's) in equation (5) are consistent with the patterns in Appendix Figure A.1.

With two simplifying assumptions, we can derive a closed-form relationship showing that F (P) increases (decreases) the rate of national identity adoption. First, we assume that intergroup antagonism, D_k, is a linear function of group shares: $D_k = 4\psi p_k$ for all $k = 1, ..., J$. This is consistent with larger groups asserting cultural dominance. Additionally, if $\pi_j = \pi$ for all $j = 1, ..., J$,[15] we show in Appendix B.3 that equation (5) simplifies to:

$$\begin{aligned}\dot{G}^N &= \pi\,(1-\pi)\left\{\theta\pi\left(1 - \sum_{j=1}^{J} p_j^2\right) - \psi\,(1-\pi)\left[4\sum_{j=1}^{J} p_j^2(1-p_j)\right] - \gamma\right\} \\ &= \beta_0 + \beta_1 F - \beta_2 P \end{aligned} \tag{6}$$

where $\beta_0 = -\pi\,(1-\pi)\,\gamma < 0$, $\beta_1 = \theta\pi^2\,(1-\pi) > 0$, and $\beta_2 = \psi\pi\,(1-\pi)^2 > 0$. In the case of matching under segregation, the expression becomes $\beta_0 + \beta_1(1-\sigma)F - \beta_2(1-\sigma)P$. In other words, an increase in segregation, σ, dampens the positive effects of F and negative effects of P.

To summarize, in a *fractionalized* community, there are multiple options for a common culture, and agreeing upon one may be difficult. If nation-builders promote the adoption of a neutral national identity, they can help groups coordinate on a single culture to maximize the gains from intergroup contact. However, coordination may be more elusive in a *polarized* community. With a few large groups, each is more likely to assert its own culture. This can sharpen ethnic cleavages and deepen intergroup antagonism. Ethnic loyalty shields members from such antagonism, further entrenching ethnic tribalism. The relative strengths of these competing forces determine the speed of diffusion.

Finally, note that our theoretical results describe the instantaneous growth rate of national identity.

[14] This P is a special case in the more general class of polarization indices introduced in Esteban and Ray (1994). Empirically, we follow Esteban et al. (2012) and consider generalizations that account for variable intergroup distances (see Section 7.1).

[15] This approximates the initial conditions in Transmigration settlements where the π-shares were likely small for all groups.

Empirically, we are able to estimate the level effects of initial diversity three decades after resettlement. These relationships are informative about the long-run process of identity change. In general, the model displays multiple evolutionary stable equilibria; some villages will converge to the national identity while others will feature persistent ethnic entrenchment. Appendix B.4 explores these equilibria and how they depend on initial conditions. Using an approximation argument and simulations, we show that as F increases, this widens the basin of attraction to N, and as P increases, the basin of attraction to N becomes smaller.

4 Data

This section describes several data sources that we use to measure diversity and proxies for nation building in Transmigration villages. Appendix D provides further details on the data.

4.1 Transmigration Census

To identify Transmigration villages, we digitized the 1998 Transmigration Census, produced by the Ministry of Transmigration. This provides the number of transmigrants assigned, the settlement year, and the location of each unique settlement village based on 2000 boundaries. Our main sample comprises 817 Transmigration villages (outside of Papua) settled from 1979 to 1988. These villages are dispersed across the Outer Islands (see Figure 1) and initially received 1,872 transmigrants from Java/Bali on average (with a range of 350 to 8,500). Many villages are located in contiguous settlement clusters, and village boundaries have changed over time. We account for both features of the data in robustness checks.[16]

4.2 Ethnic Diversity and Segregation

We use individual-level data from the 2010 Population Census to measure ethnic diversity and segregation in Transmigration villages. This complete-count census includes a single, self-reported ethnic identity for over 234 million individuals across Indonesia and over 2 million in Transmigration villages. There are more than 1,330 different ethnicities. We exploit this full granularity in our main analysis but also consider aggregations based on linguistic similarities between groups (Fearon, 2003) and classifications by Indonesian demographers (Ananta et al., 2013).[17] In Transmigration villages, the baseline fractionalization index (F), as defined in Section 3.2, ranges from 0 to 0.88 with a median of 0.40. Polarization (P) ranges from 0 to 0.99 with a median of 0.62. We also measure within-village ethnic segregation using enumeration details to pinpoint household residential locations (see Section 7.1).

[16] The 2000 village boundaries are our main spatial units of analysis as the policy varied at this level. While 254 Transmigration villages are isolated villages, the remainder are part of clusters containing 2–18 villages with half of those containing 2–4. Some of these clusters contain villages settled in the same year while others contain villages settled over multiple years from 1979 to 1988. What's more, by 2010, 141 of the 817 villages had split into two or more additional villages by 2010, for a total sample of 987 villages if defined using 2010 boundaries. In Table 6, we consider the effects of diversity at different levels of spatial aggregation, which ensures that our findings are not driven by the baseline village boundaries in 2000.

[17] In Transmigration villages, we see 16 Inner-Island and 700 Outer-Island ethnic groups. Inner-Island groups include all ethnicities native to Java/Bali with the top four—Javanese, Sundanese, Madurese, and Balinese—comprising over 99 percent of Inner Islanders. Meanwhile, the top 50 Outer Island ethnicities comprise over 84 percent of Outer Islanders.

4.3 Nation Building Outcomes

We consider several outcomes aimed at capturing the long-run, nation-building process. Like other recent literature, we view language, marriage, and name choices as leading indicators of culture and identity (Abramitzky et al., 2018; Giuliano and Nunn, 2018). We also explore broader measures of social capital and public goods using survey and administrative data.

Language Use at Home. Our main nation-building outcome is an indicator for whether or not individuals primarily speak Indonesian at home. In the 2010 Census, all individuals age 5 and above answer two questions: (i) Are you able to speak Indonesian? (ii) What is your primary language used at home?[18] That there are two questions about language in a short-form Census questionnaire is indicative of how important it is to the government. In Transmigration villages, 97.2 percent are able to speak Indonesian, but only 15.4 percent use it as their primary language at home. The majority instead speak their native ethnic language at home (76.4 percent).

Indonesian use at home can be seen as a choice by parents to socialize a common national identity. Because nearly everyone is able to speak Indonesian, its use at home likely reflects deeper beliefs and preferences rather than simply a desire to improve fluency—a claim on which we provide empirical evidence below. As further evidence of revealed preference, nationally-representative survey data (*Susenas* 2015) identifies an important distinction between Indonesian use *outside* versus *inside* the home. While 35 percent of Indonesians speak the national language *outside the home* on a daily basis (e.g., at work, school, etc.), one-third of those switch to their native ethnic language as the primary one *inside the home*.

Indonesian Language and Identity. Is language important for social identity?[19] We marshal evidence from two independent surveys to show that Indonesian use at home is associated with weaker ethnic attachment and stronger national integration. First, using the 2009 *Asian Barometer* survey, a cross-sectional analysis shows that home use of Indonesian is associated with a relatively stronger sense of national identity. Conditional on age, gender, education, and region fixed effects, individuals who primarily or exclusively speak Indonesian at home are 10 p.p. more likely to choose the national identity over their ethnic or other identity, relative to a mean of 63 percent.[20]

Second, panel evidence from the Indonesia Family Life Survey (IFLS) shows how Indonesian use at home may contribute to an intergenerational process of nation building. We examine how people using Indonesian at home as children (observed in 1997) made different language and identity choices after forming new households as adults more than a decade later (in 2014). The regressions in Table 2 control for age, gender, education, and village fixed effects, thus comparing observably similar individuals except for differences in parental Indonesian use as a child.

[18] Enumerators record a native ethnic language if individuals respond to (ii) with both Indonesian and an ethnic language. According to the IFLS (see Table 2), which records multiple languages at home, 56.9 percent of those speaking Indonesian at home also speak an ethnic language at home. Hence, those speaking exclusively Indonesian at home are a distinct group.

[19] As Kramsch and Widdowson (1998) argue, "There is a natural connection between language and identity insofar as language often defines membership to a specific group to the exclusion of nonmembers. Through language the group manifests 'personal strength and pride' and a 'sense of social importance and historical continuity' and most of all belonging to an 'imagined community' that shares a common worldview and that commands allegiance to it. . . ." Simpson (2007) notes that "Indonesian has also become positively valued as the primary shared component of the country's emerging national identity."

[20] The question reads: "Let us suppose you had to choose between being an Indonesian and being [own ethnic group], which of these do you feel most strongly attached to?" Responses include "Indonesian", "Own Ethnic group", and "Another Identity".

Panel A shows that adults who grew up speaking Indonesian at home are: (i) nearly 50 percent more likely to speak Indonesian at home in their new households (column 1); (ii) more likely to report a different ethnicity in 2014 than in 1997, reflecting a more fluid self-concept of ethnic identity (column 2); (iii) around 55 percent more likely to marry a non-co-ethnic (column 3); and (iv) significantly less likely to trust co-ethnics more than others (column 4). Panel B shows that these patterns are not driven solely by individuals that grew up with multiethnic parents, which is similarly and independently correlated with these four outcomes. Together, these patterns suggest that using Indonesian at home may weaken ethnic attachments and help to socialize a shared national identity across generations.

Other Nation Building Outcomes. In addition to language use at home, we examine several different measures of integration, social capital, conflict, and development. We use the 2010 Census to construct two proxies for ethnic attachment: interethnic marriage and the ethnic content of children's name choices. We use a 2012 household survey (*Susenas*) to examine subjective intergroup preferences, including trust, tolerance and willingness to contribute to local public goods, among others. Finally, we explore village-level public good provision, ethnic conflict, electoral outcomes, and development using several sources. We describe these outcomes at length when presenting results in Section 7.2.

5 Empirical Strategy

We develop our empirical strategy in four steps. First, we explain how the resettlement process generated plausibly exogenous variation in initial diversity. Second, we show that this policy-induced variation persisted over the long run. Third, we describe the variation in fractionalization (F) and polarization (P) and provide motivating evidence on how they relate to Indonesian use at home. Finally, we present a formal identification strategy to estimate causal, long-run effects of diversity.

5.1 Transmigrant Assignment and Ethnic Diversity in the New Settlements

The Transmigration program's rapid expansion beginning in 1979 contributed to an as-if-random initial assignment of transmigrants. As planners rushed to meet lofty annual targets set by the central government, institutional frictions and bottlenecks were rife, with many reports describing the haphazard implementation as a "plan-as-you-proceed" approach (Hardjono, 1988; World Bank, 1988). Coordination problems between government agencies made it infeasible to systematically match transmigrants to settlements. One agency was responsible for recruiting transmigrants in the Inner Islands, while another was tasked with clearing sites in the Outer Islands.

In practice, the arbitrary timing of transmigrants' arrival to transit camps in Java/Bali played an important role in shaping diversity in the new settlements. There were four main transit camps where transmigrants would gather for brief pre-departure orientations. Participants could not choose their destinations and, even upon departure, were often ill-informed about the conditions they would face (see Kebschull, 1986, for survey evidence). Importantly, because the camps were large, each would collect transmigrants from many different areas of Java/Bali. This resulted in ethnic mixing within the camps that would carry over to the new settlements. Suppose a new village just opened up with slots for 400 households. Given the haste in implementation, the given set of households in the camp departure queue would be sent to that settlement. At some times, that queue was ethnically homogenous, while at others it was more mixed. This arbitrariness explains why in some villages, all transmigrants are Javanese while in others one finds a mix of Inner-Island ethnic groups (see Section 5.3).

A second source of variation in diversity comes from the APPDT quotas for native Outer Islanders from nearby areas. These quotas were designed to encourage Inner–Outer ethnic mixing, to avoid Inner-Island ethnic enclaves, and to forestall local grievances in resettlement areas. Officially, these slots were reserved for residents of other villages within the same province. In 1979, *de jure* guidelines required each village's APPDT share to be 10 percent, and this increased to 25 percent in 1982. However, *de facto*, APPDT shares varied across locations and were often set by provincial officials, including the governor and local MOT leadership (Rigg, 2013; Tirtosudarmo, 1990). Some villages had APPDT shares of 50–80 percent (Tanasaldy, 2012, p. 191). In sum, APPDT shares varied due to policy rules changing over time and to discretion by provincial officials, two sources of variation borne out in 2010 Census data.[21]

Finally, planners had little scope to match Inner-Island ethnic groups to culturally similar destinations. If, for example, many Javanese arrived in a transit camp just before a new settlement opened in an ethnically Kutai region of Kalimantan, then such groups would have been forced to mix in the new settlement even if the Sundanese, who arrived later to the transit camp, would have been less culturally distant (Clauss et al., 1988). What's more, the *de facto* APPDT share for that settlement was set before local officials knew which ethnic groups would be departing from the transit camp in Java/Bali.

5.2 Persistent, Policy-Induced Diversity

The resettlement process described above resulted in persistent variation in diversity. Using 2010 Census data, Figure 2 plots the distribution of village-level F and P across Transmigration program (solid line) and non-program (dashed line) villages in the Outer Islands.[22] For Transmigration villages, we see a continuum of diversity and significant mass at relatively higher F and P. Migration frictions and land-market imperfections likely contributed to the persistence of the initial program-induced diversity.[23] In typical settings with free labor mobility, segregation and tipping forces will render such high F and P unstable. The solid density in Figure 2 shows that non-program villages—settled organically over time—are generally less diverse as people tend to self-segregate across villages along ethnic lines.

The long-run diversity in Transmigration villages is rooted in the initial policy variation. A Shapley decomposition suggests that 46 (52) percent of variation in F (P) is explained by diversity among Inner-Island ethnicities (F_{inner} and P_{inner}). This is consistent with the mixing and queuing process in the transit camps of Java/Bali naturally shuffling the ethnic mix of Inner Islanders. Another 50 (48) percent of variation in F (P) is explained by the Inner-Island ethnic share, which varies with the APPDT allocations. Most Outer Islanders in Transmigration villages either belong to one large group or many very small groups local to the settlement area, which partly explains why diversity among Outer-Island ethnicities (F_{outer} and P_{outer}) does not constitute a large share of overall variation.

[21] The data show (i) a significant increase in the mean Inner-Island ethnic share for villages settled after 1982, (ii) sizable variation around that mean across provinces, and (iii) less variation within than between provinces.

[22] See Appendix Figure A.2 for the joint distribution of F and P and Appendix Table A.1 for analogous evidence showing that the housing lottery induced lower long-run segregation within Transmigration villages (conditional on F and P).

[23] Weak property rights and missing land markets are often a barrier to migration in rural areas (see, e.g., De Janvry et al., 2012).

5.3 Motivating Evidence on Fractionalization and Polarization

The fact that F and P are highly correlated at low levels makes it difficult to separately identify their causal impacts. Figure 3 plots F against P for all Transmigration villages, reproducing a familiar shape from the cross-country figures in Montalvo and Reynal-Querol (2005). At low levels of diversity—where $F < 0.2$ and $P < 0.4$—F and P are nearly collinear. Beyond this region, the two measures are positively correlated when P is high but negatively correlated when F is high, making it difficult to determine the sign of the omitted variable bias if one of the two diversity measures is excluded.

To illustrate how independent variation in F and P affects our main outcome, we present three examples of Transmigration villages. These villages are depicted in Figure 3, where the different shapes and colors across villages correspond to different quintiles of Indonesian use at home. Some villages, like Tanjung Gading (TG), were settled with many small groups. TG is home to 43 ethnic groups, including three large Inner-Island groups (42% Javanese, 21% Banten, 9% Sundanese), one large local Outer-Island group (11% Lampung), and many other small groups. Consequently, TG has a very high F of 0.76.

By contrast, Bukit Kemuning (BK) was settled by only 14 ethnic groups. TG and BK have similar polarization levels (0.63 in TG and 0.59 in BK), but BK has a much lower F of 0.41. The model in Section 3 suggests TG will have more Indonesian use at home because, in communities with many small groups, the gains from coordinating are high while the benefits from self-segregating are low. Indeed, 95% of the population chose Indonesian as the primary language at home in TG compared to 22% in BK.

Now, consider the village of Tri Dharma Wirajaya (TDJ) and compare it to the prior village, BK. Both villages have the same fractionalization (0.41), a similar number of groups (17 and 14), and a large majority (around 75% Javanese). The key difference is that TDJ also has a large minority group (21% Sundanese) whereas BK has many small minorities (each with less than 10%). Accordingly, polarization is substantially higher in TDJ (0.71) relative to BK (0.59). The model suggests that intergroup antagonism is more intense in polarized villages with a few large groups, reducing the incentive to integrate. Only 7% of TDJ speaks Indonesian at home, compared to 22% in BK.

5.4 Identifying the Effects of Diversity

Our main specification regresses nation building outcome y on diversity in Transmigration village v:

$$y_v = \alpha + \beta_f F_v + \beta_p P_v + \mathbf{x}_v' \boldsymbol{\beta} + \varepsilon_v. \tag{7}$$

All regressions include controls in $\mathbf{x}_v$ for 21 predetermined measures of geography and agroclimatic conditions used by planners to select sites and determine the population size of the new settlements. These include, among others, several natural advantages typically associated with diversity and openness.[24] We also include island fixed effects to account for broad regional differences. In robustness checks, we further control for island-, province-, or even district-by-year-of-settlement fixed effects to rule out variation in program implementation across space and time that may be confounded with latent integration. We cluster standard errors by district, of which there are 84.[25] The model implies $\beta_f > 0$ and $\beta_p < 0$.

[24] See the notes to Table 3 for a complete elaboration of the components of $\mathbf{x}$.

[25] Appendix Table A.3 shows that inference is robust to four alternative procedures: (i) spatial HAC (Conley, 1999), (ii) wild cluster bootstrap (Cameron et al., 2008), (iii) effective degrees-of-freedom adjustment (Young, 2016), and (iv) multi-way clus-

For our core outcome (*homeIndo*) and a few others, we estimate individual-level analogues to equation (7) with up to 1.8 million people. We include an array of fixed effects to address confounders. For example, ethnicity FE address the possibility that some groups may be more open to integration and more likely to live in diverse communities. These specifications help but do not fully resolve endogeneity in today's F and P. We take three additional steps to address remaining sources of bias.

First, Appendix Table A.2 offers *prima facie* evidence against *ex post* sorting. Panel A shows that F and P in Transmigration villages today are uncorrelated with location fundamentals associated with nation building. These include (i) natural advantages (e.g., distance to district capitals and roads) and (ii) proxies for the national integration, including *homeIndo*, of populations living in nearby areas before the program. By contrast, (i) and (ii) are highly correlated with F and P in other Outer-Island, non-Transmigration villages (see Panel B), which is what we expect with endogenous sorting. The weaker and null correlations in Panel A suggest limited sorting after the initial policy assignment in the 1980s.

Second, we provide direct evidence on the plausibly exogenous assignment of initial diversity. Recall that the policy-induced variation in F and P comes from (i) the Inner-Island ethnic share and (ii) ethnic diversity among Inner Islanders. On (i), Appendix Figure A.4 shows that planners did not systematically assign more transmigrants to locations that were more nationally integrated in the 1970s or inherently attractive to (linguistically similar) migrants thereafter. Appendix XII discusses similar null results for other confounders. There, we also provide analogous evidence on the unconfoundedness of (ii) as proxied by F_{inner} and P_{inner} in 2000 for those born in Java/Bali before resettlement.

Third, we develop an instrumental variables (IV) strategy that isolates variation in initial diversity. We pin down the Inner-Island ethnic share using a flexible function of the number of transmigrants in the initial year. The $\mathbf{x}$ vector in equation (7) proxies for the policy rule determining the carrying capacity and potential population in each village. Therefore, conditional on $\mathbf{x}$, a larger initial stock of transmigrants implies a higher Inner-Island ethnic share. We pin down the ethnic mix among transmigrants using ethnic group shares for Inner Islanders born in Java/Bali (based on the 2000 Census).[26] Appendix XII shows that the two sets of instruments are strong predictors of F and P in 2010. The exclusion restriction requires that planners did not create more diverse settlements in locations that were unobservably more prone to integration. The abovementioned tests provide supportive evidence, suggesting that the instruments are uncorrelated with a large number of historical correlates of nation building.

6 Results: Diversity and National Language Use at Home

In this section, we estimate the effects of diversity on national integration, as proxied by Indonesian use at home. First, we present baseline results and evidence consistent with a social identity interpretation. Second, we address threats to causal identification.

tering on birthplace and ethnicity (Cameron et al., 2011). Given this robustness, we opt for the baseline clustering by district in all other tables throughout the paper and appendix.

[26] These data provide the best available proxy for the initial ethnic composition of transmigrants from Java/Bali. They are of course limited by the possibility that death and re-migration rates may differ across ethnic groups between the year of settlement and 2000. However, we do not think this is a major source of bias. If it were, diversity among those born before the year of settlement would differ substantially from diversity among those born after. Instead, F and P for the older generation nearly perfectly predict F and P for the younger generation with a coefficient that is indistinguishable from one.

6.1 Main Results

Table 3 presents our core results for *homeIndo*. We focus on OLS specifications and leave IV results to robustness checks below. We begin with village-level regressions where the dependent variable is the share of individuals in the village who mainly speak Indonesian at home.

Columns 1 to 3 demonstrate the importance of estimating conditional effects of F and P. Recall from Figure 3 that F and P are positively correlated as diversity increases from very low levels but negatively correlated in other regions. In columns 1 and 2, we estimate significant positive unconditional effects of F and P, but these effects may be coming primarily from the first region with low levels of diversity. In column 3, where we include both measures, the sign reverses for P and the coefficient on F increases substantially. Hence, important independent variation in F and P was not captured in the unconditional estimates. But this is precisely the variation needed to identify the two distinct forces in the model of Section 3. Including one measure of diversity but not the other confounds this distinction.[27]

The estimates in column 3 of Table 3 imply significant effects of ethnic diversity on Indonesian use at home. A one standard deviation (s.d.) increase in F (holding P constant) leads to 12.9 p.p. greater *homeIndo*. By contrast, a one s.d. increase in P (holding F constant) leads to 8.1 p.p. lower *homeIndo*. These results are consistent with the predictions of our model, which suggest that F (P) captures benefits (costs) of intergroup contact. These are also large effects relative to the mean village where 14.4 percent of individuals speak Indonesian at home. For reference, a one s.d. increase in F equals 0.21 relative to a mean of 0.41, and a one s.d. increase in P equals 0.23 relative to a mean of 0.57.

Together, the opposing effects of F and P in Table 3 suggest that national integration is stronger in villages with many small groups relative to villages with a few large groups. Taking the relationship between Indonesian use at home and national identity from the *Asian Barometer* (see Section 4.3), the standardized effects of F and P in column 3 imply, respectively, a 14.1 (8.9) percent increase (decrease) in national relative to ethnic identity.

We validate these opposing forces of diversity in Figure 4 by estimating a flexible specification with quintiles of F and P and an exhaustive set of interactions thereof:

$$y_v = \alpha + \sum_{i=1}^{5}\sum_{j=1}^{5}\theta_{ij}\mathbf{1}\Big\{F_v \in [\kappa_{i-1},\kappa_i) \text{ and } P_v \in [\rho_{j-1},\rho_j)\Big\} + \mathbf{x}_v'\boldsymbol{\beta} + \varepsilon_v. \tag{8}$$

where $\kappa_0 = \rho_0 = 0$,and κ_i and ρ_j for $i,j = 1,...,5$ respectively index the upper bounds of quintiles of F and P across Transmigration villages. This specification provides a richer approximation of the underlying variation and a stricter comparison across villages with similar F but different P (and vice versa), along the lines of the motivating village examples in Section 5.3. Figure 4 plots results for each ij cell, adding the estimated $\widehat{\theta}_{ij}$ to the mean Indonesian use at home of 0.036 for the bottom reference quintiles of F and P ($i = j = 1$). The estimates follow a similar pattern as that seen in the raw data plotted in Figure 3. There is a roughly monotonic increase in Indonesian use at home moving towards more fractionalized villages at a given level of polarization and vice versa moving towards more polar-

[27] Furthermore, we show in Appendix Table A.4 that the effects of F and P are robust to controlling for the size of one's own ethnic group in the village. Of course, for homogenous villages or those with just two groups, the own-group share is sufficient to identify the relationship of interest. However, the considerable variation in the number and size of groups suggests that both F and P are necessary to capture the effects of ethnic composition on individual behavior.

ized villages at a given level of fractionalization. The point estimates and standard errors can be seen in Appendix Table A.5, which shows that all but a few lower quintile interactions are significantly different at the 1% level from the reference villages with very little to no diversity ($i = j = 1$).

Returning to the linear specification in Table 3, we estimate individual-level regressions using the full Census microdata. Column 4 shows that the analogous baseline estimate is indistinguishable from the village-level estimate in column 3. Column 5 includes exhaustive fixed effects (FE) for the 95 ages, 2 genders, and 716 ethnicities in Transmigration villages. Thus, we compare, for example, Javanese living in villages with different F and P. Column 6 additionally controls for 496 birth district and 84 current district FEs. This compares, for example, individuals born in the same district of Java/Bali or residing in the same Outer-Island district today but living in villages with different F and P. The full set of FEs in column 6 cuts the baseline effects in half, but the effects remain sizable despite the more limited identifying variation.[28] Overall, these demanding FEs ensure that the effects are not driven by compositional differences associated with a proclivity for Indonesian use (e.g., younger people, ethnicities with native languages closer to Indonesian, or immigrants from or living in tolerant regions).

Together, the results in Table 3 provide strong evidence that national language use at home is increasing in ethnic fractionalization and decreasing in ethnic polarization. Appendix Table A.6 shows that these effects are driven by individuals switching out of their own native ethnic language (rather than choosing another ethnic group's language). We interpret these results as evidence of fractionalization (polarization) weakening (deepening) attachment to one's native ethnic identity.

National Language Use as National Identity? While individuals may face immediate economic incentives to speak Indonesian at home, this choice may also reflect deeper, long-run investments in identity. We present evidence here consistent with that interpretation. If, for example, individuals were speaking Indonesian at home only to increase their fluency or improve their skills in the local labor market, we would observe sharp differences in the effects of F and P across individuals with different education levels or who sorted into occupations on the basis of comparative advantage in Indonesian.

Instead, we find stable effects of F and P across different education levels and employment sectors. Appendix Table A.7 splits the sample in column 6 of Table 3 across six education levels, ranging from no schooling in column 2 to some post-secondary schooling in column 7. Compared to the baseline estimate reproduced in column 1, we see similar standardized effects of F and P across education levels. This suggests that different degrees of fluency and exposure to Indonesian in schools cannot fully explain the effects. Appendix Table A.8 presents related insights based on sample splitting by sector of employment for all working-age individuals (column 1). We restrict to those not working in column 2 and then to those working in six broad sectors: (3) agriculture and mining, (4) manufacturing, (5) other manual (e.g., construction), (6) trade and services, (7) white collar (e.g., banking), and (8) other. Sectors differ along many dimensions, but two important ones are skill requirements and the likelihood of mixing with other groups in the workplace, both of which are plausibly lowest for agriculture and highest for white collar, trade, and services. Yet, individuals exhibit similar responses to F and P across sectors.[29]

[28] In Appendix Table A.9, we find very similar results when including (i) FE for each birth district–current district pair (16,109 in total), or (ii) FE for each ethnicity–current district pair (4,575 in total). These similar results suggest that sorting along particular origin–destination corridors or particular ethnicity–destination matches cannot explain our findings. We opt for the more parsimonious, additive FEs in column 6 of Table 3 as the main individual-level specification moving forward.

[29] Although some of the sample splits in Tables A.7 and A.8 may be endogenous outcomes of F and P, we view the stable

Interpreted through the model in Section 3, the results thus far suggest that exposure to diversity may change one's incentives to invest in forms of identity conducive to integration. For some, the incentives to embrace Indonesian may be strongly economic; for others, less so. However, from the nation building perspective, what matters first and foremost is that local diversity affects ethnic attachment. Whether that occurs as a result of initial economic or non-economic incentives is less first order, and, in fact, both forces are at play in our model.

We close this section with two results that further point to a social identity motive for *homeIndo* that goes beyond economic incentives to improve fluency. Both use the individual-level specification in column 6 of Table 3. First, regressing an indicator for one's ability to speak Indonesian on diversity yields small standardized effects of 0.007(0.001) and -0.003(0.002) for F and P, respectively. While Indonesian ability is responsive to diversity, the effects are orders of magnitude smaller than those for its use at home. This is not surprising since nearly everyone can speak Indonesian.

Second, in an even stronger test, we find large effects of diversity on *homeIndo* for the ethnic Malay population.[30] This group has little economic incentive to speak Indonesian at home given that they have native fluency already (as Indonesian is based on the Malay language). Yet, F and P have large standardized effects in a regression restricted to ethnic Malay in Transmigration villages: 0.104(0.021) for F and -0.050(0.022) for P relative to a mean of 21.3%. This suggests that *homeIndo* must be capturing something deeper than latent fluency or a desire to improve one's skills thereof. For Malays to report Indonesian rather than their mutually intelligible mother tongue, they plausibly feel more invested in the national identity. What's more, as seen in Figure 5, the effects of F and P are similar for several other large ethnic groups in Transmigration villages, including the most numerous Javanese and Sundanese.

6.2 Addressing Threats to Identification

This section presents addresses key concerns about endogenous sorting and other confounders.

Instrumental Variables Estimates. Table 4 shows that the IV procedure detailed in Section 5.4 delivers similar estimates as the baseline OLS results. We re-estimate columns 3–6 of Table 3 using instruments that isolate the policy-induced variation in initial diversity across Transmigration settlements.[31] These IV-GMM estimates are generally larger than the corresponding OLS in Table 3. In the village-level specification in column 1, the coefficient on F increases from 0.637 to 1.017 and on P from -0.362 to -0.793. However, we cannot reject that the IV estimates are different from the OLS estimates (based on a Hausman-type GMM test). Similar patterns hold for individual-level regressions in columns 2–4.

The similarity between OLS and IV estimates points to the persistent impact of the initial settlers on diversity 2–3 decades later. Coupled with earlier evidence against endogenous initial assignments, this reinforces the notion that the program generated plausibly exogenous variation in diversity. While

effects across different sub-populations as informative nonetheless.

[30] Malay comprise 5% of the total population and 16.7% of the native Outer-Island ethnic population in Transmigration villages.

[31] Given the many instruments, we estimate the 2SLS equations using Generalized Method of Moments (GMM) for greater efficiency. At the bottom of the table, the Sanderson and Windmeijer (2016) Wald statistics reject the null of weak instruments on the two endogenous variables. Based on the Hansen (1982) test, we cannot reject the null hypothesis that the instruments are uncorrelated with the error term and are correctly excluded from the second stage. Coupled with the rejection of the null under the Anderson and Rubin (1949) test (that the coefficients on the endogenous variables jointly equal zero and the overidentifying restrictions are valid), these diagnostics point to a well-specified IV model.

the IV and OLS results are statistically indistinguishable, larger IV point estimates are consistent with a Local Average Treatment Effect (LATE) in which the instruments isolate policy-induced compliers who are more responsive to diversity. In contrast, OLS could be capturing tolerant individuals who endogenously sorted and hence are less affected by diversity, biasing the OLS estimates towards zero.

To clarify the LATE-based interpretation, we shut down one source of endogenous deviation from the policy rules by only considering Transmigration villages where the Outer-Island ethnic share today is below the *de jure* APPDT quotas (see Section 5.1). For these villages, most of the overall diversity comes from ethnic differences among the transmigrants who, unlike Outer-Island natives, could not choose their destination village. Compared to the baseline OLS estimates from column 3 of Table 3, the effects for F increase from 0.637(0.073) to 1.238(0.164) and for P from -0.362(0.051) to -0.676(0.108). These magnitudes are similar to the IV estimates. To better understand why, we turn to a more thorough investigation of sorting using the individual-level Census data.

Further Checks on Sorting. Several results suggest that endogenous sorting is unlikely to explain the main findings in Table 3. First, in Table 5, we separately estimate the effects of diversity for different ethnic and immigrant sub-populations in Transmigration villages. For reference, column 1 reproduces the estimate from column 6 of Table 3, but we standardize coefficients within-sample for ease of comparison across columns.[32] Column 2 restricts to Inner-Island ethnics, most of whom are first- or second-generation transmigrants from Java/Bali assigned by planners to the given village. Column 3 restricts further to first-generation transmigrants born in Java/Bali before the given Transmigration village was created. These two samples exhibit similar responses to F and P as the full sample in column 1. This suggests that our main findings are driven largely by the initial transmigrants and their children.

Columns 4 and 5 estimate analogous specifications, respectively, for Outer-Island ethnics and first-generation residents born in the Outer Islands before their village was created. Here, the effect sizes for both F and P are smaller. Native Outer Islanders could be less responsive to local diversity because they have more proximate “exit” options: greater potential to interact with fellow Outer Islanders outside the settlements and easier access to their (nearby) origin villages.

We sharpen this sorting interpretation in the remaining columns of Table 5. We split the Outer-Island natives in column 5 into those born in nearby districts eligible for inclusion in the APPDT quota for the given village (column 6) and those born in faraway districts and hence ineligible for APPDT (column 7).[33] The latter are likely to have migrated over long distances to reach the given Transmigration village and hence are more likely to exhibit stronger endogenous sorting on unobservables. Therefore, it is not surprising that these long-distance “sorters” have considerably higher Indonesian use at home today (32% versus 16%) *and* are less responsive to local diversity.

Finally, column 8 restricts to plausible children of initial transmigrant or APPDT settlers. These individuals were born in the given district after the year of settlement, but we cannot say for certain

[32] While mean Indonesian use at home differs across columns, the baseline fixed effects in column 1 (for ethnicity, age, gender, current district and birth district), used in every column thereafter, make it possible to compare standardized coefficients.

[33] Tirtosudarmo (1990) discusses other non-APPDT categories of Outer-Island natives that joined Transmigration settlements through official means in certain regions (e.g., *swakarsa*, ‘resettlement’, or *sisipan*). These groups were less numerous than the APPDT and received less government assistance during the relocation process. Importantly, though, mean F and P are statistically indistinguishable across APPDT-eligible and -ineligible groups in columns 6 and 7, which suggests that these long-distance sorters are not more prevalent in villages with particular types of diversity.

whether they moved to the village from elsewhere in the district. The effects are similar to those in column 2. Akin to the first-generation transmigrants from Java/Bali, these individuals were plausibly exposed to diversity as a result of others' choices (their parents). Together, the results in columns 2 and 8 show that the baseline effects are due to initial program assignments rather than sorting.[34]

It is still possible that long-distance sorters exhibit strong spillover effects that explain our overall findings. We address this concern in Appendix Table A.10 by controlling flexibly for the share of the village population that we classified in column 7 of Table 5 as ineligible for the APPDT. Doing so leaves the main results unchanged. In other words, the baseline effects of diversity for program-assigned migrants are not confounded by the prevalence of those settling in the village through endogenous sorting.

Confounding Variation in Assignment Rules. It is also possible that planners created more or less diverse villages in locations deemed more suitable for nation building. We know that *de jure* APPDT quotas increased over the 1980s, and sensitive regions were allowed higher APPDT quotas *de facto*. There may be other confounding sources of variation in local assignment rules that remain unobservable. Appendix Table A.11 addresses such concerns using our village-level specification (baseline in column 1).

Columns 2–5 include separate year-of-settlement FE in different islands, provinces, and districts. These specifications compare across villages created in the same year within the same region. Column 5 includes 303 district-by-year-of-settlement FE, effectively bringing us close to a matching-type estimator that compares the effects of F and P across a few nearby villages. Consistent with the reduced variation in F and P, the effects fall by 20–30% but still remain sizable.

Column 6 includes FE for the 102 indigenous ethnolinguistic homelands that span Transmigration villages. This addresses the possibility that planners created more diverse villages in regions where the local ethnic group is more culturally similar to Inner-Island ethnic groups. Column 7 interacts these FE with year-of-settlement FE. Again, the effects slightly fall but remain sizable. Together, these results in Appendix Table A.11 mitigate the concern that planners may have learned over time or space about which locations were more amenable to the nation-building goals of ethnic mixing.

Summary. Overall, this section provided evidence against important identification concerns about *ex post* sorting and confounding variation in initial assignments. While reassuring, these results may raise the question of what identifying variation remains when comparing villages settled at the same time in the same region. Our baseline controls ($\mathbf{x}$) absorb much of the potentially concerning residual local variation (e.g., proximity to roads). Moreover, results are unchanged when adding further village-level controls capturing predetermined natural advantages associated with agricultural development (i.e., potential crop yields) or the disease environment (i.e., a malaria index from 1978). These checks suggest that the effects of diversity on *homeIndo* are not driven by individuals or places with unobservable predisposition to national integration. Rather, the haphazard resettlement process generated significant variation in diversity even across nearby settlements with similar natural advantages. Unanticipated exposure to such diversity then shaped identity choice as seen through language use at home.

[34] We further validate this point by examining the 103,338 individuals that immigrated into Transmigration villages from other districts in the last five years. These individuals have predictably higher *homeIndo* on average but are also less sensitive to local diversity. Reproducing a version of Table 5 based on this sample yields small and insignificant effects of diversity, particularly for native Outer Islanders. Omitting these individuals from the baseline regressions leaves results unchanged.

7 Mechanisms and Other Outcomes

This section provides deeper insight into why ethnic diversity affects nation building. We first show how different dimensions of intergroup distance—spatial, economic, and cultural—shape the relationship between diversity and national language use at home. We then demonstrate effects of diversity on several other outcomes related to the nation-building process.

7.1 Mechanisms: Intergroup Distance and the Salience of Ethnic Divisions

Our baseline findings suggested that F is conducive to national integration while P deepens ethnic attachment. We provide three sets of results that clarify how underlying ethnic divisions become salient, driving these results. First, residential segregation determines the scope for intergroup contact to change behavior in diverse communities. Second, interethnic inequality undermines the benefits of diversity. Third, cultural distance between ethnic groups amplifies the effects of diversity.

Residential Segregation and Intergroup Contact. In Table 6, we use the full spatial detail in the 2010 Census to show that diversity at the neighbor(hood) level may be more important in shaping behavior than diversity at more aggregate levels. We measure F and P at the sub-village administrative level and also identify the ethnic mix of next-door neighbors by leveraging the zigzag enumeration method (see Appendix D).[35] In each case, we estimate individual-level regressions based on the FE specification in column 6 of Table 3 and standardize the diversity measures for comparability.

Moving from left to right in Table 6, the diversity measures become increasingly local while the effect sizes grow larger. Column 1 examines diversity at an aggregate level that includes all contiguous Transmigration villages (see Section 4.1). Contiguous-cluster-level diversity has somewhat weaker effects than village-level diversity in our baseline estimate, reproduced in column 2. Column 3 then looks at diversity across neighborhoods (*rukun tetangga* or RT) within villages.[36] The resulting effects of F and P are significantly larger than at the village level. Column 4 goes even more local by examining ethnic differences with neighbors in the two adjacent housing units. Relative to households with both neighbors of the same ethnicity, those with one (both) neighbor(s) from a different ethnicity are 6.6 (19.1) p.p. more likely to speak Indonesian as the main language at home. The results in columns 3 and 4 are robust to the use of village fixed effects (that absorb village-level diversity). Column 5, which includes all diversity measures simultaneously, shows that neighborhood and within-neighborhood diversity are the strongest drivers of *homeIndo*. Cluster- and village-level diversity have small and mostly insignificant effects when included alongside these more localized diversity measures.

Column 6 rounds out these findings with a village-level summary measure of ethnic segregation due to Alesina and Zhuravskaya (2011).[37] As neighborhood-level ethnic shares differ from village-level shares, the segregation index, S, increases. With full segregation, each neighborhood contains a separate

[35] This zigzag approach is similar to Logan and Parman (2017) who study racial segregation in historical U.S. Censuses.

[36] These neighborhoods are the lowest level of governance, with leaders responsible for facilitating public good provision in tandem with the village government. Across Transmigration villages, the median has 15 RT, while the maximum has 59 RT.

[37] The index for village v is given by $S_v = \frac{1}{I-1}\sum_{i=1}^{I}\sum_{b=1}^{B}\left(\frac{n_{bv}}{N_v}\right)\frac{(\pi_{ibv}-\pi_{iv})^2}{\pi_{iv}}$, where $i = 1,\ldots,I$ denotes ethnicities, and (n_{bv}/N_v) measures the population of census block $b = 1,\ldots,B$ as a fraction of the total village population. S_v is the squared coefficient of variation between block-level ethnic shares, π_{ibv} and village-level ethnic shares, π_{iv}.

ethnic group, and S equals one. If every block has the same ethnic mix as the overall village, S equals zero. In Transmigration villages, S ranges from 0 to 0.27 with a mean and standard deviation of 0.03. We find that a one s.d. increase in S reduces *homeIndo* to the same extent as a one s.d. increase in P.

Although the housing lottery generated exogenous variation in initial segregation, it is possible that newly formed households endogenously sorted within the village over subsequent years. We address this *ex post* residential sorting by instrumenting for overall ethnic segregation (S) with the ethnic segregation among the original parental cohort (S^{old}). This older cohort—born $\geq$ 15 years before the year of settlement—is likely to be living in the same house assigned upon arrival, whereas their *young* (grand)children may have established new houses elsewhere in the village. In practice, though, S^{old} is highly correlated with S^{young} ($\rho = 0.86$). This is consistent with inheritance norms in rural areas where land is passed on to children who then form households near their parents. As a result, instrumenting S with S^{old} does not materially change the coefficient on S in column 6 of Table 6.

In Table 7, we make the intergroup contact mechanism even more precise by showing how segregation attenuates the effects of overall diversity. Column 1 runs a village-level specification analogous to the individual-level regression in column 6 of Table 6. Column 2 then adds interactions of S with F and P, with each measure standardized (pre-interaction). Segregation dampens both the positive effects of F and the negative effects of P. By limiting local contact, segregation makes fractionalized communities seem more ethnically homogenous at the neighborhood level, thereby increasing incentives for ethnic attachment. Analogously, in villages where different groups are isolated from each other, the negative effects of polarization are more muted as there are fewer venues for intergroup antagonism to materialize. Together, these results are consistent with a model extension where segregation alters the matching function, making it less likely for one to meet non-co-ethnics (see equation (B.4) in the Appendix).

Interethnic Inequality. Beyond physical proximity, economic inequality is another potentially important means by which ethnic divisions become salient. We explore this mechanism in columns 3–4 of Table 7 using a measure of interethnic inequality. Although settlers received the same quantity and expected quality of assets upon arrival in the new settlements, there may have been initial differences in human capital endowments across groups as a result of the arbitrary assignment process. We capture these differences using a measure of location-specific human capital that is predetermined and exogenous with respect to initial diversity: agroclimatic similarity ($\mathcal{A}_{od}$) between an initial settler's origin o and the given destination d.[38] $\mathcal{A}_{od}$ measures the extent to which the agroclimatic environment in an individual's district of birth was similar to the environment where that individual was placed. As shown in our prior work (Bazzi et al., 2016), agroclimatic similarity is a good proxy for skill transferability and hence an important determinant of economic well-being. Inequality between ethnic groups in this skill might exacerbate ethnic differences to the extent that it leads, for example, to inequality in economic opportunities or in the ability to cope with shocks.

Interethnic inequality reduces national language use at home. The estimate in column 3 implies 3.3 p.p. lower *homeIndo* for a one s.d. increase in interethnic inequality. These results are conditional on F, P,

[38] We construct an index of inequality in agroclimatic similarity between all ethnic groups i and j in village v, *Between-Group Agroclimatic-Similarity Inequality*: $\text{BGAI}_v = \frac{1}{2\bar{a}} \sum_{i=1}^{I} \sum_{j=1}^{J} n_i n_j |a_i - a_j|$, where n_i is the relative size of ethnic group i, a_i is the average agroclimatic similarity within each ethnic-group i and $\bar{a}$ is the average agroclimatic similarity within each village v. This measure is akin to the between-group rainfall inequality index in Guariso and Rogall (2017).

village-level average agroclimatic similarity, and overall inequality in agriclimatic similarity (regardless of ethnicity); these last two measures have small and null effects (not shown). Column 4 then introduces interaction terms showing that interethnic inequality reduces the positive effects of F on integration but exhibits no significant heterogeneity with respect to polarization. Overall, these results suggest that interethnic inequality changes the type of contact, making it potentially more antagonistic (Lowe, 2018).

Ethnolinguistic Distance. The remaining columns of Table 7 explore how native linguistic differences may accentuate ethnic divisions. Our baseline measures of diversity treat every self-reported ethnicity as equally distant from every other ethnicity. Thus, for example, the Batak Tapanuli are equidistant from the Javanese and the Batak Toba even though both Batak sub-groups have mutually intelligible languages and similar cultures. Columns 5 and 6 of Table 7 explore whether this simplification obscures aspects of ethnic diversity that are important in shaping identity choices.

Column 5 shows that the baseline effect sizes for F and P are unchanged when using a coarser definition of ethnic identity that consolidates the 1,330 self-reported ethnicities into 44 broad groups stipulated by Indonesian demographers (Ananta et al., 2013). We cannot reject that the coefficients are different from those in column 3 of Table 3 (appropriately standardized). This coarse grouping, which mostly obscures diversity among Outer Island ethnicities, seems to capture the leading sources of variation in diversity in Transmigration villages. This suggests that deeper ethnic divisions are driving differences in national language use at home. In other words, it is the differences between Javanese and Batak rather than between Batak Toba and Batak Tapanuli that matters for *homeIndo*.

We validate this interpretation in column 6, which adjusts F and P for the linguistic distance between each ethnic group based on native language classifications. This generalization of F is given by the Gini-Greenberg index, $F_v(\delta) = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{iv} p_{jv} \delta_{ij}$, where δ_{ij} measures the linguistic distance between groups i and j. The generalization of P is given by the Esteban and Ray (1994) formulation: $P_v(\delta) = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{iv}^2 p_{jv} \delta_{ij}$.[39] We follow the literature in defining $\delta_{ij} = 1 - \left(\frac{\text{branch}_{ij}}{\max(\text{branch}_i, \text{branch}_j)} \right)^{\kappa}$ based on the fraction of possible shared branches on linguistic classification trees from the *Ethnologue* database (see Appendix D.4). We set κ to 0.05 as in Esteban et al. (2012). This low κ amplifies deeper ethnolinguistic cleavages by accentuating, for example, the Javanese–Batak difference more than the Javanese–Sunda difference because the Sunda language is more similar to Javanese than is Batak.

These linguistic-distance-adjusted diversity measures have slightly larger effects on *homeIndo*, and we can reject at the 5% level that $P(\delta)$ has the same effect as P. Increasing κ brings us closer to the results for baseline F and P as expected. For example, with $\kappa = 0.5$ (as in Desmet et al., 2009), the coefficient on fractionalization (polarization) is 0.142 (-0.088) compared to 0.144 (-0.092) for $\kappa = 0.05$ in column 6 and 0.135 (-0.084) for the baseline $\kappa = \infty$. By down-weighting culturally similar groups, these δ adjustments make clear that deep-rooted linguistic differences between ethnic groups are an important factor shaping the relationship between diversity and integration.

7.2 Other Evidence on Local Diversity and Nation Building

The results thus far suggest that national language use at home reveals weaker attachment to one's ethnic group and perhaps a greater affinity for the national identity. This section provides corroborating

[39] If $\delta_{ij} = 1$ when $i \neq j$, and if $\delta_{ii} = 0$ for all i, then $F_v(\delta) = \sum_{i=1}^{I} \sum_{i \neq j}^{J} p_{iv} p_{jv} = F_v$ and $P_v(\delta) = \sum_{i=1}^{I} \sum_{i \neq j}^{J} p_{iv}^2 p_{jv} = P_v$.

evidence using a host of other outcomes. These other proxies for nation building provide (i) further validation of the revealed preference interpretation of national language use as an identity choice, and (ii) evidence of broader economic and social implications of diversity.

Intermarriage. Intermarriage has long been viewed as a leading indicator of integration, and officials in the Ministry of Transmigration monitored marriage between Inner and Outer Islanders in the new settlements (Babcock, 1986). Such marriages may be important for nation building: children in intermarried households exhibit greater tolerance and weaker ethnic attachment later in life (see Table 2).

We use 2000 and 2010 Population Census data to measure intermarriage, focusing on young cohorts plausibly married after resettlement.[40] Despite Indonesia's diversity, intermarriage is rare: across the country, only 10 percent marry outside their ethnic group. For young households, Transmigration villages had an average intermarriage rate of 15.2 percent (17.8 percent in 2010). As a benchmark, the intermarriage rate in the capital city of Jakarta for roughly the same age cohort is 34.2 percent. At the average rate of increase in intermarriage across Transmigration villages, 2.6 p.p. per decade, it will take 58 years for the mean Transmigration village to arrive at the intermarriage rate in Jakarta.

Table 8 shows that F hastens and P hinders the otherwise slow process of integration through marriage. We estimate results using the 2000 and 2010 Censuses, defining diversity in the given year. Columns 1 and 2 show that a one s.d. increase in F (P) is associated with roughly 50 (15) percent higher (lower) intermarriage rates. These patterns are consistent with weaker ethnic attachment in more fractionalized communities and stronger ethnic attachment in more polarized communities.

Of course, intermarriage rates reflect both demand for and supply of non-co-ethnic spouses in the village. We use a simple reduced form approach to adjust for these supply effects at the village level. We divide the actual intermarriage rate by the average intermarriage rate from 10,000 simulations of random matching among the young, married population in each village.[41] In 2000, for example, the actual intermarriage rate is only 38.8 percent of the average rate from random matching.

Columns 3 and 4 of Table 8 show that polarization still has a statistically and economically significant negative effect, even after adjusting for the random intergroup matching rate. We find very similar results when including a quadratic or cubic polynomial in potential intermarriage rates on the right-hand side of columns 1 and 2 (instead of adjusting the left-hand side as in columns 3 and 4). The effect of F is no longer significant as it is highly correlated with the random matching rate ($\rho = 0.987$).[42] Fractionalization increases the likelihood of intermarriage by increasing the potential for intergroup contact, but polarization captures intergroup antagonism, above and beyond changes in the potential supply of different groups in the local marriage market.

To better understand this result, consider two Transmigration villages. Terusan Makmur is somewhat fractionalized ($F = 0.60$) but very polarized ($P = 0.90$) with 47.3% Balinese, 41.6% Javanese, and 8.0% local native Banjar. The high F implies a high supply of potential non-co-ethnic partners. Yet, the

[40] In practice, we restrict to households where the head is younger than the legal marriage age (15) in the year of settlement.

[41] We treat the village as the marriage market. If we used the district instead, we would have smaller supply adjustments. This is because supply effects due to the program are concentrated at the village level, and quite muted at the district level. Hence, supply adjustments at the village level are more conservative. Note also that these adjustments are based on the married population (of household heads and spouses) whereas the diversity regressors are based on the entire population.

[42] The probability of a mixed marriage is equal to a weighted average of $p_g \times (1 - p_g)$, where p_g is the population share of group g. The weights, based on the gender-specific marriage-age population of g, explain the lack of a perfect correlation with F.

actual intermarriage rate of 0.07 is only a small fraction of the potential intermarriage rate of 0.59. By comparison, the village of Rimba Beringin has similar $F = 0.59$ but much lower $P = 0.68$, and an actual intermarriage rate of 0.19. These examples and the results in Table 8 suggest that the choice of marriage partners varies with attachment to one's ethnic identity, which is fueled by polarization.

In closing the discussion, it is important to note that intermarried households do not explain the overall effects of diversity on Indonesian use at home. Appendix Table A.12 shows that F and P have similar effects on children with and without intermarried parents. If anything, the effects are slightly smaller for children in intermarried households, which is consistent with the message from Table 6 insomuch as intrahousehold diversity has more proximate effects than village-level diversity.

Children's Name Choices. In addition to language and marriage choices, children's names can be informative about nation building. This is arguably the first act of intergenerational cultural transmission, reflecting parents' preferences and expectations about the value of different identities. Using the 2010 Census, we construct four indices measuring the extent to which a name conveys weaker ethnic attachment and stronger national integration.[43] Importantly, while these indices are correlated with *homeIndo* and intermarriage, names are an additional margin of identity choice. Many children have names evocative of integration even though they live in an ethnically homogenous household where everyone speaks the native ethnic language.

Our first index associates children's names with speaking Indonesian at home. Similar to the "black name index" in Fryer and Levitt (2004), for each first name n, we calculate the relative likelihood, between 0 and 1, that n is associated with someone who speaks Indonesian at home:

$$\textsc{Indo score}_n = \frac{\mathbb{P}\left(name = n \,|\, homeIndo = 1\right)}{\mathbb{P}\left(name = n \,|\, homeIndo = 1\right) + \mathbb{P}\left(name = n \,|\, homeIndo = 0\right)}. \tag{9}$$

For example, consider the name, Asep. The numerator measures the fraction of people named Asep among those speaking Indonesian at home. The denominator is the sum of this term and the probability that someone has this name if they do not not speak Indonesian at home. If everyone named Asep speaks Indonesian at home, the index equals 1. We construct this likelihood for everyone living outside Transmigration villages (more than 200 million people), and then apply the score to children in Transmigration villages born after resettlement. We standardize the proxy for ease of interpretation.

In Table 9, we relate INDO SCORE and three other indices to village-level diversity. We estimate individual-level regressions for plausible second-generation immigrants in Transmigration villages. This is the same sample as column 8 of Table 5 but now includes those under 5. These regressions include ethnicity FE, which subsume unobservable, ethnicity-specific naming conventions. In column 1 of Table 9, we see that fractionalization is associated with children's names that are more predictive of *homeIndo*. Polarization acts in the opposite direction with effects of a similar magnitude. In column 2, our second index associates names with the likelihood of living in an intermarried household. We see again that F leads to greater integration, while P has the opposite effect. In column 3, our third index associates names with the likelihood of living in an urban area. While nearly all Transmigration villages are in

[43] We focus on measures based on individual names but exclude those with names that are not shared by at least 100 people in the entire country. Fryer and Levitt (2004) implement a similar cutoff rule, and our results are robust to other cutoffs. Appendix Table A.13 estimates a similar set of regressions for all children's names using a double-metaphone adjustment that groups similar-sounding names prior to calculating the indices and hence does not require stipulating such a cutoff.

rural areas, diversity may lead parents to give their children names that are more indicative of the types of names given in cosmopolitan urban areas. The coefficients on F and P for this measure are similar.

Our fourth index, the dependent variable in column 4, associates names with ethnic attachment by generalizing the procedure above to allow for many identity groups. For each individual with name n and ethnic group g, we calculate the relative likelihood that n is associated with g. For example, suppose Asep is highly indicative of the Sundanese ethnic group but mildly indicative of the Batak ethnic group (i.e., few Batak choose this name). Then, a Sundanese person named Asep will have a high own-ethnic-index value (i.e., his name reveals a strong own-ethnic-attachment to Sundanese), but a Batak person named Asep will have a low own-ethnic-index value (i.e., his name reveals a weak attachment to Batak). Compared to the indices in prior columns, we see mirror image effects using this measure: F reduces the precision of a child's name in identifying his or her ethnic group while P increases it.

Overall, the lessons from Table 9 are clear: fractionalization leads parents to choose names more evocative of national integration while polarization fuels more insular name choices. To be sure, the results in Table 9 are capturing much of the same variation as the baseline findings for *homeIndo* and intermarriage. It is nevertheless reassuring to find similar effects of F and P on name choice. Next, we use survey data to validate the findings from these revealed preference measures of integration.

Social Capital. Table 10 provides new evidence that polarization undermines social capital. We explore eight questions from the sociocultural module of the 2012 National Socioeconomic Survey (*Susenas*): (1) willingness to contribute to voluntary public goods; (2) participation in neighborhood social activities; (3) tolerance of non-co-ethnics in the village; (4) trust of neighbors to watch one's house; (5) trust of neighbors to care for one's children; (6) how safe one feels; (7) how easy it is to obtain help from neighbors; and (8) willingness to assist unfortunate neighbors. These measures provide a window into subjective intergroup preferences and interaction. Each outcome, in rows, is reported on a 1 to 4 scale with higher numbers indicating greater support for the given statement. The columns report beta coefficients on F and P for ease of interpretation. One limitation is that because this is a national survey, it only covers 87 Transmigration villages with around 10 respondents (households heads) per village.

Individuals in polarized villages are less likely to contribute to public goods (row 1) or to join community groups (row 2), though the latter is statistically insignificant.[44] Polarization also reduces tolerance of non-co-ethnic activities in the village (row 3), trust in neighbors (rows 4 and 5), feelings of safety (row 6), helpfulness of neighbors (row 7), and support for poorer neighbors (row 8). These effects are sizable, but some are imprecisely estimated. Combining all eight measures into a mean index suggests that a one s.d. increase in P reduces social capital by -0.340 s.d. and is significant at the 1% level.

While polarization has significant adverse effects, fractionalization has more muted and in some cases positive effects. For example, individuals in high-F villages are *more* likely to contribute to voluntary public goods (row 1) and to assist poorer neighbors (row 8). The other outcomes exhibit less clear patterns and noisier estimates. These findings are at odds with prior literature on diversity, trust, and public goods, which shows that F is associated with adverse outcomes. Our results differ in part

[44] Our estimates are based on individual-level specifications that control for predetermined covariates analogous to those in the individual-level regressions using Census data (i.e., gender, age and age squared). The 2012 *Susenas* is the only available data to study reported preferences in a large enough sample of Transmigration villages to yield reliable estimates. Even so, the estimates are relatively noisy with only 6 out of 16 being significant at conventional levels. This is due to limited statistical power, and not because of limited coverage over certain parts of the joint distribution of F and P.

because we are able to distinguish F from P.[45] That P rather than F undermines social capital is in line with our findings for language, marriage, and name choices.

Aggregate Outcomes. In Table 11, we further corroborate these lessons using village-level outcomes associated with integration. While some measures exhibit little variation, most results are in line with earlier findings, which suggest that F is conducive to nation building while P undermines this process.

Columns 1–3 show that F leads to more growth-enhancing public goods while P works against such investments. Column 1 considers a summary index of five public goods provided by village governments and recorded triennially in *Podes* data from 2002 to 2014: safe drinking water, garbage collection, public toilet facilities, 4-wheel road access, and streetlights.[46] The positive effects of F and negative effects of P on measured public goods are consistent with the individual responses in Table 10. Column 2 considers the share of the village with any visible nighttime lights in 2010, a proxy for local development (Henderson et al., 2012). A one s.d. increase in F (P) increases (reduces) light coverage by nearly one-third. Column 3 provides similar insights using a survey-based measure of mean household expenditures per capita, pooling annual *Susenas* data from 2000 to 2014. The outcome is in logs, and a one s.d. increase in F (P) increases (reduces) expenditures 6.7 (3.8) percent.

Consistent with these patterns of (under)development, columns 4 and 5 show that F reduces the likelihood of ethnic conflict while P increases it. Column 4 uses triennial *Podes* data from 2002 to 2014 covering all villages. Column 5 uses event-level data from the National Violence Monitoring System (*Sistem Nasional Pemantauan Kekerasan Indonesia* or *SNPK*), which covers incidents in high-conflict regions from 2000 to 2014. In both sources, ethnic conflict is a rare event, and the signs on F and P are similar. However, the estimates are larger and more precisely estimated for the media-reported events in *SNPK* (column 4) compared to events reported by the village head in *Podes* (column 5).[47]

We interpret these results through our model in Section 3 and the Esteban and Ray (2011) theory of ethnic conflict. In the latter, F amplifies conflict over private goods, and P amplifies conflict over public goods. Transmigration villages fostered equality in private access to land and housing. Our model implies benefits of intergroup contact in villages with many small groups (high F). With fewer reasons to fight over private resources, these benefits may be more salient in shaping interethnic interactions than in Esteban and Ray (2011). On the other hand, Transmigration villages still have a host of contestable public resources and institutions (subject to recurring elections). These "public prizes" may fuel interethnic antagonism, which drives the adverse effects of P in our model and in Esteban and Ray (2011).

Finally, columns 6 and 7 of Table 11 explore civic capital and support for inclusive, nationally-oriented political parties. Column 6 considers voter turnout in the first democratic election in 1999, recorded in *Podes* in the same year. Given such high turnout (95% in the mean village), there is little scope for diversity to matter. However, column 7 shows that P reduces support for political parties that embrace the Indonesian state ideology of *Pancasila* (see Section 2.1). The outcome, from *Podes* 2002 data, takes a value of one for villages where *Pancasila*-adhering parties finished first, second and third, and

[45] To be sure, regressions including F or P but not the other measure yield systematically negative estimates for both. For example, in row 3, F has a coefficient(std. error) of -0.369(0.191) on its own, and P has a coefficient of -0.501(0.184) on its own.

[46] These locally-provided public goods have more scope to vary over time and across villages than those provided by the Ministry of Transmigration in the 1980s (e.g., the number of schools and health clinics, see Section 2.2).

[47] The results from *Podes* are similar when restricted to the villages covered by *SNPK*: -0.010(0.010) for F and 0.005(0.009) for P.

zero otherwise. The effect size is meaningful—a 10% reduction in support for a one s.d. increase in P—though imprecise. Meanwhile, F also has negative effect, but it is smaller and even less precise.

Summary. Combined with our earlier results for *homeIndo*, the findings in Tables 8–11 provide new evidence on (i) how diversity shapes integration across ethnic groups and (ii) downstream consequences of integration for public goods, development, and conflict. By changing incentives to maintain one's ethnic identity, diversity has the potential to either undermine or reinforce this nation building process.

7.3 Mitigating Ethnic Divisions: National Language Use and Shared Identity

In this final section, we illustrate how the national language can mitigate ethnic divisions over the long-run. As more people speak Indonesian at home, the linguistic distance between ethnic groups falls, thereby reducing the effective polarization in society. Figure 6 plots the density of polarization, $P(\delta)$, across Transmigration villages, adjusted for *exogenous* native linguistic distances between groups as in Section 7.1. The dashed line presents another polarization measure, $P(\tilde{\delta})$, which uses the primary language spoken at home to compute *endogenous* linguistic distance between groups. Intuitively, if more people within a village speak the same language, then the effective polarization, $P(\tilde{\delta})$, between groups would be lower. This shift is evident in Figure 6 where the $P(\tilde{\delta})$ density lies to the left of the $P(\delta)$ density. A Kolmogorov-Smirnov test rejects the null that the two distributions are identical ($p < 0.001$). These differences are important. If everyone spoke their native language, there would be no shift. If only a few people from different groups spoke a common language, the shift would be much less pronounced.

Put differently, adoption of Indonesian at home is helping to integrate ethnic groups that would otherwise remain divided along deep linguistic cleavages inherited over many generations. While Figure 6 is based on a single cross-section in 2010, it hints at the possibility of national language use facilitating cultural integration across time. Indeed, this is the message of Table 2, which showed that children who grow up speaking Indonesian at home exhibit weaker attachment to their inherited ethnic identity and greater openness to integrating with those from other ethnic groups. Together, these results, and the nexus of findings in Section 6 are consistent with language being a key nation-building instrument.

8 Discussion

This paper offered new evidence on how intergroup contact shapes the nation-building process in diverse societies. We studied a large-scale resettlement program involving nearly two million voluntary migrants across more than 800 diverse new communities. Our findings illustrate two important dimensions of local diversity. With many small groups (high fractionalization), there are large returns to integrating through a common identity. With a few large groups (high polarization), intergroup antagonism and incentives for cultural dominance grow stronger, making coordination more difficult. These two forces shape numerous outcomes related to the nation-building process, including national language use at home, intermarriage, name choices for children, social capital, and public goods provision. Moreover, we find strong neighborhood effects of diversity and show that residential segregation undermines the benefits of F while mitigating the costs of P.

Beyond Indonesia, the distinct effects of F and P that we identify contribute to recent debates on migration and demographic change in both rich and poor countries. Several studies document potential economic benefits of migration-induced diversity (e.g., Alesina et al., 2016; Ashraf and Galor, 2013), while others emphasize the costs (e.g., Borjas, 2016). Our findings suggest a possible middle ground: F may increase the benefits while P may increase the costs. These results could inform the design of resettlement or housing policies where group composition is malleable. We further speak to the importance of a shared identity and national language to unite diverse groups. While we focus on primary identity choice, it would be interesting to explore the possibility of multiculturalism in future work.

From a policy perspective, the behavioral changes that we observe have important intergenerational implications for nation building. Although small, the mixed Transmigration communities may affect aggregate policy outcomes insomuch as local cultural change spills over onto the broader political environment (Giuliano and Nunn, 2013). Because Transmigration settlements arose at a critical juncture of development in these frontier areas of the country, it is possible that their impacts on cultural formation and evolution were quite sizable in the long-run (Bazzi et al., 2018). A growing literature on culture and institutions suggests potential channels for such persistence (see Alesina and Giuliano, 2015). This should be further explored in future work along with a rigorous investigation of spillovers.

The potential spillovers beyond Transmigration settlements are also important for understanding the legacy of this controversial resettlement program. While policymakers viewed Transmigration as a tool for nation building, critics accused the government of Javanese imperialism in the Outer Islands (Hoshour, 1997). Even today, popular accounts remain colored by egregious cases of failed integration.[48] However, Barter and Côté (2015) provide ethnographic evidence against this popular view, arguing that state-sponsored Transmigration communities were not associated with the ethnic violence that erupted in the Outer Islands in the 1990s. Ultimately, our findings offer support for this more sanguine view of the program. While some villages may have achieved limited integration over the long-run, this was but one possible outcome. For others, we find national integration of the sort one only sees in Indonesia's most diverse and vibrant cities. That such outcomes can also be realized in remote and underdeveloped rural areas is a testament to the importance of intergroup contact in the nation-building process.

[48] For example, Pisani (2014) details a visit to a particularly unsuccessful Transmigration settlement in the conflict-ridden province of Aceh in the 1990s: "But even where transmigrants rubbed along well enough with their neighbors, they carried on speaking their mother tongue, they cultivated crops they grew back home, they set up the gamelan gong orchestras that mirrored those of Java or Bali. It was more transplantation than transmigration, hardly a homogenizing force. . . . Transmigration was a rare failure in Suharto's nation building efforts." (pp. 36-7,). Similar anecdotes abound in the literature on the program and are part of broader concerns about "sons of the soil" conflict in Indonesia (Fearon and Laitin, 2011).

References

Abramitzky, R., L. P. Boustan, and K. Eriksson, "Cultural Assimilation during the Age of Mass Migration," *NBER Working Paper 22381*, 2018.

Advani, A. and B. Reich, "Melting pot or salad bowl: The formation of heterogeneous communities," Technical Report, Institute for Fiscal Studies 2015.

Alesina, A. and B. Reich, "Nation Building," *Unpublished Manuscript*, 2015.

_ **and E. LaFerrara**, "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, September 2005, *43* (3), 762–800.

_ **and E. Zhuravskaya**, "Segregation and the Quality of Government in a Cross-Section of Countries," *American Economic Review*, 2011, *101*, 1872–1911.

_ **and P. Giuliano**, "Culture and Institutions," *Journal of Economic Literature*, 2015, *53* (4), 898–944.

_ **, G. Tabellini, and F. Trebbi**, "Is Europe and Optimal Political Area?," *Brookings Papers on Economic Activity*, 2017, *BPEA Conference Drafts, March 23-24, 2017.*

_ **, J. Harnoss, and H. Rapoport**, "Birthplace Diversity and Economic Prosperity," *Journal of Economic Growth*, 2016, *21* (2), 101–138.

Algan, Y., C. Hémet, and D. D. Laitin, "The social effects of ethnic diversity at the local level: A natural experiment with exogenous residential allocation," *Journal of Political Economy*, 2016, *124* (3), 696–733.

Alisjahbana, S. T., *Indonesian language and literature: Two essays*, Yale University, Southeast Asia Studies, 1962.

Allport, G. W., *The nature of prejudice*, Cambridge, UK: Cambridge University Press, 1954.

Ananta, A., E. N. Arifin, M. S. Hasbullah, N. B. Handayani, and A. Pramono, "Changing ethnic composition: Indonesia, 2000-2010," in "XXVII IUSSP International Population Conference" 2013, pp. 26–31.

Anderson, B., *Imagined Communities*, Verso, 1983.

Anderson, T. W. and H. Rubin, "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 1949, *20* (1), 46–63.

Ashraf, Q. and O. Galor, "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *The American Economic Review*, 2013, *103* (1), 1–46.

Babcock, T., "Transmigration as a regional development strategy," in C. MacAndrews, ed., *Central Government and Local Development in Indonesia*, Oxford: Oxford University Press, 1986.

Bandiera, O., M. Mohnen, I. Rasul, and M. Viarengo, "Nation-Building Through Compulsory Schooling During the Age of Mass Migration," *Economic Journal*, forthcoming.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein, "Improving refugee integration through data-driven algorithmic assignment," *Science*, 2018, *359* (6373), 325–329.

Barbour, S. and C. Carmichael, *Language and nationalism in Europe*, OUP Oxford, 2000.

Barter, S. J. and I. Côté, "Strife of the soil? Unsettling transmigrant conflicts in Indonesia," *Journal of Southeast Asian Studies*, 2015, *46* (1), 60–85.

Bayer, P., S. L. Ross, and G. Topa, "Place of work and place of residence: Informal hiring networks and labor market outcomes," *Journal of Political Economy*, 2008, *116* (6), 1150–1196.

Bazzi, S., A. Gaduh, A. Rothenberg, and M. Wong, "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia," *American Economic Review*, 2016, *106* (9), 2658–2698.

_ **, M. Fiszbein, and M. Gebresilasse**, "Frontier Culture: Historical Roots and Persistence of 'Rugged Individualism' in the United States," *NBER Working Paper 23997*, 2018.

Bertrand, J., *Nationalism and ethnic conflict in Indonesia*, Cambridge, UK; New York: Cambridge University Press, 2004.

Bleakley, H. and A. Chin, "Age at arrival, English proficiency, and social assimilation among US immigrants," *American Economic Journal: Applied Economics*, 2010, *2*, 165.

Blouin, A. and S. Mukand, "Erasing Ethnicity? Nation Building, (Mis)Trust and the Salience of Identity in Rwanda," *Unpublished Manuscript*, 2016.

Blumer, H., "Race prejudice as a sense of group position," *Pacific Sociological Review*, 1958, *1* (1), 3–7.

Boisjoly, J., G. J. Duncan, M. Kremer, D. M. Levy, and J. Eccles, "Empathy or antipathy? The impact of diversity," *American Economic Review*, 2006, *96* (5), 1890–1905.

Borjas, G., *We Wanted Workers: Unraveling the Immigration Narrative*, W.W. Norton & Company, 2016.

Cameron, A. C., J. B. Gelbach, and D. L. Miller, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

_ , _ , and _ , "Robust Inference with Multiway Clustering," *Journal of Business & Economic Statistics*, 2011, *29* (2).

Chetty, R. and N. Hendren, "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects," *The Quarterly Journal of Economics*, 2018, *133* (3), 1107–1162.

Clauss, W., H.-D. Evers, and S. Gerke, "The Formation of A Peasant Society: Javanese Transmigrants in East Kalimantan," *Indonesia*, 1988, (46), 79–90.

Clingingsmith, D., A. I. Khwaja, and M. Kremer, "Estimating the impact of the Hajj: religion and tolerance in Islam's global gathering," *Quarterly Journal of Economics*, 2009, *124* (3), 1133–1170.

Clots-Figueras, I. and P. Masella, "Education, Language and Identity," *The Economic Journal*, 2013, *123* (570), F332–F357.

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

Cornes, R. and T. Sandler, *The theory of externalities, public goods, and club goods*, 2nd ed ed., Cambridge ; New York: Cambridge University Press, 1996.

Darity Jr., W. A., P. L. Mason, and J. B. Stewart, "The economics of identity: The origin and persistence of racial identity norms," *Journal of Economic Behavior & Organization*, July 2006, *60* (3), 283–305.

Dell, M. and P. Querubin, "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies," *Quarterly Journal of Economics*, forthcoming.

Desmet, K., I. Ortuño-Ortín, and S. Weber, "Linguistic Diversity and Redistribution," *Journal of the European Economic Association*, 2009, *7* (6), 1291–1318.

_ , J. Gomes, and I. Ortuño-Ortín, "The geography of linguistic diversity and the provision of public goods," *Unpublished Manuscript*, 2016.

Duggan, C., *The Force of Destiny: A History of Italy Since 1976*, Penguin Books, London, 2007.

Enos, R. D., *The space between us: Social geography and politics*, Cambridge University Press, 2017.

Esteban, J. and D. Ray, "Polarization, fractionalization and conflict," *Journal of Peace Research*, 2008, *45* (2), 163–182.

_ and _ , "Linking conflict to inequality and polarization," *American Economic Review*, 2011, *101* (4), 1345–1374.

_ and _ , "Conflict and Development," *Annual Review of Economics*, 2017, *9* (1).

_ , L. Mayoral, and D. Ray, "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, 2012, *102* (4), 1310–1342.

Esteban, J.M. and D. Ray, "On the measurement of polarization," *Econometrica*, 1994, pp. 819–851.

Fearon, J. D., "Ethnic and Cultural Diversity by Country," *Journal of Economic Growth*, 2003, *8* (2), 195–222.

__ **and D. D. Laitin**, "Sons of the soil, migrants, and civil war," *World Development*, 2011, *39*, 199–211.

Feith, H., *The decline of constitutional democracy in Indonesia*, 1st equinox ed ed., Jakarta: Equinox Pub, 1962.

Fouka, V., "Backlash: The Unintended Effects of Language Prohibition in US Schools after World War I," *Stanford Center for International Development Working Paper*, 2016, (591).

Frey, W. H., *Diversity Explosion: How New Racial Demographics are Remaking America*, Brookings Institution Press, 2014.

Fryer, R. G. and S. D. Levitt, "The Causes and Consequences of Distinctively Black Names," *Quarterly Journal of Economics*, 2004, *119* (3), 767–805.

Ginsburgh, V. and S. Weber, "The economics of language," 2018.

Giuliano, P. and N. Nunn, "The transmission of democracy: from the village to the nation-state," *American Economic Review*, 2013, *103* (3), 86–92.

__ **and** __ , "Understanding cultural persistence and change," *Harvard Universtity, mimeo*, 2018.

Gordon, M. M., *Assimilation in American life: The role of race, religion, and national origins*, New York: Oxford University Press, 1964.

Guariso, A. and T. Rogall, "Rainfall inequality, political power, and ethnic conflict in Africa," 2017.

Hansen, L. P., "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Hanson, G. and C. McIntosh, "Is the Mediterranean the New Rio Grande? US and EU Immigration Pressures in the Long Run," *The Journal of Economic Perspectives*, 2016, *30* (4), 57–81.

Hardjono, J., "The Indonesian Transmigration Program in Historical Perspective," *International Migration*, 1988, *26* (4), 427–439.

Henderson, J. V., A. Storeygard, and D. N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, 2012, *102* (2), 994–1028.

Hoey, B. A., "Nationalism in Indonesia: Building Imagined and Intentional Communities Through Transmigration," *Ethnology*, 2003, *42* (2), 109–126.

Hoshour, C. A., "Resettlement and the Politicization of Ethnicity in Indonesia," *Bijdragen tot de Taal-Land-en Volkenkunde*, 1997, (4de Afl), 557–576.

Huntington, S. P., *Who are We?: America's Great Debate*, Free Press, 2004.

Iannaccone, L. R., "Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives," *The Journal of Political Economy*, April 1992, *100* (2), 271–291.

Janvry, A. De, K. Emerick, M. Gonzalez-Navarro, and E. Sadoulet, "Delinking Land Rights from Land Use: Certification and Migration in Mexico," *Unpublished Manuscript*, 2012.

Kebschull, D., *Transmigration in Indonesia: An Empirical Analysis of Motivation, Expectations and Experiences*, Hamburg, Germany: Transaction Publishers, 1986.

Kramsch, C. and H. G. Widdowson, *Language and Culture*, Oxford University Press, 1998.

Laitin, D. and R. Ramachandran, "Linguistic Diversity, Official Language Choice and Nation Building: Theory and Evidence," *Unpublished Manuscript*, 2015.

Laitin, D. D., *Language repertoires and state construction in Africa*, Cambridge University Press, 2007.

Lazear, E., "Culture and Language," *Journal of Political Economy*, 1999, *107*, s95–s126.

Liu, A. H., *Standardizing diversity: the political economy of language regimes*, Philadelphia: University of Pennsylvania Press, 2015.

Logan, T. D. and J. M. Parman, "The national rise in residential segregation," *The Journal of Economic History*, 2017, *77* (1), 127–170.

Lowe, M., "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration," *Unpublished Manuscript*, 2018.

Miguel, E., "Tribe or Nation?: Nation Building and Public Goods in Kenya versus Tanzania," *World Politics*, 2004, *56* (3), 327–362.

Miller, R., "The Development of European Identity/Identities: Unfinished Business," *A Policy Review*, 2012.

Montalvo, J. G. and M. Reynal-Querol, "Ethnic Polarization, Potential Conflict and Civil War," *American Economic Review*, 2005, *95* (3), 796–816.

Montgomery, J. D., "Intergenerational cultural transmission as an evolutionary game," *American Economic Journal: Microeconomics*, 2010, *2* (4), 115–36.

Okunogbe, O., "Does Exposure to Other Ethnic Groups Promote National Integration? Evidence from Nigeria," *Unpublished Manuscript*, 2015.

Paauw, S., "One land, one nation, one language: An analysis of Indonesia's national language policy," *University of Rochester Working Papers in the Language Sciences*, 2009, *5* (1), 2–16.

Paluck, E. L., S. A. Green, and D. P. Green, "The contact hypothesis re-evaluated," *Behavioural Public Policy*, 2018, pp. 1–30.

Pew Research Center, "What It Takes to Truly Be 'One of Us'," *Washington: Pew Research Center*, February 2017.

Pisani, E., *Indonesia, Etc.: Exploring the Improbable Nation*, WW Norton & Company, 2014.

Putnam, R. D., "E Pluribus Unum: Diversity and Community in the Twenty-first Century The 2006 Johan Skytte Prize Lecture," *Scandinavian Political Studies*, 2007, *30* (2), 137–174.

Rao, G., "Familiarity does not breed contempt: Diversity, discrimination and generosity in Delhi schools," *American Economic Review*, forthcoming.

Ricklefs, M. C., *A history of modern Indonesia since c. 1200*, 4th ed ed., Stanford, Calif: Stanford University Press, 2008.

Rigg, J., *Southeast Asia (Routledge Revivals): A Region in Transition*, Routledge, 2013.

Sanderson, E. and F. Windmeijer, "A weak instrument F-test in linear IV models with multiple endogenous variables," *Journal of Econometrics*, 2016, *190* (2), 212–221.

Sandholm, W. H., *Population games and evolutionary dynamics*, MIT press, 2010.

_ , "Population games and deterministic evolutionary dynamics," in "Handbook of game theory with economic applications," Vol. 4 2015, pp. 703–778.

Schelling, T. C., "Dynamic models of segregation," *Journal of Mathematical Sociology*, 1971, *1* (2), 143–186.

Simpson, A., "Indonesia," in "Language and national identity in Asia" 2007.

_ , *Language and national identity in Africa*, Oxford University Press, 2008.

Sneddon, J.N., *The Indonesian Language: Its History and Role in Modern Society*, University of New South Wales Press, 2003.

Tanasaldy, T., *Regime change and ethnic politics in Indonesia; Dayak politics of West Kalimantan*, Brill, 2012.

Thornton, D. L., *Javanization of Indonesian politics*, University of British Columbia, 1972.

Tirtosudarmo, R., "Transmigration and its centre-regional context: the case of Riau and South Kalimantan provinces, Indonesia," 1990.

Weber, E., *Peasants into Frenchmen: the modernization of rural France, 1870-1914*, Stanford University Press, 1976.

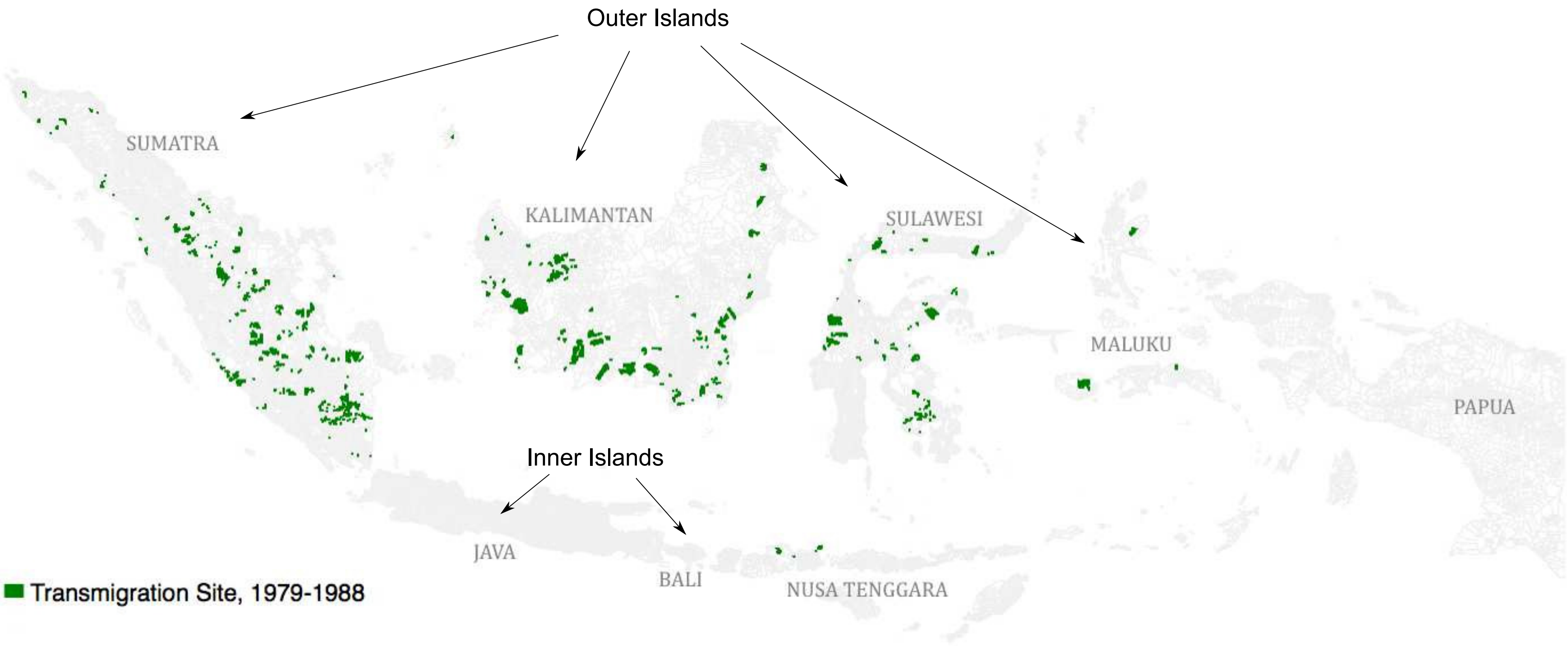
World Bank, *Indonesia: The Transmigration Program in Perspective* 1988.

Young, A., "Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections," *Unpublished Manuscript*, 2016.

Young, H. P., "The evolution of social norms," *Annual Review of Economics*, 2015, *7* (1), 359–387.

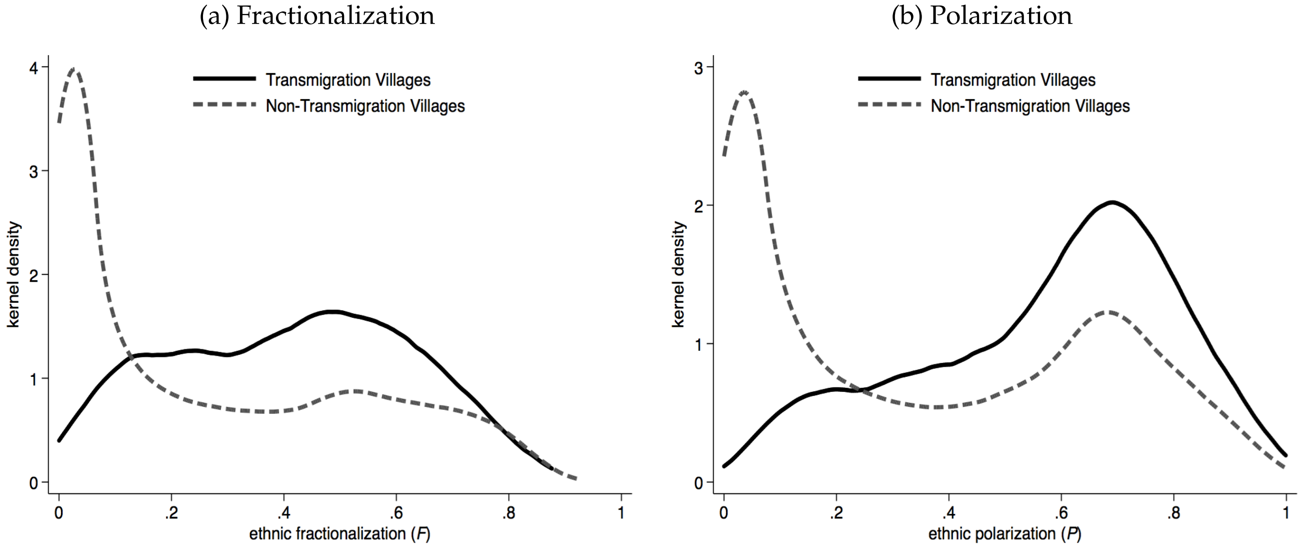
Figures

Figure 1: Map of Transmigration Villages



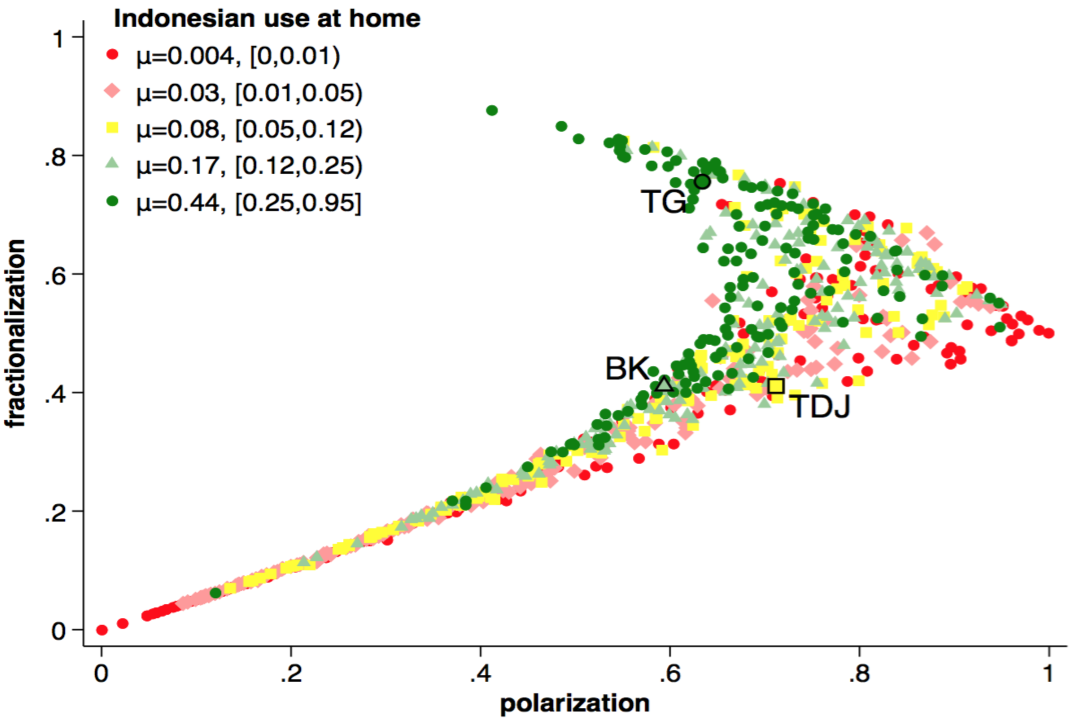
Notes: Each colored location on the map corresponds to a Transmigration village settled between 1979 and 1988. The white areas outlined in grey are other villages.

Figure 2: Ethnic Diversity in Transmigration and Non-Transmigration Villages



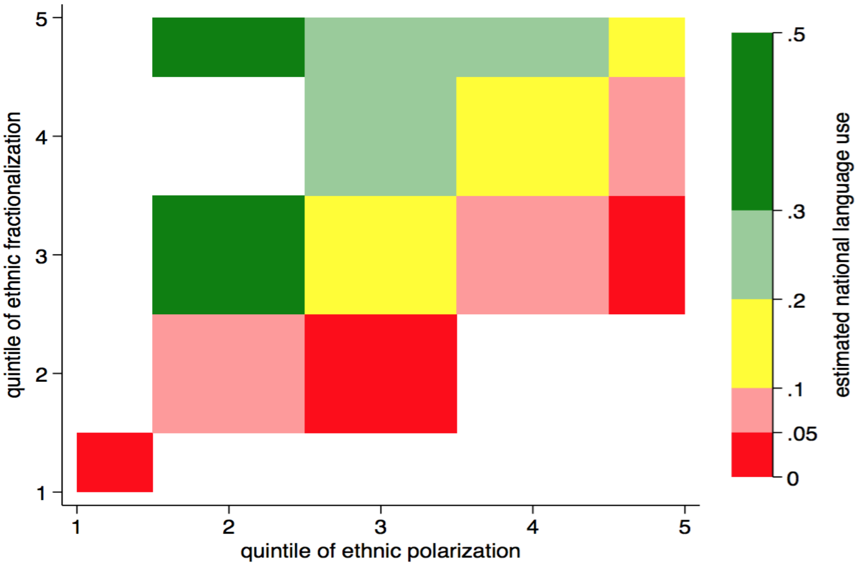
Notes: This figure plots the kernel density of ethnic (a) fractionalization and (b) polarization in 2010 for Transmigration villages and non-Transmigration villages in the Outer Islands. For both densities, we employ an Epanechnikov kernel and rule-of-thumb bandwidth.

Figure 3: Fractionalization, Polarization, and Indonesian Use at Home



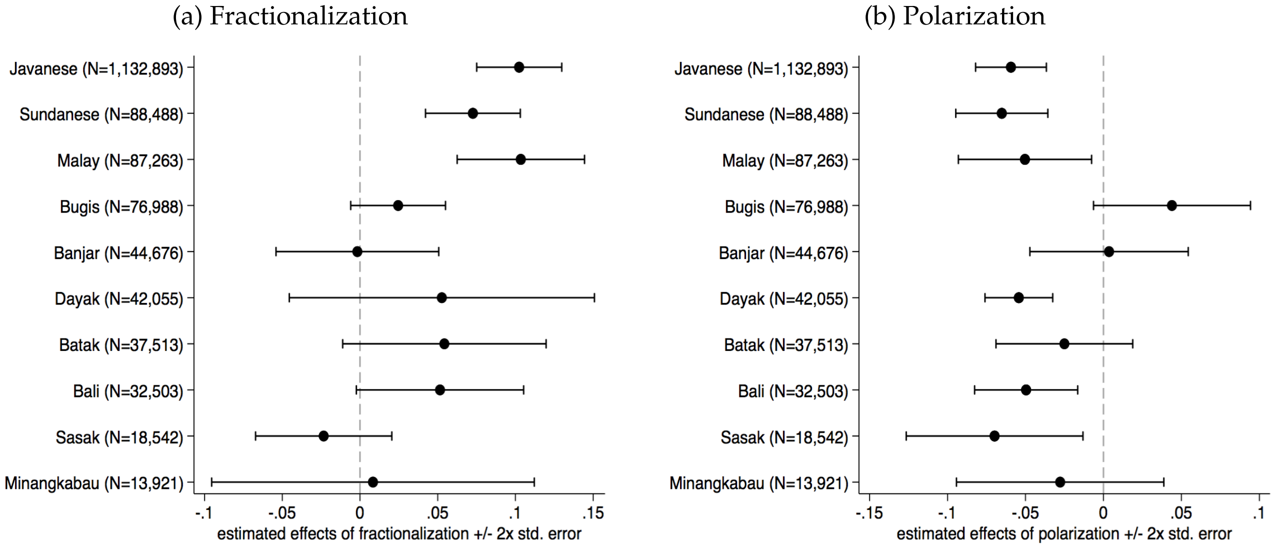
Notes: Each circle corresponds to a Transmigration village settled between 1979 and 1988. The villages are grouped into quintiles of average Indonesian use at home with μ indicating the mean and $[.,.)$ indicating the range within-quintile. See Section 5.3 for a discussion of the three case-study villages: TG is the village of Tanjung Gading, BK is Bukit Kemuning, and TDJ is Tri Dharma Wirajaya.

Figure 4: Flexibly-Estimated Effects of Diversity on Indonesian Use at Home



Notes: The figure plots the predicted national language use implied by estimating equation (8) with a full set of interactions between indicators for quintiles of fractionalization and quintiles of polarization. That is, we take the mean national language use of 0.036 in the omitted category (bottom quintiles of fractionalization and polarization) and add the $\widehat{\theta}_{ij}$ coefficient on the given interaction of quintile i of fractionalization and quintile j of polarization. The θ_{ij} estimates are in Appendix Table A.5. The quintiles of fractionalization are (1) $F \in [0, 0.196)$, (2) $F \in [0.197, 0.355)$, (3) $F \in [0.357, 0.488)$, (4) $F \in [0.488, 0.608)$, and (5) $F \in [0.608, 0.877]$. The quintiles of polarization are (1) $P \in [0, 0.351)$, (2) $P \in [0.356, 0.554)$, (3) $P \in [0.555, 0.673)$, (4) $P \in [0.673, 0.762)$, and (5) $F \in [0.763, 0.999]$. The white space indicates cells with no observations. Ten out of 25 potential ij cells are not represented as, for example, there are no villages in the first quintile of fractionalization ($i = 1$) and the top quintile of polarization ($j = 5$). The size of the squares/rectangles are arbitrary and connotes no additional information.

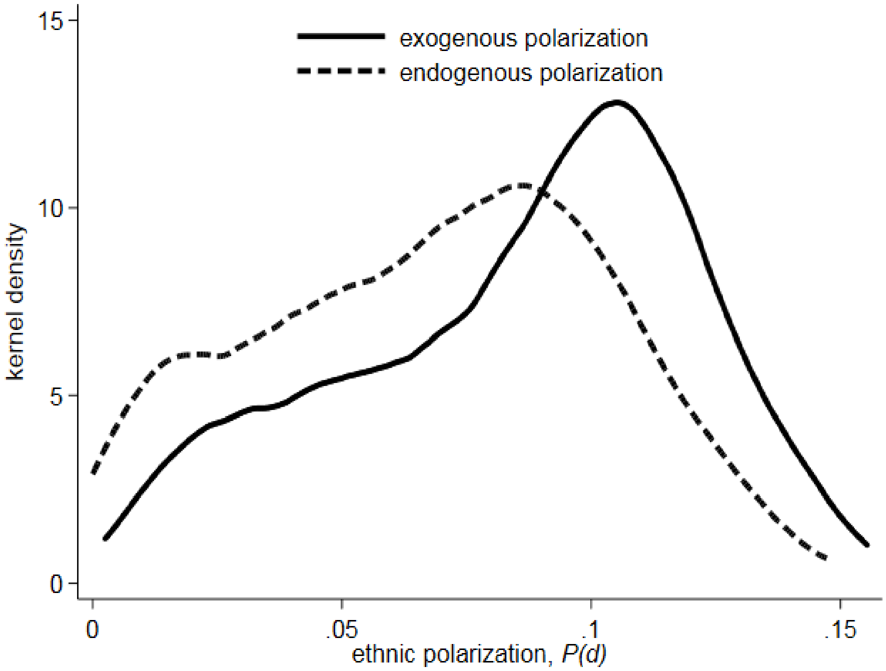
Figure 5: Effects of Diversity on Indonesian Use at Home by Ethnicity


(a) Fractionalization

(b) Polarization

Notes: This figure plots the standardized effects of (a) fractionalization and (b) polarization for the individual-level specification in column 6 of Table 3 estimated separately by ethnic group. We report results for the top 10 largest ethnicities with greater than 12,000 people across the 817 Transmigration villages. We group the Malay, Dayak and Batak sub-ethnicities into their broader ethnic groups. The graph reports point estimates $+/- 2 \times$ standard-error bars.

Figure 6: Shared Language Use and Effective Polarization



Notes: The figure plots kernel density estimates for exogenous and endogenous ethnic polarization, $P(d)$, where d captures the linguistic distance between groups (see Section 7.1). In the exogenous case, $d = \delta$ is based on predetermined linguistic classifications of each ethnic group's native language, and the endogenous measure is based on the actual language spoken at home by each member of each ethnic group. When individuals from two different groups i and j speak the same language at home, they are deemed to have zero linguistic distance. Formally, $\delta = 1 - \left(\frac{\text{branch}_{ij}}{\max(\text{branch}_i, \text{branch}_j)}\right)^{\kappa}$ where the ratio in parentheses captures the fraction of possible shared branches on linguistic classification trees from the *Ethnologue* database, and $\kappa = 0.5$ here, but the results are similar for $\kappa = 0.05$. Meanwhile, in the endogenous case, $d = \tilde{\delta}$ captures the linguistic distance between groups based on actual languages spoken at home. By construction, $\tilde{\delta}$ is must be weakly smaller than δ; if individual ℓ_i from group i and individual ℓ_j from group j speak the same language, then $\delta_{ij} = \tilde{\delta}_{\ell_i \ell_j}$ (see Appendix D.4). Polarization in village v is then defined as $P_v(d) = \sum_{i=1}^{I_v} \sum_{i \neq j}^{I_v} p_{iv}^2 p_{jv} d_{ij}$ where p captures the ethnic group shares in the village and $d = \delta$ for the solid, exogenous polarization distribution and $d = \tilde{\delta}$ for the dashed, endogenous polarization distribution.

Tables

Table 2: Nation Building and Language Use at Home

	Dep. Var. as Adult in 2014:			
	Speaks Indonesian at Home	Changes Ethnicity from 1997	In Interethnic Marriage	Trust Other Ethnic Groups (z-score)
	(1)	**(2)**	**(3)**	**(4)**
	Panel A: Baseline			
Indonesian was Primary Language at Home as Child in 1997	0.156	0.062	0.053	0.148
	(0.022)	(0.019)	(0.023)	(0.054)
	Panel B: Adding Parental Intermarriage			
Indonesian was Primary Language at Home as Child in 1997	0.151	0.045	0.046	0.131
	(0.022)	(0.019)	(0.023)	(0.054)
Parents from Different Ethnic Groups	0.053	0.177	0.092	0.160
	(0.021)	(0.030)	(0.031)	(0.055)
Number of Individuals	8,623	6,594	5,628	8,236
Dependent Variable Mean	0.369	0.114	0.103	0.00
Age, Gender, Education Fixed Effects	Yes	Yes	Yes	Yes
Village Fixed Effects	Yes	Yes	Yes	Yes

Notes: This table reports estimates of the correlation between parental daily Indonesian language use at home as a child in 1997 and the given column's dependent variable for individuals in the 2014 round of the *Indonesia Family Life Survey*. Panel B augments the baseline specification with a control for whether the child's parents hail from different ethnic groups. The sample is restricted to all individuals greater than 15 years old who live in a different household in 2014 compared to 1997. The dependent variables include in column (1) an indicator for whether the individual used the Indonesian language at home on a regular basis in 2014, (2) an indicator for whether the individual switched his/her reported ethnicity between 1997 and 2014, (3) an indicator for whether a married individual is in an interethnic marriage in 2014, (4) an index normalized to have mean zero and standard deviation one based on ordered response on a 4 point scale to the question "Do you trust people from other ethnic groups less than you trust your people from own group?". Note that the language use at home variable is distinct from the 2010 Population Census measure used elsewhere in the paper, which only lists a single, primary language at home as opposed to listing all languages used at home. All specifications include the fixed effects listed at the bottom of the table where the age FE are for each individual age. Standard errors are clustered at the village level of which there are around 1,300 across columns.

Table 3: Ethnic Diversity and National Language Use At Home

	Dep. Var.: National Language Use at Home					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Village-Level			Individual-Level		
ethnic fractionalization	0.296		0.637	0.671	0.499	0.377
	(0.041)		(0.073)	(0.075)	(0.057)	(0.051)
ethnic polarization		0.086	-0.362	-0.392	-0.302	-0.184
		(0.030)	(0.051)	(0.057)	(0.041)	(0.038)
Number of Villages	817	817	817	817	817	817
Number of Individuals	–	–	–	1,800,499	1,800,499	1,800,499
Dependent Variable Mean	0.144	0.144	0.144	0.154	0.154	0.154
R^2	0.379	0.303	0.437	0.114	0.221	0.280
Island FE, Predetermined Controls (**x**)	✓	✓	✓	✓	✓	✓
Ethnicity, Age, Relation, Gender FE					✓	✓
Birth District, Current District FE						✓

Notes: This table reports estimates of equation (7) where fractionalization and polarization are defined using the self-reported ethnicities in the 2010 Population Census. Columns 1–3 are village-level regressions where the dependent variable is the share of individuals that report Indonesian as their main language at home in 2010. Columns 4–6 are individual-level regressions where the dependent variable is an indicator equal to one if the individual reports Indonesian as their main language at home in 2010. All columns include our baseline set of predetermined controls (**x**) described in Appendix D: log village area, three measures of village slope, a ruggedness index, log altitude, three measures of soil quality, two measures of soil texture, two measures of soil drainage, mean rainfall and temperature from 1948 to 1978, distance to nearest point in Java/Bali, distance to the nearest pre-1979 road, distance to the coast, distance to the nearest river, distance to the subdistrict and district capital, and four island fixed effects. Column 5 includes exhaustive fixed effects for individual ethnicity, age, relation to household head, and gender. Column 6 further includes fixed effects for birth district and current district. The dependent variable means are across all villages in columns 1–3 and across all individuals in column 4–6. Standard errors are clustered by district, of which there are 84.

Table 4: Instrumental Variables Estimates of Diversity and Indonesian Use At Home

	Dep. Var.: National Language Use at Home			
	(1)	**(2)**	**(3)**	**(4)**
	village-	individual-level		
ethnic fractionalization	1.017	0.726	0.599	0.592
	(0.095)	(0.073)	(0.079)	(0.052)
ethnic polarization	-0.793	-0.547	-0.447	-0.420
	(0.095)	(0.061)	(0.051)	(0.046)
Number of Villages	817	817	817	817
Number of Individuals	–	1,800,499	1,800,499	1,800,499
Dependent Variable Mean	0.145	0.154	0.154	0.154
SW fractionalization, p-value	0.000	0.000	0.000	0.000
SW polarization, p-value	0.000	0.000	0.000	0.000
KP Wald stat	7.8	8.7	10.1	22.5
Hansen J test, p-value	0.607	0.253	0.411	0.470
Hausman GMM test OLS=IV, p-value	0.372	0.807	0.747	0.769
Island FE, x Predetermined Controls	✓	✓	✓	✓
Ethnicity, Age, Relation, Gender FE			✓	✓
Birth District, Current District FE				✓

Notes: This table estimates instrumental variables regressions for the village- and individual-level specifications in columns 3 and 4–6, respectively, of Table 3. The instruments include (i) dummies for each ventile of the number of transmigrants from Java/Bali in the initial year of settlement and (ii) the share of each of 15 (out of 16) Inner-Island ethnic groups among all those born in Java/Bali before the year of settlement. The latter are based on the 2000 Population Census and measure, for example, the share of Javanese in the total population of Inner-Island ethnics born in Java/Bali before the year of settlement. We estimate the 2SLS equations using generalized method of moments (GMM) given the many instruments. The null hypotheses of (i) the Sampson-Windmeijer (SW) test is that the instruments for the given endogenous variable are weak, (ii) the Hansen J test is that the instruments are uncorrelated with the error term and correctly excluded from the second stage, and (iii) the Hausman GMM test is that the OLS estimates equal the IV estimates. The Kleibergen-Paap (KP) Wald statistic is a multivariate generalization of the first-stage F statistic. Standard errors in all columns are clustered by district, of which there are 84.

Table 5: Effects of Diversity on Sub-populations within Transmigration Villages

	Dep. Var.: Individual Speaks National Language at Home							
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Sample:	baseline	inner-ethnic	inner-born $<$ yr. settled	outer-ethnic	outer-born $<$ yr. settled	outer-born APPDT	outer-born non-APPDT	born same district $\geq$ yr. settled
ethnic fractionalization	0.082	0.098	0.081	0.056	0.069	0.069	0.056	0.098
	(0.011)	(0.013)	(0.012)	(0.015)	(0.015)	(0.020)	(0.015)	(0.014)
ethnic polarization	-0.040	-0.058	-0.053	-0.028	-0.024	-0.035	0.001	-0.046
	(0.008)	(0.010)	(0.010)	(0.012)	(0.011)	(0.012)	(0.014)	(0.011)
Number of Individuals	1,800,499	1,267,946	543,655	532,486	408,751	282,030	126,721	626,772
Dependent Variable Mean	0.154	0.099	0.066	0.285	0.207	0.158	0.316	0.168
R^2	0.281	0.198	0.143	0.328	0.299	0.305	0.283	0.300

Notes: This table estimates the full fixed effects, individual-level specification from column 6 of Table 3 for different population subsamples of the Transmigration village in 2010. Column 1 reproduces the standardized estimates from column 6 of Table 3 for reference. The subsequent columns restrict the sample to individuals: (2) reporting an ethnicity native to the Inner Islands of Java/Bali, (3) born in Java/Bali before the year of settlement, (4) reporting an ethnicity native to the Outer Islands, (5) born in the Outer Islands before the year of settlement, (6) born in the Outer Islands in the same district or a neighboring one in the same province before the year of settlement, (7) born in the Outer Islands in in a different province before the year of settlement, and (8) born in the same district after the year of settlement. While mean Indonesian use at home differs across columns, the baseline fixed effects in column 1, also used in every column thereafter, make it possible to compare standardized estimates across columns. Standard errors in all columns are clustered by district, of which there are 84.

Table 6: Intergroup Contact, Segregation, and National Language Use at Home

	Dep. Var.: Individual Speaks National Language at Home					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
ethnic fractionalization, contiguous settlements	0.054				-0.006	
	(0.014)				(0.014)	
ethnic fractionalization, village		0.082			0.021	0.084
		(0.011)			(0.010)	(0.010)
ethnic fractionalization, neighborhood			0.129		0.098	
			(0.008)		(0.009)	
ethnic polarization, contiguous settlements	-0.026				0.000	
	(0.009)				(0.010)	
ethnic polarization, village		-0.040			-0.011	-0.031
		(0.008)			(0.009)	(0.008)
ethnic polarization, neighborhood			-0.064		-0.055	
			(0.008)		(0.009)	
2 out of 2 next-door neighbors of different ethnicity				0.192	0.146	
				(0.010)	(0.008)	
1 out of 2 next-door neighbors of different ethnicity				0.067	0.035	
				(0.006)	(0.003)	
ethnic segregation						-0.029
						(0.005)
Number of Individuals	1,758,030	1,758,030	1,758,030	1,758,030	1,758,030	1,758,030
Dependent Variable Mean	0.154	0.154	0.154	0.154	0.154	0.154
R^2	0.276	0.282	0.301	0.301	0.316	0.285

Notes: This table estimates individual-level regressions with different measures of diversity and segregation. The specification is based on column 6 of Table 3. The sample size is slightly smaller as the measures of neighbor ethnicity in column 4 are unavailable for a small number of households, and we want to ensure a constant sample across columns. Column 1 measures diversity at the level of contiguous Transmigration villages. While 254 Transmigration villages are isolated villages, the remainder are part of settlement blocs containing 2–18 villages with half of those containing 2–4. Column 2 is the baseline with own-village-level diversity. Column 3 measures diversity at the sub-village, neighborhood level, of which there are as many as 59 with the median village having 15. Column 4 measures diversity at the level of immediate neighbors in housing units adjacent to one's own. Column 5 includes all measures simultaneously. Column 6 augments the baseline specification in column 2 with a summary measure of ethnic residential segregation proposed in Alesina and Zhuravskaya (2011). Standard errors in all columns are clustered by district, of which there are 84.

Table 7: Intergroup Distance Mechanisms

Dep. Var.: National Language Use at Home	Distance Between Groups					
	Spatial		Economic		Linguistic	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
fractionalization	0.134	0.145	0.149	0.165		
	(0.015)	(0.015)	(0.015)	(0.019)		
polarization	-0.068	-0.084	-0.073	-0.102		
	(0.011)	(0.012)	(0.012)	(0.016)		
segregation	-0.033	-0.031				
	(0.006)	(0.006)				
fractionalization × segregation		-0.041				
		(0.010)				
polarization × segregation		0.018				
		(0.007)				
interethnic inequality, agroclimatic similarity			-0.033	-0.019		
			(0.009)	(0.012)		
fractionalization × interethnic inequality				-0.036		
				(0.015)		
polarization × interethnic inequality				0.012		
				(0.011)		
fractionalization, 44 groups					0.129	
					(0.018)	
polarization, 44 groups					-0.079	
					(0.015)	
fractionalization(δ), linguistic distance						0.144
						(0.016)
polarization(δ), linguistic distance						-0.092
						(0.013)
Number of Villages	817	817	816	816	817	816
Dependent Variable Mean	0.144	0.144	0.145	0.145	0.144	0.145
R^2	0.465	0.486	0.453	0.461	0.407	0.439
H_0: $F(44) = F$ baseline, p-value					[0.669]	
H_0: $P(44) = P$ baseline, p-value					[0.716]	
H_0: $F(\delta) = F$ baseline, p-value						[0.100]
H_0: $P(\delta) = P$ baseline, p-value						[0.022]

Notes: This table estimates regressions based on augmenting the main, village-level regression in column 3 of Table 3. All variables are standardized (prior to interacting) for ease of interpretation. Column 1 estimates the village-level analogue of column 6 in Table 6, and column 2 adds the interaction of segregation with our village-level diversity measures. Columns 3 and 4 consider interethnic inequality in agroclimatic similarity (as a proxy for human capital endowments) among initial migrants to the village. This between-group inequality measure is akin a standard Greenberg-Gini index. These columns also control for average and overall inequality in agroclimatic similarity at the village level. Column 5 redefines the diversity measures based on an aggregation of the 1,330 self-reported ethnicities into 44 broad ethnic groups determined by Indonesian demographers (Ananta et al., 2013). Column 6 adjusts the baseline diversity measures based on the linguistic distance (δ) between ethnic groups according to their native ethnic languages (see Section 7.1). We lose one observation in columns 3–4 and 6 due to merging difficulties with the underlying agroclimatic and linguistic distance data, respectively. The bracketed p-values at the bottom of the table for columns 5 and 6 are based on a test of the difference in coefficients between the given diversity measure and the baseline measure from column 3 of Table 3. Standard errors in all columns are clustered by district, of which there are 84.

Table 8: Diversity and Intermarriage

	Dep. Var.: post-settlement intermarriage rate in			
	2000	2010	2000	2010
	actual		supply-adjusted	
	(1)	**(2)**	**(3)**	**(4)**
ethnic fractionalization	0.068	0.093	-0.025	-0.006
	(0.012)	(0.008)	(0.022)	(0.013)
ethnic polarization	-0.028	-0.027	-0.081	-0.112
	(0.010)	(0.007)	(0.021)	(0.012)
Number of Villages	815	816	815	816
Dependent Variable Mean	0.152	0.179	0.388	0.482
R^2	0.258	0.560	0.114	0.317

Notes: This table estimates the baseline village-level regression from column 3 of Table 3 for interethnic marriage outcomes as observed in the 2000 and 2010 Population Census. The intermarriage rates are defined over all self-reported ethnicities by husbands and wives within a household and are restricted to those that were younger than 15 years old (or not yet born) by the year of settlement and hence plausibly married after arriving in the given Transmigration village. The diversity measures are standardized and based on the given year of the outcome listed at the top of each column. Columns 1 and 2 are the actual intermarriage rates. Columns 3 and 4 take the actual intermarriage rates and divide by the potential intermarriage rate (i.e., "supply-adjusted"), which is based on randomly matching the (young) married men and women 10,000 different times and taking the average. The sample size is smaller by a few villages in this table, two in 2000 and one in 2010, due to missing marriage data for the young cohort. Standard errors in all columns are clustered by district, of which there are 84.

Table 9: Diversity and the Identity Content of Children's Names

	Dep. Var.: precision of name in identifying . . .			
	Indonesian language home	intermarried household	urban household	own-ethnicity
	(1)	**(2)**	**(3)**	**(4)**
ethnic fractionalization	0.222	0.196	0.268	-0.215
	(0.038)	(0.041)	(0.052)	(0.042)
ethnic polarization	-0.127	-0.113	-0.161	0.081
	(0.032)	(0.034)	(0.044)	(0.032)
Number of Individuals	726,969	676,307	731,628	720,142
R^2	0.101	0.190	0.080	0.101

Notes: This table estimates the baseline individual-level regression from column 5 of Table 3 for the precision of children's names (from the 2010 Population Census) in identifying whether they belong to the given identity type listed at the top of the column. We restrict the sample to children born after the year of settlement for the given village. The index captures whether name n is likely to belong to identity type g, being in column (1) a household where the modal member speaks Indonesian at home, (2) a household with the head and spouse being of a different ethnicity, (3) an individual residing in urban areas, (4) an individual belonging to his/her native ethnicity. These measures are estimated for all individuals living outside of Transmigration villages elsewhere in Indonesia. See equation (9) and Appendix D.2 for details on the construction of these indices. We restrict the sample to children with names that are observed for at least 100 other people in Indonesia (population 234 million) to deal with unique names as in Fryer and Levitt (2004). Standard errors in all columns are clustered by district, of which there are 84.

Table 10: Diversity and Social Capital

	standardized coefficients			
	fractionalization	**polarization**	Dep. Var. Mean	No. of
Dependent Variable	(F)	(P)	(1-4 scale)	individuals
1. voluntary public good provision	0.166 (0.113)	-0.224 (0.119)	2.7	834
2. join community group(s)	0.017 (0.129)	-0.068 (0.106)	2.4	820
3. pleased with non-coethnics	0.106 (0.189)	-0.285 (0.167)	2.9	840
4. trust neighbor to watch house	0.145 (0.120)	-0.242 (0.100)	2.9	840
5. trust neighbor to tend children	-0.080 (0.149)	-0.120 (0.124)	2.7	840
6. feel safe	-0.077 (0.107)	-0.202 (0.099)	3.2	850
7. ease in obtaining neighbor assistance	0.005 (0.121)	-0.120 (0.104)	2.7	850
8. contribute to assist unfortunate neighbors	0.227 (0.097)	-0.199 (0.113)	2.9	850

Notes: This table estimates an individual-level regression using the sociocultural module of the 2012 National Socioeconomic Survey (*Susenas*). The survey covers 87 Transmigration villages with up to 850 household heads responding to the 8 questions listed in shorthand statements in each row of the table denoting a separate regression. See Appendix D.2 for the fully elaborated questions. Responses to these questions are given on a 1 to 4 integer scale, and we re-order the responses such that higher numbers indicate stronger agreement. The sample varies slightly across outcomes due to non-responses, though these are not systematic with respect to diversity. The dependent variable and the two diversity indices are standardized, leaving the coefficients in beta form for ease of interpretation. The specification is otherwise identical to those in prior tables including controls for the predetermined village-level covariates ($\mathbf{x}$) and island fixed effects as well as age, age squared and a gender dummy. Standard errors in all columns are clustered by district, of which there are 45.

Table 11: Diversity and Village-Level Outcomes: Public Goods, Development, Conflict, and Voting

Dependent Variable:	**local development and public goods**			**conflict**		**voting**	
	village	light	household	any ethnic conflict		turnout	*Pancasila*
	pub. goods	intensity	exp./capita	*SNPK*	*Podes*		party 1st-3rd
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
ethnic fractionalization	0.030	0.026	0.067	-0.062	-0.005	-0.001	-0.022
	(0.011)	(0.015)	(0.033)	(0.028)	(0.004)	(0.006)	(0.032)
ethnic polarization	-0.022	-0.025	-0.038	0.066	0.004	-0.003	-0.045
	(0.011)	(0.014)	(0.036)	(0.028)	(0.004)	(0.007)	(0.031)
Number of Villages	817	817	710	244	817	795	817
Dependent Variable Mean	0.412	0.082	12.489	0.045	0.010	0.947	0.470
R^2	0.227	0.109	0.124	0.316	0.028	0.092	0.106

Notes: This table estimates the baseline village-level regression from column 3 of Table 3 for several outcomes measured from 2000–14: column (1) an index taking the mean of five binary indicators for whether the village has a given village-provided public good, including provision of safe drinking water, garbage collection, public toilet facilities, 4-wheel road access, and streetlights on the main road as reported in the 2002 *Podes*; (2) the share of village area covered with any nighttime lights in 2010; (3) the log of mean village-level household expenditures per capita averaged across all available rounds of the National Socioceconomic Survey (*Susenas*) from 2000 to 2012, which covers a subset of Transmigration villages in at least one of those years; (4) a binary indicator for any ethnic conflict in the village from 2000 to 2014 as reported in the SNPK violence database, which only covers a subset of Indonesian provinces; (5) a binary indicator for any ethnic conflict in the village in 2002, 2005, 2008, 2011, or 2014 as reported by the village head in the *Podes* data from those years; (6) the share of the voting-age population that voted in the first democratic election in 1999 as reported in *Podes* 1999; (7) a binary indicator for whether the parties finishing in the top 3 in terms of vote shares in the 1999 national legislative election adhered to a platform based on the inclusive, nationalist ideology of the Indonesian state known as *Pancasila*. For measure (3), given the arbitrary sampling variation across *Susenas* rounds, some Transmigration villages are covered more than once while others are never covered across these 15 years. The sample in column 6 is slightly smaller than in column 7 because the measures come from different rounds of *Podes* in 1999 and 2002, respectively, and the former introduced difficulties merging to the baseline sample of Transmigration villages. The diversity indices are standardized and based on the 2000 Population Census as most outcomes are measured before 2010. Standard errors in all columns are clustered by district, of which there are 84.

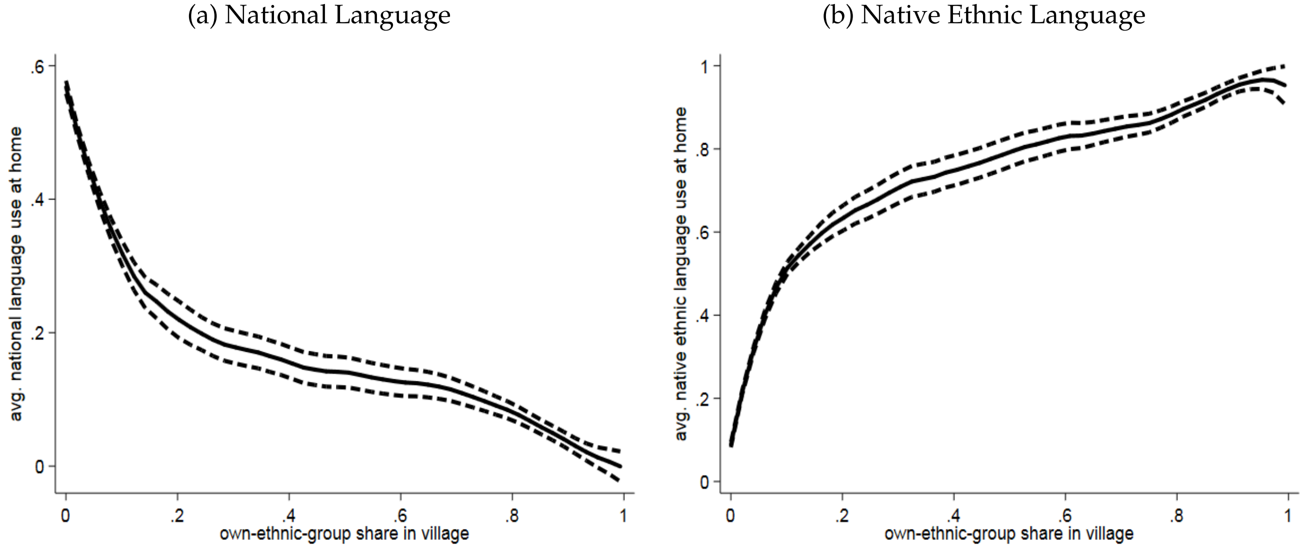
Online Appendix

A Further Empirical Results

I Background Figures

Figure A.1 plots mean (a) national and (b) native ethnic language use against the share of one's own ethnic group in the village. The local-linear regression is at the village × own-group-share level based on the full population of roughly 1.8 million individuals aged 5+ across 817 Transmigration villages.

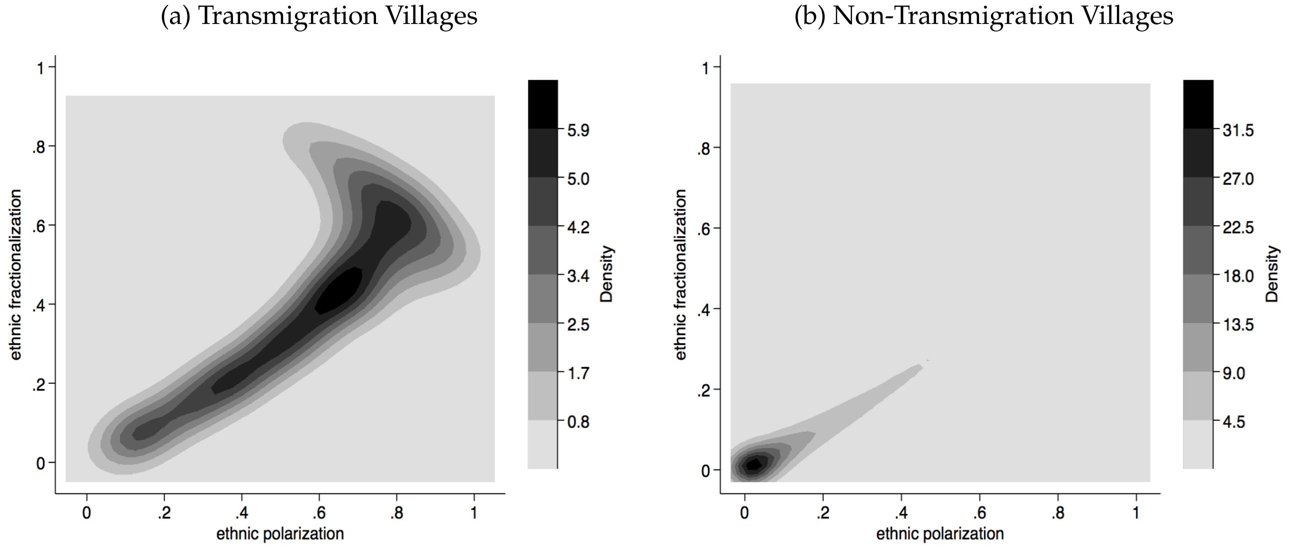
Figure A.1: Own-Group Share and Language Use at Home



Notes: These local-linear regressions use an Epanechnikov kernel and rule-of-thumb bandwidth, and the dashed lines are 95 percent confidence intervals.

Figure A.2 plots the joint kernel density of ethnic fractionalization and polarization in 2010 for (a) Transmigration villages and (b) non-Transmigration villages in the Outer Islands.

Figure A.2: Transmigration Generated Joint Variation in Fractionalization and Polarization



Notes: Both densities employ an Epanechnikov kernel and rule-of-thumb bandwidth.

II Policy-Induced Variation in Diversity and Segregation

Table A.1 shows that Transmigration villages have significantly lower residential segregation across ethnic groups compared to non-Transmigration villages with nearly identical levels of overall diversity. We measure diversity (F and P) and segregation (S, see Section 7.1) using the 2010 Census. We consider two comparison groups. Columns 1 and 2 compare Transmigration villages to all non-Transmigration villages at least 10 km from Transmigration village boundaries in 2000. Columns 3 and 4 compare Transmigration villages to planned settlements that never received the program as a result of budget cutbacks (see Bazzi et al., 2016). These "almost-treated" villages have similar natural advantages to the Transmigration villages we study, but the budget shock meant that they were gradually developed through a process of spontaneous settlement that was not managed by the federal government.

Looking across columns, Transmigration villages have around one-quarter to one-third less ethnic segregation than comparable villages with similar F and P. These conclusions hold whether we define comparable diversity using deciles or percentiles of F and P. As discussed in Section 5.2, the lottery-based assignment of housing plots (and delayed property rights) help explain the persistently lower segregation in Transmigration villages.

Table A.1: Policy-Induced Residential Segregation in Transmigration Villages

	Control Group			
	Non-Transmigration Villages		"Almost-Treated" Villages	
	(1)	**(2)**	**(3)**	**(4)**
Transmigration village	-0.006	-0.004	-0.012	-0.010
	(0.002)	(0.002)	(0.004)	(0.003)
Number of Villages	23,562	23,562	1,514	1,514
Dependent Variable Mean	0.020	0.020	0.029	0.029
R^2	0.262	0.305	0.225	0.383
Function of F, P	Decile	Percentile	Decile	Percentile

Notes: The dependent variable is the Alesina and Zhuravskaya (2011) of residential segregation in 2010. The Transmigration village indicator equals for all Transmigration villages in our study. The control group varies across columns 1–2 and 3–4 as detailed above. Columns 1 and 3 include indicators for the decile of village-level ethnic fractionalization and polarization. Columns 2 and 4 include indicators for the percentile of village-level ethnic fractionalization and polarization. These regressions also control for the same natural advantages ($\mathbf{x}$) and island fixed effects as our baseline regression. Standard errors are clustered by district.

Panel A of Table A.2 shows that diversity (F and P) in Transmigration villages in 2010 appears to be uncorrelated with natural advantages and predetermined correlates of nation building. In contrast, Panel B documents systematic correlations with diversity in non-Transmigration villages. These correlates include physical natural advantages: (1) distance to historic district capitals, (2) distance to the nearest major road, (3) distance to the coast, (4) distance to the nearest river, (5) log altitude, and (6) terrain ruggedness.[1] Other correlates measure the characteristics of populations living in nearby areas within the same district before the Transmigration program, using the 1980 Population Census and restricting to those living in the district in 1978.[2] These include: (7) total district population, (8) Indonesian use at home, (9) radio ownership, (10) television ownership, (11) agriculture, (12) trade and services, and (13) wage-based employment shares. Each column of Table A.2 regresses correlate y listed at the top of table on the ethnic fractionalization and polarization observed in each village in 2010. Together, the stark differences across Panels A and B point to the plausibly exogenous variation in long-run diversity offered by Transmigration program.

Table A.2: Long-Run Diversity, Locational Fundamentals, and Pre-Program Development

							District-Level Population Characteristics, 1978						
Dependent Variable:	distance to				log	ruggedness	total	Indonesian	radio	television	agriculture	trade/service	wage
	district cap.	major road	coast	river	altitude	index	population	use at home	ownership	ownership	empl. share	empl. share	empl. share
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**
							Panel A: Transmigration Villages						
ethnic fractionalization	0.146	0.019	-0.498	0.048	-1.061	0.018	-0.267	0.034	0.009	-0.005	0.028	-0.002	-0.019
	(0.528)	(0.041)	(0.402)	(0.299)	(1.286)	(0.047)	(0.351)	(0.038)	(0.040)	(0.022)	(0.044)	(0.033)	(0.027)
ethnic polarization	-0.241	-0.008	0.654	0.182	0.899	-0.030	-0.178	-0.020	-0.006	0.008	-0.034	0.002	0.047
	(0.432)	(0.031)	(0.307)	(0.257)	(1.093)	(0.045)	(0.254)	(0.024)	(0.030)	(0.016)	(0.032)	(0.022)	(0.021)
Number of Villages	817	817	817	817	817	817	817	817	817	817	817	817	817
Dependent Variable Mean	4.122	0.079	10.557	8.084	3.284	0.311	12.505	0.072	0.463	0.069	0.780	0.150	0.121
R^2	0.014	0.011	0.216	0.063	0.007	0.046	0.240	0.473	0.556	0.087	0.032	0.024	0.034
							Panel B: Non-Transmigration Villages in the Outer Islands						
ethnic fractionalization	-2.166	-0.048	-0.898	-0.141	-2.140	-0.007	-0.436	0.165	0.032	0.109	-0.166	0.144	0.114
	(0.288)	(0.016)	(0.263)	(0.161)	(0.412)	(0.026)	(0.233)	(0.051)	(0.021)	(0.043)	(0.086)	(0.073)	(0.047)
ethnic polarization	1.465	0.027	0.503	0.164	0.701	-0.005	0.294	-0.043	0.016	-0.053	0.109	-0.097	-0.054
	(0.207)	(0.012)	(0.276)	(0.124)	(0.357)	(0.021)	(0.163)	(0.034)	(0.016)	(0.029)	(0.059)	(0.050)	(0.032)
Number of Villages	26,119	29,158	29,158	29,158	26,119	29,158	22,400	22,400	22,400	22,400	22,400	22,400	22,400
Dependent Variable Mean	3.517	0.069	9.727	7.977	3.804	0.277	12.667	0.084	0.427	0.072	0.759	0.166	0.133
R^2	0.067	0.136	0.271	0.041	0.090	0.071	0.235	0.329	0.689	0.146	0.077	0.080	0.069

Notes: The dependent variable is as defined at the top of each column. Sample sizes vary across columns due to matching original data sources with contemporary villages. Standard errors are clustered by district.

[1] See Appendix D.3 for a discussion fo these variables.

[2] These variables are based on data from the 1980 Census sample available on IPUMS International, (ii) measured at the district level based on 1980 district boundaries, (iii) computed using the sampling weights needed to recover district-level population summary statistics, and (iv) restricted to the population in each district that did not arrive as immigrants in 1979 or earlier in 1980 (i.e., the still living population residing in the district in 1978).

III Robust Inference

Table A.3 shows that our qualitative takeaways are not sensitive to the cluster-based inference procedure. Recall that our baseline approach clusters standard errors by 2000 district, of which there are 84. We reproduce the point estimates for our baseline village- and individual-level regression. The 95% confidence intervals are in rows 1 and 6, respectively. Rows 2 and 3 use the Conley (1999) approach to allow for arbitrary correlation across all villages within 50 or 150 km of the given village, respectively. This provides a more flexible clustering procedure that cuts across district boundaries. Row 4 uses the Cameron et al. (2008) wild-cluster bootstrap to account for small-cluster biases and, here, is based on 9,999 replications and uses Webb weights in resampling. Row 5 uses the Young (2016) estimator to adjust the variance-covariance matrices by empirical degrees-of-freedom that account for the realized (correlated) variation in diversity across villages. Rows 7 and 8 uses multi-way clustering on districts and ethnicity based on the procedure in Cameron et al. (2011).

Table A.3: Robustness of Baseline Estimates to Alternative Inference Procedures

	fractionalization	polarization
village-level regression, Column 3 of Table 3	0.636	-0.362
95% confidence interval		
1. baseline, clustering by current district	(0.492, 0.781)	(-0.463, -0.262)
2. Conley (1999) spatial HAC, 50 km bandwidth	(0.490, 0.781)	(-0.463, -0.262)
3. Conley (1999) spatial HAC, 150 km bandwidth	(0.480, 0.793)	(-0.438, -0.286)
4. Cameron et al. (2008) wild cluster bootstrap	(0.480, 0.791)	(-0.467, -0.259)
5. Young (2016) effective degrees-of-freedom adjustment	(0.485, 0.789)	(-0.468, -0.257)
individual-level point estimate, Column 4 of Table 3	0.671	-0.392
95% confidence interval		
6. baseline, clustering by current district	(0.522, 0.820)	(-0.506, -0.278)
7. 2-way clustering: current district + birth district	(0.521, 0.821)	(-0.506, -0.278)
8. 3-way clustering: current district + birth district + ethnicity	(0.520, 0.823)	(-0.498, -0.286)

Notes: This table presents alternative approaches to inference on the baseline results from columns 3 and 4 of Table 3. The estimates are based on unstandardized coefficients.

IV Own-Group Share and Overall Diversity

Table A.4 shows that our results for F and P are not an artifact of variation in the size of one's own ethnic group in the village. Rather, in multi-ethnic communities like Transmigration villages, F and P convey additional information about the size of one's own group relative to multiple other groups. Columns 1, 4 and 7 reproduce the baseline individual-level estimates from columns 4,5, and 6 of Table 3, respectively. Columns 2, 5, and 7 control for the share of an individual's ethnic group in the village. Columns 3, 6, 9 control for the decile of that share with the top decile being the highest shares. Looking across columns, we find that conditioning on own-group-share reduces the effect of F but leaves the effects of P mostly unchanged. Both F and P retain their economic significance.

Table A.4: Distinguishing the Effects of Own-Group Share

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ethnic fractionalization	0.146	0.062	0.104	0.108	0.049	0.085	0.082	0.026	0.056
	(0.016)	(0.016)	(0.016)	(0.012)	(0.013)	(0.014)	(0.011)	(0.011)	(0.010)
ethnic polarization	-0.086	-0.086	-0.093	-0.066	-0.063	-0.071	-0.040	-0.038	-0.042
	(0.013)	(0.013)	(0.014)	(0.009)	(0.009)	(0.012)	(0.008)	(0.009)	(0.010)
own-group share		-0.387			-0.371			-0.357	
		(0.021)			(0.032)			(0.026)	
bottom decile, own-group share			0.429			0.384			0.367
			(0.036)			(0.037)			(0.035)
2nd decile, own-group share			0.222			0.220			0.214
			(0.038)			(0.035)			(0.034)
3rd decile, own-group share			0.101			0.109			0.127
			(0.040)			(0.036)			(0.036)
4th decile, own-group share			0.117			0.109			0.106
			(0.043)			(0.037)			(0.034)
5th decile, own-group share			0.106			0.084			0.081
			(0.043)			(0.037)			(0.037)
6th decile, own-group share			0.105			0.087			0.072
			(0.034)			(0.029)			(0.029)
7th decile, own-group share			0.110			0.089			0.077
			(0.033)			(0.028)			(0.026)
8th decile, own-group share			0.053			0.039			0.034
			(0.024)			(0.021)			(0.021)
9th decile, own-group share			0.023			0.016			0.023
			(0.014)			(0.012)			(0.016)
Number of Individuals	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499
Dependent Variable Mean	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154
R^2	0.114	0.178	0.199	0.223	0.246	0.256	0.281	0.302	0.308
Island FE, x Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ethnicity, Age, Gender FE				✓	✓	✓	✓	✓	✓
Birth District, Current District FE							✓	✓	✓

Notes: The dependent variable is national language use at home. Standard errors are clustered by district.

V Probing Nonlinearities in F and P

Table A.5 reports the point estimates on the indicators for interactions of fractionalization F quintile i and polarization P quintile j ($FiPj$) according to equation (8). These point estimates are used to generate Figure 4(b) by adding the mean for $F1P1$ at the bottom of the table to each coefficient estimate.

Table A.5: Regression Results Underlying Figure 4

	(1)
F1P2	0.036
	(0.025)
F2P1	-0.035
	(0.032)
F2P2	0.050
	(0.014)
F2P3	-0.017
	(0.016)
F3P2	0.366
	(0.173)
F3P3	0.106
	(0.022)
F3P4	0.044
	(0.019)
F3P5	-0.034
	(0.020)
F4P3	0.210
	(0.040)
F4P4	0.140
	(0.034)
F4P5	0.061
	(0.022)
F5P2	0.415
	(0.074)
F5P3	0.263
	(0.043)
F5P4	0.166
	(0.030)
F5P5	0.080
	(0.023)
Number of Villages	817
Dep. Var. Mean: F1P1	0.036
R^2	0.457

Notes: Standard errors are clustered by district.

VI Native and Other Ethnic Language Use

Table A.6 reproduces the baseline individual-level estimates from columns 4–6 of Table 3 for national language use at home. After these first three columns 1-3, columns 4–6 (7–9) change the dependent variable to indicate whether the individual speaks his/her native ethnic (another group's ethnic) language at home. The three columns are mutually exhaustive of potential language choices.

Table A.6: Ethnic Diversity and Language Use At Home

	Dep. Var.: Individual Speaks [. . .] as Main Language at Home								
	Indonesian			**Native Ethnic**			**Other Ethnic**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
ethnic fractionalization	0.146	0.108	0.082	-0.182	-0.117	-0.080	0.036	0.008	-0.002
	(0.016)	(0.012)	(0.011)	(0.015)	(0.010)	(0.009)	(0.012)	(0.008)	(0.008)
ethnic polarization	-0.086	-0.066	-0.040	0.088	0.066	0.042	-0.002	-0.000	-0.002
	(0.013)	(0.009)	(0.008)	(0.011)	(0.008)	(0.008)	(0.010)	(0.008)	(0.006)
Number of Individuals	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499	1,800,499
Dependent Variable Mean	0.154	0.154	0.154	0.764	0.764	0.764	0.082	0.082	0.082
R^2	0.114	0.221	0.280	0.129	0.323	0.370	0.071	0.249	0.294
Island FE, x Predetermined Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ethnicity, Age, Gender FE		✓	✓		✓	✓		✓	✓
Birth District, Current District FE			✓			✓			✓

Notes: Standard errors are clustered by district.

VII National Language Use by Education and Sector of Employment

Tables A.7 and A.8 estimate the full fixed effects, individual-level specification (column 6 of Table 3) separately by education and occupation, respectively. The estimates for F and P reflect standardized effects of a one s.d. increase.

In Table A.7, the baseline estimate from column 6 of Table 3 is reproduced in column 1. Each subsequent column splits the sample to include only those with the education level listed at the top of the column. An individual's education is coded as either the highest level attained or the level in which that individual is currently enrolled. We find similar effects of F and P if we restrict our specifications only to individuals who have finished schooling, or to individuals who are currently enrolled. We also find similar effects on individuals with co-resident parents who have completed different educational levels.

In Table A.8, we restrict to working-age individuals. Column 1 includes the full working-age population, and column 2 restricts to those not currently employed. Columns 3–7 consider mutually exhaustive employment sector categories: (3) agriculture and mining, (4) manufacturing, (5) electricity, construction and transport, which we group together as "manual", (6) trade and services, (7) health, education and public sector, which we group together as "white collar", and (8) all other occupations.

Table A.7: Ethnic Diversity and National Language Use At Home by Education

	baseline	no school	primary		secondary		
			some	completed	junior	senior	post-
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
ethnic fractionalization	0.082	0.057	0.082	0.072	0.088	0.095	0.057
	(0.011)	(0.009)	(0.012)	(0.010)	(0.013)	(0.013)	(0.016)
ethnic polarization	-0.040	-0.029	-0.036	-0.042	-0.042	-0.028	-0.006
	(0.008)	(0.007)	(0.010)	(0.007)	(0.010)	(0.013)	(0.014)
Number of Individuals	1,800,499	141,545	408,269	650,912	336,498	198,334	64,070
Dependent Variable Mean	0.154	0.116	0.165	0.102	0.156	0.260	0.347
R^2	0.281	0.324	0.308	0.250	0.276	0.294	0.304

Notes: Following the specification in column 6 of Table 3, these regressions include the baseline village-level $\mathbf{x}$ controls as well as fixed effects for individual age, gender, ethnicity, birth district, origin district, and relation to the household head. Standard errors are clustered by district.

Table A.8: Ethnic Diversity and National Language Use At Home by Sector of Employment

	baseline	not working	agri/mine	manuf.	manual	trade/svc	white collar	other
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
ethnic fractionalization	0.080	0.089	0.058	0.075	0.107	0.081	0.071	0.092
	(0.011)	(0.013)	(0.008)	(0.016)	(0.015)	(0.012)	(0.016)	(0.017)
ethnic polarization	-0.041	-0.042	-0.034	-0.026	-0.057	-0.035	-0.018	-0.028
	(0.008)	(0.010)	(0.007)	(0.012)	(0.014)	(0.011)	(0.015)	(0.015)
Number of Individuals	1,590,709	685,523	640,488	21,372	27,246	97,930	87,272	10,374
Dependent Variable Mean	0.143	0.165	0.085	0.163	0.152	0.191	0.305	0.205
R^2	0.276	0.286	0.241	0.336	0.327	0.280	0.313	0.325

Notes: Following the specification in column 6 of Table 3, these regressions include the baseline village-level $\mathbf{x}$ controls as well as fixed effects for individual age, gender, ethnicity, birth district, origin district, and relation to the household head. Standard errors are clustered by district.

VIII Addressing Sorting

Table A.9 includes additional fixed effects to control for confounding effects of endogenous sorting along origin–destination or ethnicity–destination pairs. Column 1 reproduces column 6 of Table 3.

Table A.9: Additional Fixed Effects

	(1)	**(2)**	**(3)**
ethnic fractionalization	0.082	0.083	0.081
	(0.011)	(0.011)	(0.011)
ethnic polarization	-0.040	-0.039	-0.040
	(0.008)	(0.009)	(0.009)
Number of Individuals	1,800,499	1,800,499	1,800,499
Dependent Variable Mean	0.154	0.153	0.153
R^2	0.282	0.318	0.344
Ethnicity Fixed Effects	✓	✓	
Birth District + Current District Fixed Effects	✓		
Birth District × Current District Fixed Effects		✓	
Ethnicity × Current District Fixed Effects			✓

Notes: Standard errors are clustered by district.

Table A.10 augments the full fixed effects, individual-level specification in column 6 of Table 3 (reproduced in column 1 below) to account for the share of the population that may have endogenously sorted. We identify as sorters the share of the village population that we classified in column 7 of Table 5 as long-distance sorters. This includes all individuals born in other Outer-Island provinces, which would not have been eligible to join the given village as part of the APPDT allotment. These long-distance migrants that plausibly arrived after the initial year of settlement include individuals of Outer- and Inner-Island ethnicities. The latter include non-indigenous ethnic communities in the Outer Islands, some of whom may have resided there for several generations. We control for ventiles of the village-level population shares of each of these groups in columns 2–4. This slightly reduces the effects of F and P but mostly leaves the results unchanged.

Table A.10: Further Checks on Sorting

	Dep. Var.: National Language Use at Home			
	(1)	**(2)**	**(3)**	**(4)**
ethnic fractionalization	0.082	0.063	0.075	0.061
	(0.011)	(0.012)	(0.011)	(0.012)
ethnic polarization	-0.040	-0.036	-0.036	-0.033
	(0.008)	(0.008)	(0.009)	(0.009)
Number of Individuals	1,800,499	1,800,499	1,800,499	1,800,499
Dependent Variable Mean	0.154	0.154	0.154	0.154
R^2	0.281	0.285	0.287	0.290
Ventiles of Share of Outer Ethnicity Sorters		✓		✓
Ventiles of Share of Inner Ethnicity Sorters			✓	✓

Notes: Standard errors are clustered by district.

IX Addressing Location-by-Time Variation in Program Implementation of Diversity

Table A.11 includes an array fixed effects that account for unobservable variation in program implementation and local conditions. In column 1, we reproduce the baseline village-level specification in column 3 of Table 3. In subsequent columns, we add fixed effects for (2) the year of settlement, (3) the year of settlement by island, (4) the year of settlement by province, (5) the year of settlement by district, (6) the ethnolinguistic homeland, and (7) the ethnolinguistic homeland by year of settlement. We define the ethnolinguistic homeland of each village based on the ethnolinguistic group whose homeland polygon covers the most area of the village. These homelands correspond to the group that is native to the given region, according to the Ethnologue and World Language Mapping Study (WLMS). We are missing this homeland polygon information for a few villages due to omissions in the WLMS shapefiles (see Appendix D).

Looking across columns, the effects of F and P remain stable. This suggests that there is limited region-specific confounding of the sort that one might worry about, e.g., if planners adjusted diversity to better match local receptiveness to integration.

Table A.11: Robustness to Confounding Variation in Program Implementation and Local Conditions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ethnic fractionalization	0.135	0.129	0.130	0.123	0.114	0.121	0.127
	(0.015)	(0.015)	(0.015)	(0.015)	(0.023)	(0.015)	(0.022)
ethnic polarization	-0.083	-0.081	-0.082	-0.073	-0.058	-0.069	-0.071
	(0.012)	(0.012)	(0.012)	(0.012)	(0.019)	(0.012)	(0.019)
Number of Villages	817	817	817	817	817	813	813
Dependent Variable Mean	0.144	0.144	0.144	0.144	0.144	0.145	0.145
R^2	0.437	0.447	0.477	0.648	0.795	0.556	0.704
Year Placed FE		✓					
Island × Year Placed FE			✓				
Province × Year Placed FE				✓			
District × Year Placed FE					✓		
Ethnolinguistic Homeland FE						✓	
Ethnolinguistic Homeland × Year Placed FE							✓

Notes: Standard errors are clustered by district.

X Parental Diversity

Table A.12 shows that the effects of diversity on national language use at home are not driven solely by intermarried households. We retain the full fixed effects specification from column 6 of Table 3 but restrict the sample to children of the household head and to households with both a head and spouse. Column 2 restricts to children with parents in an interethnic marriage while column 5 looks at children with parents of the same ethnicity.

Table A.12: Ethnic Diversity and National Language Use At Home by Parental Diversity

	baseline	parents interethnic	
		yes	no
	(1)	**(2)**	**(3)**
ethnic fractionalization	0.093	0.062	0.084
	(0.014)	(0.018)	(0.014)
ethnic polarization	-0.042	-0.010	-0.043
	(0.011)	(0.014)	(0.011)
Number of Individuals	585,318	76,830	508,423
Dependent Variable Mean	0.182	0.486	0.136
R^2	0.300	0.332	0.275

Notes: Standard errors are clustered by district.

XI Adjusting Children's Names

In Table A.13, we consider alternative indices that are based on an aggregation of similar-sounding children's names using a double-metaphone procedure detailed in Appendix D.2. The effects of F and P are somewhat smaller than with the unadjusted names we use as a baseline in Table 9. This is not surprising given that adjustment procedure reduces the amount of variation across names.

Table A.13: Double Metaphone Adjustment of Children's Names in Table 9

	Dep. Var.: precision of name in identifying . . .			
	Indonesian language home	intermarried household	urban	own-ethnicity
	(1)	**(2)**	**(3)**	**(4)**
ethnic fractionalization	0.109	0.106	0.157	-0.133
	(0.026)	(0.026)	(0.032)	(0.037)
ethnic polarization	-0.064	-0.070	-0.111	0.056
	(0.022)	(0.022)	(0.026)	(0.027)
Number of Individuals	790,705	789,234	790,739	776,205
R^2	0.064	0.080	0.063	0.081

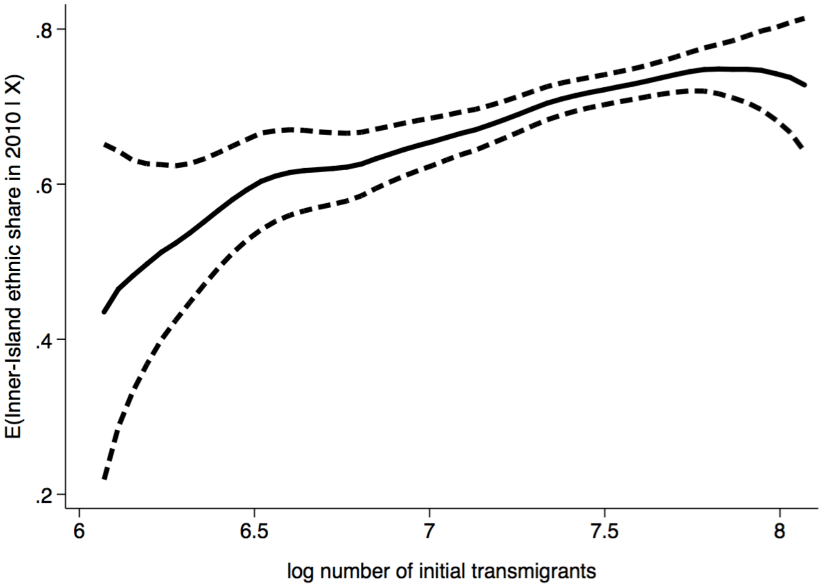
Notes: Standard errors are clustered by district.

XII Further Results on Instrument Strength and Exogeneity

This section provides additional details on the two sets of instruments isolating policy-induced variation in F and P in 2010 as detailed at the end of Section 5.4: (i) the number of initial transmigrants from the Inner Islands of Java/Bali, and (ii) the ethnic composition of those transmigrants from Java/Bali.

Appendix Figure A.3, estimated using the Robinson (1988) semiparametric approach conditional on $\mathbf{x}$, shows that the initial assignment of transmigrants strongly predicts Inner-Island ethnic shares in 2010. This strong relationship is consistent with barriers to mobility making it harder for settlers to leave their initially-assigned communities. Together, these frictions limited tipping, as evidenced by the roughly (log-)linear relationship.

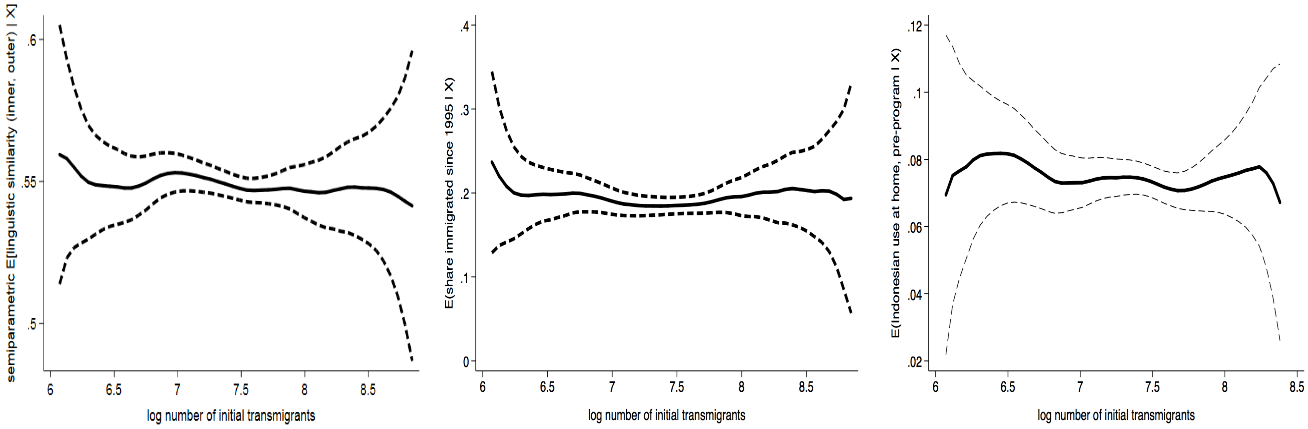
Figure A.3: Initial Transmigrant Assignment and Long-Run Inner-Island Ethnic Share



Notes: This figure reports a semiparametric Robinson (1988) regression and 95% confidence interval of the Inner-Island ethnic share in 2010 on the log of the transmigrant population from Java/Bali placed in that village in the initial year of settlement. The local linear regression is conditional on island fixed effects and the vector $\mathbf{x}$ of predetermined site selection variables described in the paper, and it is estimated based on an Epanechnikov kernel, Fan and Gijbels (1996) rule-of-thumb bandwidth, and trimming of the top 5th and bottom 1st percentile for presentational purposes.

Subsequent results in Appendix Figures A.4 and A.5 provide evidence supporting the exogeneity of the initial number of transmigrants. Figure A.4 shows that planners did not systematically assign more transmigrants to locations with greater (a) the linguistic similarity between the indigenous Outer-Island group and Inner-Island settlers, (b) Indonesian use at home in 1980 in areas near the eventual Transmigration village, or (c) post-program immigration between 1995 and 2000. As discussed in Section 5.4, this suggests that more transmigrants were not sent to locations with an initial predisposition towards national integration or immigrants. Figure A.5 shows that the instrument is uncorrelated with other predetermined proxies for development not captured in the $\mathbf{x}$ vector used for site selection. These proxies include measures of potential agricultural yields, malaria suitability in 1978, agroclimatic similarity (see Bazzi et al., 2016), and a host of district-level characteristics of the population residing within these areas (but not in the immediate settlements) as of 1978, including information on wealth, infrastructure access, schooling, and sector of work. Note that we estimate these relationships flexibly so as to capture the variation underlying our instruments in Table 4, which is based on ventiles of the number of initial transmigrants. What would be concerning in these figures is if we saw an inverted-U relationship since F and P are highest at intermediate levels of initial transmigrants (conditional on village carrying capacity implied by **x**). We find no systematic evidence of such patterns looking across this large set of outcomes.

Figure A.4: Initial Transmigrant Allocation Uncorrelated with Proxies for Sorting

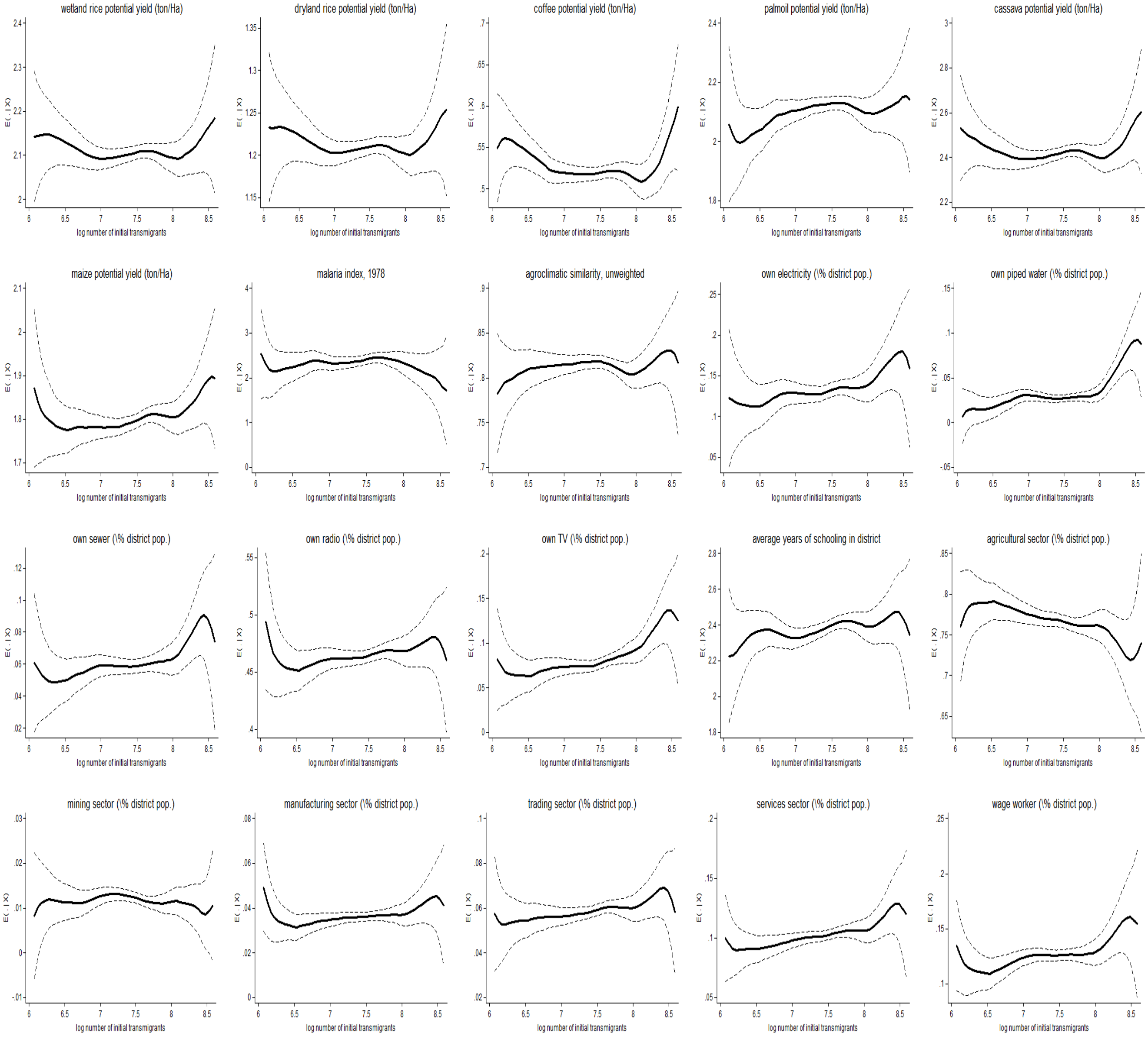


(a) Linguistic Similarity w/ Natives (b) Post-Program Immigration Share (c) Pre-Program Indonesian Use

Notes: This figure reports a semiparametric Robinson (1988) regression and 95 confidence intervals of the Inner-Island ethnic share in 2000 (based on the Population Census) on (a) the linguistic similarity between the Inner-Island ethnic population and the indigenous Outer-Island group according to the *Ethnologue* and World Language Mapping System, (b) the share of the population that immigrated to the village between 1995 and 2000, and (c) the share of the district that spoke the national language at home in 1978 based on the population residing in the given village's district at the time according to the 1980 Population Census. We use a local linear regression with island fixed effects and the vector **x** of predetermined site selection variables, an Epanechnikov kernel, Fan and Gijbels (1996) rule-of-thumb bandwidth, and trimming of the top and bottom percentiles for presentational purposes.

We perform a related set of tests for the second set of instruments capturing the ethnic composition of the initial transmigrants from Java/Bali. In particular, we examine whether the ethnic fractionalization and polarization among those born in Java/Bali before the year of settlement (i.e., plausible first-generation transmigrants) are systematically related to the same predetermined development and nation building proxies in Figures A.4 and A.5. We estimate these regressions conditional on the ventiles of the number of initial transmigrants used in the first-stage regressions in Table 4. We then test of whether the coefficients on these Inner-Island ethnic F and P (based on first-generation transmigrants still alive in 2010) are significantly different from zero in a regression with the given proxy on the left-hand side. Across these 22 outcomes, we only find two p-values less than 0.1, which is what one expects by chance.

Figure A.5: Initial Transmigrant Allocation Uncorrelated with Predetermined Development



Notes: These figures report additional semiparametric regression tests relating the instrument to other predetermined measures of political and economic development. The specifications are otherwise akin to those in the prior figure. Potential yields are obtained from FAO-GAEZ. The malaria suitability index is based on work by Gordon McCord, who generously provided us with the data. The variables beginning with "own electricity" are (i) based on data from the 1980 Population Census (available on IPUMS International), (ii) measured at the district level based on 1980 district boundaries, (iii) computed using the sampling weights needed to recover district-level population summary statistics, and (iv) restricted to the population in each district that did not arrive as immigrants in 1979 or earlier in 1980 (i.e., the still living population residing in the district in 1978). Standard errors in parentheses are clustered at the 1980 district level.

B Model Appendix

This appendix derives the core results for the model in Section 3. Section B.1 presents identity-choice payoffs with a general matching function. a general matching function. Section B.2 describes how the model's revision protocol leads to the replicator dynamic equation governing the evolution of identity choices over time. Section B.3 aggregates these equations over multiple groups to arrive at a village-level expression for growth of the national identity as a function of initial ethnic composition. Section B.4 characterizes the evolutionary equilibria and offers a richer set of results and examples than discussed in the paper.

B.1 Intergroup Contact with a General Matching Function

The model in Section 3 assumes that individuals are randomly matched, so that a group-j individual's probability of meeting a co-ethnic equals her ethnic share p_j. We can model segregated communities by introducing a segregation parameter that changes the matching process. Let m_j denote the probability that a member of group j meets a member of that same group, and let m_k denote the probability that a group j member meets a member of group k. We assume:

$$\begin{aligned} m_j &= p_j + (1-p_j)\,\sigma_j \\ m_k &= (1-\sigma_j)\,p_k \end{aligned}$$

where $0 \leq \sigma_j \leq 1$.[1] At $\sigma_j = 0$, the ethnic group is fully integrated with other groups, and match probabilities are governed by group sizes as in Section 3. As σ_j approaches 1, the ethnic group becomes more segregated, and group j members are more likely to meet their own group members and less likely to meet members of other groups.

For simplicity, we assume that the segregation parameter is identical across groups, so that $\sigma_j = \sigma$ for all $j = 1, ..., J$. Given the payoff structure of Table 1, the expected payoffs of a group-j individual for playing N and E become:

$$\begin{aligned} \text{Nationalist (N):} \quad w_j^N &= \theta m_j + (1-\sigma)\,\theta \sum_{k\neq j} p_k \pi_k - (1-\sigma) \sum_{k \neq j} (1-\pi_k) p_k D_k^N - \gamma_N \\ \text{Ethnic loyal (E):} \quad w_j^E &= \theta m_j \qquad\qquad\qquad\qquad\quad - (1-\sigma) \sum_{k \neq j} (1-\pi_k) p_k D_k^E - \gamma_E \end{aligned}$$

B.2 Pairwise Proportional Imitation and the Replicator Dynamic

Let M denote the total population living in the community. Each person in the community is endowed with a fixed, unchanging ethnicity, belonging to one of $j = 1, ..., J$ ethnic groups. Apart from ethnicity, individuals make identity choices, deciding whether or not to identify with their ethnic culture or to instead adopt the national identity. Revisions to identity choices are made only occasionally, and as we describe in more detail, the infrequent process of identity revision leads to the replicator dynamic we use.

For simplicity, we assume that each person lives forever, so that the village's population is fixed and ethnicity shares are stable. Initially, some fraction of the population chooses whether to retain their own *ethnic identity* (E) or to adopt the *national identity* (N). Let $\pi_j(0)$ denote this initial fraction of group j's population that chooses N. We take these initial conditions as exogenous, but in newly created Transmigration villages, $\pi_j(0)$ was most likely low across groups. In each period, given strategies chosen

[1] Another way of viewing the expression for m_k is that it equals the probability of meeting a non co-ethnic multiplied by the share of group k individuals among non-co-ethnics: $m_k = (1 - m_j)\frac{p_k}{1-p_j}$.

previously, players are randomly matched, according to the process described above. Depending on the outcome of the matching process, payoffs are realized in accordance with Table 1.

After payoffs are realized, some fraction of individuals decides to switch identities, imitating a random sample of strategies played by those around her. As in Sandholm (2010), we assume that the times between when players are allowed to revise their strategies are independent draws from an exponential distribution with rate R.[2] This infrequent process of identity switching delivers significant inertia and makes convergence to an evolutionarily stable equilibrium relatively slow.

In time periods when agents are allowed to revise their strategies, we assume that they adopt the *pairwise proportional imitation* revision protocol (Schlag, 1998; Sandholm, 2010).[3] That is, a player will revise their strategy by imitating a randomly selected strategy played by others around them. She will do so only if the payoff from that strategy exceeds her own payoff.[4] The probability that this revision occurs is proportional to the individual differences in payoffs.

Note that this imitative revision protocol, by its nature, assumes that agents are myopic in their decision making. Instead of forming beliefs or expectations about the evolution of the community's identity, individuals revise their identity decisions based only on current information. They need not be able to observe π_j for other groups or for their own group; instead, they only need to know whether their payoff exceeds that of a randomly sampled strategy when revisions are allowed to occur. As the village becomes larger and individuals become more anonymous, this proposition becomes more sensible.[5]

Given the structure of the matching process from Appendix B.1, and the payoffs in Table 1, Sandholm (2010) shows that this *pairwise proportional imitation* revision protocal leads to the mean replicator dynamic:

$$\begin{aligned}\dot{g}_j^N &= \pi_j\left(w_j^N - w_j\right)\\ &= \pi_j\left(w_j^N - \pi_j w_j^N - (1-\pi_j)\,w_j^E\right)\\ &= \pi_j\left((1-\pi_j)\,w_j^N - (1-\pi_j)\,w_j^E\right)\\ &= \pi_j\,(1-\pi_j)\left(w_j^N - w_j^E\right)\end{aligned} \tag{B.1}$$

B.3 Village-level Growth Rate of Nationalists

General Case. Using the expected payoffs from the different identity choices, we can write the growth rate in the share of group-j adopting the national identity as:

$$\begin{aligned}\dot{g}_j^N &= \pi_j\,(1-\pi_j)\left(w_j^N - w_j^E\right)\\ &= \pi_j\,(1-\pi_j)\left[(1-\sigma)\,\theta\sum_{k\neq j} p_k\pi_k + (1-\sigma)\sum_{k\neq j}(1-\pi_k)p_k\left(D_k^N - D_k^E\right) - (\gamma_N - \gamma_E)\right]\end{aligned}$$

[2] Formally, let T denote the time an individual must maintain their identity choice, after which revisions are allowed to occur. We assume that $T \sim \exp(R)$, so that $\mathbb{P}(T \leq t) = 1 - e^{-Rt}$. This means that the number of identity revisions that are allowed to occur during the time interval $[0,t]$ follows a Poisson distribution, with mean Rt.

[3] Another revision protocol that would lead to the same replicator dynamic would be *imitation driven by dissatisfaction*. In this protocol, agents that are allowed to revise their strategies compare their current payoff to some ideal payoff K. The probability of revisions is proportional to the difference between K and their current payoff (Sandholm, 2010).

[4] Note that a slightly different setup would consider the revision timing to reflect a birth and death process. Instead of living forever, individuals have survival probabilities of time T which is distributed exponential with rate R. When individuals die, they are replaced through asexual reproduction, and the probability that newly born agents decide to switch their identities is proportional to the relative fitness of individual payoffs in the population. This *natural selection* revision protocal leads to a slightly different replicator dynamic, but Sandholm (2010) argues that it only differs from B.1 by a change in speed.

[5] An important limitation of this approach is that it rules out the possibility that certain village leaders or "norm entrepreneurs" may steer the village towards a new identity within a fairly short time period. See Young (2015) for more discussion.

$$
\begin{aligned}
&= \pi_j\left(1-\pi_j\right)\left[\left(1-\sigma\right)\theta\sum_{k\neq j}p_k\pi_k - \left(1-\sigma\right)\sum_{k\neq j}(1-\pi_k)p_kD_k - \gamma\right] \\
&= \pi_j\left(1-\pi_j\right)\Bigg[\underbrace{\left(1-\sigma\right)\theta\left(\overline{\pi}-p_j\pi_j\right)}_{\substack{\text{relative gain from}\\ \text{productive interactions}}} - \underbrace{\left(1-\sigma\right)\sum_{k\neq j}(1-\pi_k)p_kD_k}_{\substack{\text{relative interethnic}\\ \text{antagonism}}} - \underbrace{\gamma}_{\substack{\text{relative}\\ \text{identity}\\ \text{cost}}}\Bigg] \qquad \text{(B.2)}
\end{aligned}
$$

where $\overline{\pi} = \sum_k p_k\pi_k$. So, this growth rate depends on whether the gains from productive interactions are greater than the costs of intergroup antagonism and national-identity adoption. Segregation dampens the effects of these benefits and costs on the growth of the nationalist share. The values of those terms crucially depend on the nationalist shares of all other groups, π_k for $k \neq j$, pointing to social externalities.

To obtain the village-level growth rate of nationalists, we take the sum of $\dot{g}_j^N$ weighted by its group share. Denoting $A_i = \pi_i(1-\pi_i)$ to simplify notation, we obtain:

$$
\begin{aligned}
\dot{G}^N = \sum_j p_j\dot{g}_j^N &= \sum_j p_jA_j\left(1-\sigma\right)\theta(\bar{\pi}-p_j\pi_j) - \left(1-\sigma\right)\sum_j\sum_{k\neq j}[p_jp_jA_j(1-\pi_k)D_j] - \sum_i p_jA_j\gamma \\
&= \left(1-\sigma\right)\theta\Big(\bar{A}\bar{\pi} - \sum_i A_j\pi_jp_j^2\Big) - \left(1-\sigma\right)\sum_i\sum_{j\neq i}[p_jp_jD_jA_j(1-\pi_k)] - \bar{A}\gamma \\
&= \left(1-\sigma\right)\theta\Phi\Big(1-\sum_j\phi_jp_j^2\Big) - \left(1-\sigma\right)\sum_j\sum_{k\neq j}p_jp_kT_{jk} - \bar{A}\gamma \qquad \text{(B.3)}
\end{aligned}
$$

where $\bar{A} = \sum_j p_jA_j$, $\overline{\pi}$ is defined above, $\Phi = \bar{A}\overline{\pi}$, $\phi_j = (A_j\pi_j)/\Phi$, and $T_{jk} = A_j(1-\pi_k)D_k$.

Exact Approximation. If we make two simplifying assumptions, we can derive a closed-form solution for the aggregate growth rate of nationalists. The first is that each group has an identical nationalist share (i.e., $\pi_j = \pi$ for all $j = 1, ..., J$). The second is that the relative antagonism term for group k, D_k, is a linear function of that group's shares: $D_k = 4\psi p_k$ for all $k = 1, ..., J$. If these hold, we have:

$$
\begin{aligned}
\dot{G}^N &= \sum_{j=1}^{J}p_j\dot{g}_j^N \\
&= \sum_{j=1}^{J}p_j\left(\pi_j\left(1-\pi_j\right)\right)\left[\left(1-\sigma\right)\theta\sum_{k\neq j}p_k\pi_k - \left(1-\sigma\right)\sum_{k\neq j}(1-\pi_k)p_kD_k - \gamma\right] \\
&= \pi\left(1-\pi\right)\left\{\sum_{j=1}^{J}p_j\left[\left(1-\sigma\right)\theta\pi\sum_{k\neq j}p_k - \left(1-\sigma\right)\left(1-\pi\right)\sum_{k\neq j}p_kD_k - \gamma\right]\right\} \\
&= \pi\left(1-\pi\right)\left\{\sum_{j=1}^{J}p_j\left[\left(1-\sigma\right)\theta\pi\left(1-p_j\right) - \left(1-\sigma\right)4\psi\left(1-\pi\right)\sum_{k\neq j}p_k^2 - \gamma\right]\right\} \\
&= \pi\left(1-\pi\right)\left\{\left(1-\sigma\right)\theta\pi\left[\sum_{j=1}^{J}\left(p_j-p_j^2\right)\right] - \sum_{j=1}^{J}p_j\left[\left(1-\sigma\right)4\psi\left(1-\pi\right)\sum_{k\neq j}p_k^2\right] - \sum_{j=1}^{J}p_j\gamma\right\} \\
&= \pi\left(1-\pi\right)\left[\left(1-\sigma\right)\theta\pi\left(1-\sum_{j=1}^{J}p_j^2\right) - \left(1-\sigma\right)4\psi\left(1-\pi\right)\sum_{j=1}^{J}\sum_{k\neq j}p_jp_k^2 - \gamma\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \pi\left(1-\pi\right)\left[\left(1-\sigma\right)\theta\pi\left(1-\sum_{j=1}^{J}p_j^2\right)-\left(1-\sigma\right)4\psi\left(1-\pi\right)\sum_{j=1}^{J}p_j^2\sum_{k\neq j}p_k-\gamma\right] \\
&= \pi\left(1-\pi\right)\left\{\left(1-\sigma\right)\theta\pi\left(1-\sum_{j=1}^{J}p_j^2\right)-\left(1-\sigma\right)\psi\left(1-\pi\right)\left[4\sum_{j=1}^{J}p_j^2(1-p_j)\right]-\gamma\right\} \\
&= \beta_0+\left(1-\sigma\right)\beta_1 F-\left(1-\sigma\right)\beta_2 P \qquad \text{(B.4)}
\end{aligned}
$$

where, as in equation (3), $\beta_0 = -\pi\left(1-\pi\right)\gamma < 0$, $\beta_1 = \theta\pi^2\left(1-\pi\right) > 0$, and $\beta_2 = \psi\pi\left(1-\pi\right)^2 > 0$. Equation (3) is the special case of full integration, where the segregation parameter σ is equal to 0. All else constant, the effect of F and P on the aggregate growth rate of nationalist-identity adopters is weaker in more segregated communities (as σ goes to 1).

B.4 Evolutionary Equilibria

Proposition B.1 characterizes the evolutionary equilibria of the system of differential equations formed by (B.2), when the segregation parameter, $\sigma_j = \sigma$ for all j.

Proposition B.1. *With matching segregation parameter $\sigma_j = \sigma$ for all j, the system of J differential equations formed by* (B.2) *has three unique steady states, of which only two are asymptotically stable:*

1. (National Convergence): $\pi_j = 1$ *for all* $j = 1, ..., J$.
2. (Ethnic Backlash): $\pi_j = 0$ *for all* $j = 1, ..., J$.
3. (Unstable Tipping Point): $\pi_j = \pi_j^*$ *for all* $j = 1, ..., J$*, where we have*

$$
\pi_j^* = \left(\frac{\gamma\left(1-\sigma\right)^{-1}\left(J-1\right)^{-1}+D_j p_j}{\theta p_j + D_j p_j}\right) \quad . \qquad \text{(B.5)}
$$

When each group j's national identity shares are equal to π_j^*, the term in brackets of (B.2) *is equal to zero for all j.*

Proof. Note that if $\pi_j = 1$ for all j, $\dot{g}_j^N = 0$ for all j, so this is clearly a fixed point of the system of differential equations. Similarly, if $\pi_j = 0$ for all j, $\dot{g}_j^N$ is also equal to 0 for all j.

To solve for the unstable tipping point in closed form, we use an add-subtract trick. If all of the terms in brackets of (B.2) were equal to 0, the following J equations must hold:

$$
\theta\left(\overline{\pi}-p_1\pi_1\right) = \sum_{\substack{k=1\\k\neq 1}}^{J}\left(1-\pi_k\right)p_k D_k + \left(\frac{\gamma}{1-\sigma}\right) \qquad \text{(1*)}
$$

$$
\theta\left(\overline{\pi}-p_2\pi_2\right) = \sum_{\substack{k=1\\k\neq 2}}^{J}\left(1-\pi_k\right)p_k D_k + \left(\frac{\gamma}{1-\sigma}\right) \qquad \text{(2*)}
$$

$$
\vdots
$$

$$
\theta\left(\overline{\pi}-p_J\pi_J\right) = \sum_{\substack{k=1\\k\neq J}}^{J}\left(1-\pi_k\right)p_k D_k + \left(\frac{\gamma}{1-\sigma}\right) \qquad \text{(J*)}
$$

where again we're assuming that $\sigma_j = \sigma$ for all j.

This add-subtract trick can be explained as follows:

1. First, add up both sides of all equations that contain the π_j terms that we want to isolate:

$$(1^*) + (2^*) + \ldots + (J^*) \quad \text{dropping equation} \quad (j^*) \tag{B.6}$$

   Notice that when we add up (1^*), (2^*), ..., (J^*) but do not add equation (j^*), we will have an expression with both sides containing $(J-2)$ terms of the form $\kappa\pi_k$ where $k \neq j$, but $(J-1)$ terms of the form $\kappa\pi_j$ on both sides.

   Collecting terms, we can rewrite (B.6) as:

$$(J-2)\,\theta\overline{\pi} - \theta p_j \pi_j = \underbrace{\left[(J-2)\sum_{\substack{k=1\\k\neq j}}^{J}(1-\pi_k)\,p_k D_k\right]}_{\text{terms not containing } j} + (J-1)\,(1-\pi_j)\,p_j D_j + (J-1)\left(\frac{\gamma}{1-\sigma}\right)$$

2. Next, we subtract $(J-2)$ times equation (j^*) on both sides to remove the terms not containing j:

$$\begin{aligned}(J-2)\,\theta\overline{\pi} - \theta p_j\pi_j - (J-2)\left[\theta\left(\overline{\pi} - p_j\pi_j\right)\right] = {} & \underbrace{\left[(J-2)\sum_{\substack{k=1\\k\neq j}}^{J}(1-\pi_k)\,p_k D_k\right]}_{\text{terms not containing } j} + (J-1)\,(1-\pi_j)\,p_j D_j \\ & + (J-1)\left(\frac{\gamma}{1-\sigma}\right) \\ & - (J-2)\left[\underbrace{\sum_{\substack{k=1\\k\neq j}}^{J}(1-\pi_k)\,p_k D_k +}_{\text{terms not containing } j}\left(\frac{\gamma}{1-\sigma}\right)\right]\end{aligned}$$

Cancelling and rearranging, the expression above simplifies to the following:

$$(J-1)\,\theta p_j\pi_j = (J-1)\,(1-\pi_j)\,D_j p_j + \left(\frac{\gamma}{1-\sigma}\right)$$

Solving for π_j, we obtain the unstable tipping point:

$$\pi_j^* = \left(\frac{\gamma\,(1-\sigma)^{-1}\,(J-1)^{-1} + D_j p_j}{\theta p_j + D_j p_j}\right)$$

□

Note that $\pi^* = (\pi_1^*, \pi_2^*, \ldots, \pi_J^*)'$ represents the unstable tipping point level of national-identity adoption. If ethnic group adoption shares are greater than these values, the dynamics will push them asymptotically to the national identity equilibrium ($\pi_j = 1$ for all $j = 1, \ldots, J$). If ethnic group adoption shares are below these values, the system will converge to ethnic backlash ($\pi_j = 0$ for all $j = 1, \ldots, J$).

As π_j^* grows smaller, the basin of attraction to national identity increases. Notice that as σ increases to 1, π_j^* gets larger, so that the basin of attraction to national identity gets smaller. This is intuitive; with larger values of σ, there is less mixing of ethnic groups. This reduces the gains from national identity adoption as a coordination device, leading to more ethnic-identity choices.

B.4.1 5-Group Example

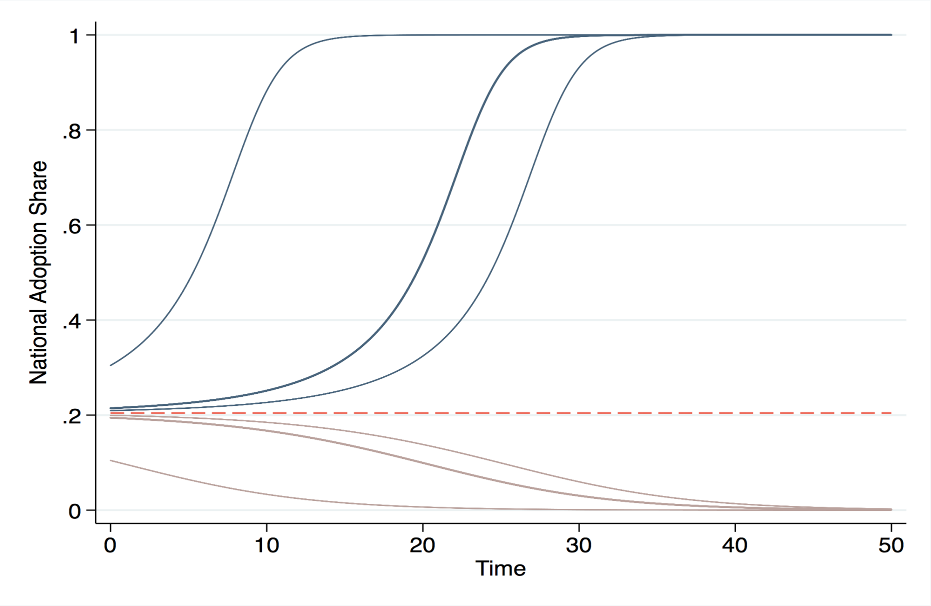
As an example, consider a 5-group village with groups of equal size, so that $p_1 = p_2 = p_3 = p_4 = p_5 = 0.2$. Fix $\theta = 1$, $D_j = 0.5p_j$, $\gamma = 0.1$, and $\sigma = 0$. Because of symmetry, we can just focus on the evolution of one group's shares. In this example, $\pi_j^* \approx 0.205$ for all j.

Figure B.1 shows the evolution of $\pi_j(t)$ when we vary initial starting values for each group (so that $\pi_j(0) = \pi(0)$ for all j). To plot this figure, we use Euler's method to approximate the system of differential equations equations. Given a choice of starting values, we can write $\pi_j(t)$ as

$$\pi_j(t) = \pi_j(t-1) + \frac{\partial \pi_j(t-1)}{\partial t}\big[t - (t-1)\big]$$

, where we evaluate $d\pi_j(t-1)/dt$ by plugging in lagged values of $\pi_j(t-1)$ for $j = 1, ..., J$ into (3) above. With relatively small step sizes, this approximates the actual function.

Figure B.1: $\pi_j(t)$ for Different Starting Values



In Figure B.1, the red dashed line depicts $\pi^* \approx 0.205$. If all group shares begin at this precise value, they will continue to stay there ad infinitum, but this is an unstable equilibrium. The blue lines reflect the path of national identity adoption with starting values that are greater than π^* by 0.1, 0.01, and 0.001. In each case, the group (and village) converges to national identity adoption. The dark red lines measure the path of national identity adoption when starting values are less than π^* by 0.1, 0.01, and 0.00. In each case, the group (and village) converge to ethnic attachment. This example illustrates the dependence of convergence to nationalism on initial national shares, and the role of the fixed point in tipping the equilibrium one way or the other.

B.4.2 Approximating Village-Level Tipping Points

In newly-created Transmigration villages in the 1980s, we expect that initial national identity adoption shares were relatively small and the same across ethnic groups. Let $\pi_j(0) = \pi(0)$ denote the precise values of these shares. A sufficient condition for convergence to the national identity is if the village-level initial share is greater than p^*, the village-level weighted average of the π_j^* tipping points.

That is, the village will converge to national identity if we have:

$$p^* \leq \pi(0)$$

where p^* is given by:

$$\begin{aligned}
p^* &= \sum_{j=1}^{J} p_j \pi_j^* \\
&= \sum_{j=1}^{J} p_j \left(\frac{\gamma \left(1-\sigma\right)^{-1} \left(J-1\right)^{-1} + D_j p_j}{\theta p_j + D_j p_j} \right)
\end{aligned}$$

As p^* gets larger, it becomes more difficult for the village to converge to the national identity. We will show that p^* can be approximated by a linear function of F and P. To do so, we make use of the properties of geometric series.

Assume that $D_j = \psi p_j$ and that $\sigma_j = \sigma$ for all j. Let:

$$\begin{aligned}
v_j &\equiv p_j \pi_j^* \\
&= p_j \left(\frac{\gamma \left(1-\sigma\right)^{-1} \left(J-1\right)^{-1} + \psi p_j^2}{\theta p_j + \psi p_j^2} \right) \\
&= \frac{\gamma \left(1-\sigma\right)^{-1} \left(J-1\right)^{-1} + \psi p_j^2}{\theta + \psi p_j} \\
&= \left(\gamma \left(1-\sigma\right)^{-1} \left(J-1\right)^{-1} + \psi p_j^2 \right) \frac{1}{\theta + \psi p_j} \\
&= \left(\left[\frac{\gamma \left(J-1\right)^{-1}}{\theta \left(1-\sigma\right)} \right] + \left[\frac{\psi}{\theta} \right] p_j^2 \right) \frac{1}{1 + \frac{\psi}{\theta} p_j}
\end{aligned}$$

From the properties of geometric series, we can write:

$$\begin{aligned}
v_j &= \left(\left[\frac{\gamma \left(J-1\right)^{-1}}{\theta \left(1-\sigma\right)} \right] + \left[\frac{\psi}{\theta} \right] p_j^2 \right) \frac{1}{1 + \frac{\psi}{\theta} p_j} \\
&= \left(\left[\frac{\gamma \left(J-1\right)^{-1}}{\theta \left(1-\sigma\right)} \right] + \left[\frac{\psi}{\theta} \right] p_j^2 \right) \sum_{k=0}^{\infty} (-1)^k \left(\frac{\psi}{\theta} \right)^k p_j^k
\end{aligned}$$

Note that this holds as long as $\left| \frac{\psi}{\theta} p_j \right| < 1$, which will be true as long as $\psi < \theta$.

Define the following constants:

$$\begin{aligned}
A &= \left[\frac{\gamma \left(J-1\right)^{-1}}{\theta \left(1-\sigma\right)} \right] \\
B &= \left[\frac{\psi}{\theta} \right]
\end{aligned}$$

where both $A > 0$ and $B > 0$, because $\psi > 0$, $\theta > 0$, $\gamma > 0$, and $0 < \sigma < 1$. Using this notation, we can write:

$$\begin{aligned}
v_j &= \left(A + Bp_j^2\right) \sum_{k=0}^{\infty} (-1)^k B^k p_j^k \\
&= \left(A + Bp_j^2\right) \left(1 - Bp_j + B^2 p_j^2 - B^3 p_j^3 + ...\right) \\
&= A + Bp_j^2 - ABp_j - B^2 p_j^3 + AB^2 p_j^2 + B^3 p_j^4 - AB^3 p_j^3 - B^4 p_j^5 + ...
\end{aligned}$$

If we ignore terms of p_j^κ for $\kappa \geq 4$, we can approximate v_j as follows:

$$\begin{aligned}
v_j &\approx A + Bp_j^2 - ABp_j - B^2 p_j^3 + AB^2 p_j^2 - AB^3 p_j^3 \\
&= A - ABp_j + \left(B + AB^2\right) p_j^2 - \left(B^2 + AB^3\right) p_j^3 \\
&= A - ABp_j + \left(B + AB^2 - B^2 - AB^3\right) p_j^2 + \left(B^2 + AB^3\right) p_j^2 - \left(B^2 + AB^3\right) p_j^3 \\
&= A - ABp_j + \Bigg[\left(B - B^2\right) + A\left(B^2 - B^3\right) \Bigg] p_j^2 + \Bigg[B^2 + AB^3 \Bigg] \left(p_j^2 - p_j^3\right)
\end{aligned}$$

where we add and subtract $\left(B^2 + AB^3\right) p_j^2$ between lines 2 and 3 above. Define the following parameters:

$$\begin{aligned}
-C &= \Bigg[\left(B - B^2\right) + A\left(B^2 - B^3\right) \Bigg] \\
D &= \Bigg[B^2 + AB^3 \Bigg]
\end{aligned}$$

So, we can write our approximation for v_j as:

$$v_j \approx A - ABp_j - Cp_j^2 + D\left(p_j^2 - p_j^3\right)$$

Summing across groups in the village, we have:

$$\begin{aligned}
p^* &= \sum_{j=1}^{J} v_j \\
&\approx \sum_{j=1}^{J} \left(A - ABp_j - Cp_j^2 + D\left(p_j^2 - p_j^3\right)\right) \\
&= JA - AB - C\sum_{j=1}^{J} p_j^2 + D\sum_{j=1}^{J} \left(p_j^2 - p_j^3\right) \\
&= JA - AB - C + C - C\sum_{j=1}^{J} p_j^2 + D\sum_{j=1}^{J} \left(p_j^2 - p_j^3\right) \\
&= \Bigg[JA - AB - C \Bigg] + C\left(1 - \sum_{j=1}^{J} p_j^2\right) + D\sum_{j=1}^{J} p_j^2 \left(1 - p_j\right) \\
&= \Bigg[JA - AB - C \Bigg] + C\,\mathbf{F} + D\,\mathbf{P} \qquad \text{(B.7)}
\end{aligned}$$

where $\mathbf{F}$ is the village-level fractionalization index, and $\mathbf{P}$ is the polarization index.

Note that $D > 0$, because

$$D = \Bigg[B^2 + AB^3 \Bigg] = B^2 \Bigg[1 + AB \Bigg]$$

and both B and A are greater than 0. So, the coefficient on $\mathbf{P}$ is positive. Note also that we can rewrite C as follows:

$$\begin{aligned}-C &= \left[\left(B - B^2\right) + A\left(B^2 - B^3\right)\right] \\ -C &= \left[B\left(1 - B\right) + AB^2\left(1 - B\right)\right] \\ -C &= \left(1 - B\right)\left[B + AB^2\right] \\ \implies C &= -\left(1 - B\right)\left[B + AB^2\right]\end{aligned}$$

So, $C < 0$ if $B < 1$, which will occur when $\psi < \theta$. This means that as long as the interethnic antagonism term is smaller than the gains from trade, the coefficient on fractionalization will be negative.

B.4.3 5-Group Simulations

At the village level, the total tipping-point threshold is given by:

$$\begin{aligned}p^* &= \sum_{j=1}^{J} p_j \pi_j^* \\ &= \sum_{j=1}^{J} p_j \left(\frac{(J-1)^{-1}\gamma + \psi p_j^2}{\theta p_j + \psi p_j^2}\right)\end{aligned}$$

To see how p^* varies with F and P at the village level, we first simulated 10,000 villages, each with 5 different ethnic groups.[6]

To simulate the effects of F and P on village-level tipping behavior, we estimated regressions of the following form,

$$p_v^* = \beta_0 + \beta_1 F_v + \beta_2 P_v + \mathbf{x}_v'\theta + \varepsilon_v \,,$$

where we vary $\psi = 0, 0.1, 0.2, 0.3, 0.4, 0.5$, and $\mathbf{x}_v$ contains the levels of group shares for all groups. Results are available upon request, but overall, we found that $\widehat{\beta}_1$ was negative and $\widehat{\beta}_2$ was positive, as expected. Additionally, the coefficients C and D from the approximation formula (B.7) did a good job of reflecting the magnitudes of the coefficients estimated from the simulation data.

[6] The group shares are drawn uniformly from the unit simplex ($1 = \sum_j p_j$). We follow Rubin (1981) in drawing from the simplex. First, we draw uniform random variables for each group, denoted by $U_{\dot{g}}$ for $g = 1, ..., 5$. Then, we form $y_{\dot{g}} = -\ln\left(U_{\dot{g}}\right)$. Finally, we normalize: $\widetilde{p}_{\dot{g}} = \frac{y_{\dot{g}}}{\sum_j y_j}$.

C The Transmigration Program: Policy Implications and External Validity

We discuss here several potential ways in which our study may matter for policy and for understanding migration and nation building in other contexts.

First, our results are particularly informative for rural-to-rural migration, which comprises population flows that are 1.5 to 2 times greater than those from rural-to-urban migration (Young, 2013). Yet, there is little research on migration between rural areas. This is an important gap in the literature given mounting concerns about the effects on climate change on agricultural viability. The International Organization for Migration estimates that 200 million people may become environmentally-induced migrants by 2050. Many of those displaced from rural areas will likely move initially to other rural areas and may not be that different from transmigrants in terms of their limited resources and need for government support to migrate. It is therefore important to understand how such migrants react to diversity. Our setting and results suggest that planning for such shocks should account for the differential effects of migration-induced changes in fractionalization and polarization.

Second, a recent refugee crisis has also stoked debate over how to design resettlement policies to facilitate the integration of diverse groups. Refugee flows are likely to continue and perhaps even grow in the foreseeable future as extreme weather events, climate change, and conflict become more pervasive and frequent (Hsiang et al., 2013; Harari and LaFerrara, 2018; Sherbinin et al., 2011). Part of the policy challenge, discussed by Bansak et al. (2018), lies in assigning migrants to locations where their cultural background and linguistic skills are best matched. Our results shed light on how to optimally mix refugees from different backgrounds and how to organize housing within new settlements. It seems better to send many different groups of refugees to a given destination rather than a few large groups. With many small groups, it will be important to design housing schemes that encourage intergroup contact both among refugees and with natives. However, if a few large groups must be resettled in the same area, it may be best to limit the scope for intergroup contact through more segregated housing. More generally, our findings, and the Pew survey noted above, point to the importance of helping refugees learn the national language.

Third, while the Transmigration program is unique in certain respects, it shares features with other major rural resettlement schemes across the developing world. As referenced in Bazzi et al. (2016), these include the Polonoroeste program in Brazil that relocated 300,000 migrants between 1981 and 1988 at a cost of US$ 1.6 billion, villagization programs in Ethiopia that relocated 440,000 households between 2003 and 2005, the resettlement of 400,000 individuals in Africa due to dam construction, the resettlement of 4 million migrants in Mozambique between 1977 and 1984, and another 43,000 households that were relocated following floods in the 2000s (Arnall et al., 2013; de Wet, 2000; Hall, 1993; Taye and Mberengwa, 2013; World Bank, 1999).

Fourth, the Transmigration program also has parallels in historical efforts to settle frontier areas through state-sponsored migration. Poland, for example, implemented a large-scale resettlement effort post-WWII to populate its newly acquired and depopulated Western Territories from Germany (Becker et al., 2018). Other examples abound across the Americas during the age of mass migration as central governments sought to expand the scope of the state by facilitating Westward expansion of the rural frontier. Diversity (or lack thereof) in the newly settled areas may have contributed to nation building in interesting ways. To date, there is little work on this question in the historical context. There may be similar forces to the ones we identify on the rural frontier in Indonesia.

Finally, there are a number of desegregation policies in both rich and poor countries that affect community-level diversity with implications for integration and identity formation. These policies generally place quotas on ethnic, religious, or immigrant groups in neighborhoods in Singapore (Wong, 2013), India (Barnhardt et al., 2017), Germany (Glitz, 2012), and Denmark (Dutch Refugee Council, 1999). See Polikoff (1986) and Boustan (2011) for a review of residential desegregation policies.

D Data Appendix

Table D.1 summarizes the main datasets used in the paper. We describe each of these sources in the following sections.

Table D.1: Summary of Main Datasets

Dataset	Description	Obs. Unit
Transmigrant placement		
Transmigration Census, 1998	location of Transmigration village; the number of individuals resettled, and year of settlement	village
Demographics		
Population Census 2010	relationship to household head, ethnicity, highest level of schooling, sectoral employment, birth information (year and month, district), district of residence in 2005, (sub-)village administrative identifiers	individual
Population Census 2000	relationship to household head, ethnicity, highest level of schooling, sectoral employment, birth information (year and month, district), district of residence in 1995	individual
Social and Economic Outcomes		
Population Census 2010	primary language at home, Indonesian speaking ability, full name, intermarriage	individual
Population Census 2000	intermarriage	individual
Podes 2002	distance to (sub)district capital, top 3 parties in 1999 election, village-provided public goods (safe drinking water, garbage collection, public toilet facilities, 4-wheel road access, and streetlights), ethnic conflict	village
Podes 2005, 2008, 2011, 2014	ethnic conflict	village
Podes 1999	voter turnout	village
Susenas 2000–12	mean household expenditures per capita	village
Susenas 2012	social attitudes: contribute to public goods, community group participation, tolerance of non-coethnics, trust of neighbors to watch children and house, feeling of safety, ease of obtaining help of neighbors, contribute to help misfortunate neighbors.	individual
SNPK 2000–14	ethnic conflict	village
Indonesia Family Life Survey (IFLS) 1997, 2014	language use at home (1997, 2014); own, mother's, and father's ethnicity; relative trust of non-coethnics.	Individual
NOAA Light Intensity	light intensity data, 2010	30-arc-second grid
Agroclimatic characteristics		
GIS Map - Dept. Public Works	village area, distance to coast, roads and others.	village
Harmonized World Soil Database	elevation, ruggedness, soil quality (organic carbon, topsoil characteristics, texture, drainage).	30-arc-second grid
Terrestrial Precipitation and Temperature Data	rainfall (Matsuura and Wilmott, 2012b) and temperature (Matsuura and Wilmott, 2012a), 1948-1978.	$0.5^\circ \times 0.5^\circ$ grid

D.1 Transmigration Census and Maps

We employ the Ministry of Manpower and Transmigration's 1998 Census of Transmigration sites established between 1952 and 1998 to obtain details about the placements of transmigrants. The census identifies the physical locations and names of realized transmigration sites, years of establishment, and the number of transmigrants at the time of the initial settlement. Our main sample comprises 817 Transmigration villages established in Indonesia's Third and Fourth Five-Year Development Periods (1979–1988) in the Outer islands, excluding Papua. The 1998 Transmigration Census identifies villages that correspond to those in the 2000 Population Census shapefile. These 2000 village boundaries are the level at which the program varied and form our core spatial unit of analysis.[1] For some analyses, including column 1 of Table 6, we redefine the spatial unit of analysis (for defining F and P) to group all Transmigration villages that share a boundary and hence are part of the same cluster.

D.2 Demographic and Socioeconomic Variables

We link several census-, administrative- and survey-based data sources to Transmigration and other villages.

Population Census Data, 2010. The 2010 Population Census contains information on 237,641,326 Indonesian residents, and was produced by BPS-Statistics Indonesia (or BPS). We use a version of the census available at the Harvard Library Government Documents Group. This dataset includes village and sub-village identifiers, complete individual names, and a host of individual characteristics, including gender, relation to household head, birth information (month-year and district), marital status, education, and district of residence in 2005, educational level, sector of employment, religion, ethnicity, ability to speak Indonesian, and primary language spoken at home. The latter two questions on language use were not asked in the last complete-count Census in 2000 (described below). For ethnicity, each individual is asked to report the single ethnicity to which they feel closest. This was a free-response question and resulted in over 1,330 unique ethnic identities, 716 of which have at least one individual in Transmigration villages.

We use the Census records to compute several measures of local diversity. First, we construct measures of village-level fractionalization (F) and polarization (P) based on self-reported ethnicities, native linguistic-distance-adjusted ethnic groups, and aggregated super-ethnic groups determined by Indonesian demographers (see Section 7.1). Second, we construct sub-village-level F and P using neighborhood identifiers (*satuan lingkungan setempat* or SLS) reported by enumerators. Third, we construct indicators for whether one's two next-door neighboring households have the same ethnicity. We define household ethnicity by taking the modal ethnicity within each household. We define next-door neighbors as those two over in the listing roster within each neighborhood are on either side of a given unit. For example, my household number is 5, then I am adjacent to households 3 and 7 with 4 and 6 being across the street. This is in line with the zigzag enumeration method described in the Census enumerator's manual. Fourth, we follow the literature on segregation and use information on Census blocks, which partially overlap with SLS, to construct the Alesina and Zhuravskaya (2011) measure of ethnic segregation within the village as detailed in Section 7.1.

We also use the Census to construct four nation-building outcomes. First, we construct an indicator for the national language (*Bahasa Indonesia*) being the primary one used at home. All individuals over the age of 5 respond to this question. Note that while individuals could report Indonesian as a primary language at home, they could not report Indonesian as their ethnicity. Second, we construct an indicator for the native ethnic language being the primary one use at home. We define the native language of ethnic group e as the modal language other than Indonesian spoken by members of e in the whole of

[1] In the online appendix of Bazzi et al. (2016), we describe in detail how we constructed this dataset from the original Transmigration census.

Indonesia. Third, for each household head, we can identify whether they are in an interethnic marriage. We identify such status for 453,300 couples in Transmigration villages but restrict attention in the empirical analysis to those that were below the legal minimum age of marriage in the year of settlement to ensure that we identify plausibly new marriages.

Fourth, we use individual names to construct indices measuring how precise a child's name is in identifying his/her membership to one of four groups: (i) Indonesian-speaking households, (ii) intermarried households, (iii) urban households, and (iv) one's native ethnic group. We describe index (i) with reference to equation (9). The procedure for constructing indices (ii) and (iii) is identical but just replaces *homeIndo* with intermarried and urban residence indicators, respectively where intermarried equals one if the child's parents are intermarried. For (iv), we generalize the likelihood expression in equation (9) as follows:

$$\text{ETHNIC SCORE}_n = \frac{\mathbb{P}\left(name = n \,|\, own\text{-}ethnicity = 1\right)}{\sum_e \mathbb{P}\left(name = n \,|\, ethnicity = e\right)}, \tag{D.1}$$

where distinct names are indexed by n. We construct the probabilities for each name n using the entire population of 230+ million Indonesians living outside Transmigration villages and then apply the scores to children born after the year of settlement for the given Transmigration village. Note that we focus on measures based on individual names but exclude those with names that are not shared by at least 100 people in the entire country. Fryer and Levitt (2004) implement a similar cutoff rule, and our results are robust to other cutoffs.

For robustness, in Appendix Table A.13, instead of using actual names, we used each name's metaphone and double metaphone scores as arguments in ETHNIC SCORE$_n$ (Philips, 2000). Lawrence Philips' metaphone and double metaphone algorithms take each name and return a rough approximation for how each name sounds. By grouping together similar-sounding names prior to calculating the indices, we avoid issues related to misspellings in the Census data, common spelling differences, and consequently, we reduce problems related to unique names.[2]

Population Census Data, 2000. The 2000 Population Census, also fielded by BPS, contains similar information as the 2010 Census except that it does not include questions about language or individual names. This too was meant to be a complete-count, universal coverage census, but the provincial offices of BPS had to estimate the data for some of the areas due to to communal violence following the 1998 political transition.[3] This was the first Census since the 1930 Census conducted by the Dutch colonial authority to ask about ethnicity. Like the 2010 Census, ethnicity is self-reported, and at the time, individuals reported 1,033 unique ethnic identities. We use the 2000 Census to construct the population share of ethnic groups native to Java/Bali and restricting to those born in Java/Bali. These are used as instrumental variables to capture ethnic diversity among the original transmigrants from Java/Bali. We also construct measures of F and P as well as segregation and intermarriage rates.

Village Potential (*Podes*), 1999, 2002, 2005, 2008, 2011, 2014. We use multiple rounds of the triennial *Podes* to construct outcomes of interest. First, we construct an index of five village-provided public goods: safe drinking water, garbage collection, public toilet facilities, 4-wheel road access, and streetlights. We construct binary indicators for each from every year beginning in 2002 and then take the average of the year-specific average across the five indicators. Second, we construct an indicator for the occurrence of any ethnic conflict from 2002 to 2014. Third, we measure voter turnout in the first democratic election of 1999 as reported in *Podes* of that year. Fourth, we measure political preferences in

[2] We used an open-source implementation of these algorithms in python, which can be found here: https://pypi.org/project/Metaphone/.

[3] The areas where data were estimated instead of enumerated are in the provinces of Aceh, Maluku, Papua, and Central Sulawesi (Surbakti et al., 2000).

that election using *Podes* 2002, which records the top-3 parties in terms of national legislative vote shares at the village level. We classify these parties based on whether they espoused the inclusive national ideology of *Pancasila* (Baswedan, 2004). Most non-*Pancasila* parties adhered to Islam as their ideology. Finally, we also use *Podes* 2002 to measure distance to the district capital, which is based on reported travel distance by the village head.

***Susenas*, 2000–2012.** We use data from the annual National Socioeconomic Survey (*Susenas*) to examine social attitudes and household expenditures.

For social attitudes, we employ the Sociocultural (*Sosial Budaya*) Module from the 2012 round in which household heads are asked a host of questions capturing social attitudes. We explore eight questions from relevant domains of social capital, namely:

1. Do you participate in activities to provide public goods (e.g., building public facilities, communal clean up) in your community?
2. Do you participate in social activities (e.g., ROSCA, sports, arts) in your community?
3. Are you pleased with the activities of people from other ethnic groups in your community?
4. How much do you trust your neighbor to watch your children (aged 0-12) if no adult is home?
5. How much do you trust your neighbor to watch your home if all household members are away?
6. Do you feel safe living in this community?
7. How easy is it to ask neighbors (who are non-relative) for help when you have financial difficulties?
8. Do you participate to help neighbors who endured misfortunes (e.g., death, illness)?

Respondents then provided responses on a 1 to 4 integer scale indicating the strength of their agreement.

Next, we construct a measure of mean household expenditures per capita using all available years in which each village is covered by the survey from 2000 to 2012. Given the random sampling, some villages are observed multiple times, and others are not observed at all. We take a simple average across all households and all years.

Indonesia Family Life Survey (IFLS). IFLS is a longitudinal household dataset that was collected between 1993 and 2015. Five waves of data collection had been conducted in 1993, 1997, 2000, 2007, and 2014. Over the span of more than two decades, IFLS tracked all individuals from the 7,224 households in the first wave with a very low attrition rate (of less than 10 percent) between IFLS1 and IFLS5 (Strauss et al., 2016). In particular, it tracks individuals who left their original (IFLS1) households, either due to the formation of new households or emigration out of their villages within their original district.

IFLS has a rich set of variables. Included among the rich set of IFLS variables are (reported) ethnicity, the ethnicity of an individual's mother and father, language spoken at home, and discriminative attitudes (with respect to ethnicity). These variables are collected for all members of the surveyed households members. We use these variables to identify individuals who were brought up in households that use Indonesian at home, as well as those whose parents are of mixed ethnicity.

NOAA Data on Light Intensity, 2010. To proxy for economic activities at the local level, we make use of an innovative technique, developed by Henderson et al. (2012), which uses satellite data on nighttime lights. Daily between 8:30 PM and 10:00 PM local time, satellites from the United States Air Force Defense Meteorological Satellite Program (DMSP) record the light intensity of every 30-arc-second-square of the Earth's surface (corresponding to roughly 0.86 square kilometers). DMSP cleans this daily data, dropping anomalous observations, and provides the public with annual averages of light intensity from multiple satellites. After averaging the data across multiple satellites, we obtain annual estimates of light intensity for every 30-arc-second square of the Earth's surface in 2010. Henderson et al. (2012) show that across countries, growth in night-lights (measured as the change in the spatial average digital number of light intensity over time) is linearly related to growth in output.[4] See Bazzi et al. (2016) for references on the quality of this proxy for income in the Indonesian context.

D.3 Spatial, Topographical, and Agroclimatic Variables

We include geographical characteristic and climatic variables to construct the controls for natural endowments. These include measures of: (i) topography (land area, elevation, slope, ruggedness, and altitude), (ii) pre-program market access (distance to (sub)district capitals, roads, rivers, and the sea coast), and (iii) soil quality such as texture, drainage, sodicity, acidity, and carbon content. Many of these variables are explicitly listed in program manuals from 1978 in the MOT archives that provided guidance for site selection. We construct these variables from a variety of sources. Below, we briefly discuss the construction and sources of these variables. The online appendix of Bazzi et al. (2016) provides more details of the variable construction procedures.

Distances and Map Projection. Data for the shapefiles for Indonesia's rivers, roads, major cities, and coast lines were all provided by Indonesia's Department of Public Works (*Departemen Pekerjaan Umum*). Using GIS, we constructed the distance from each village polygon in the dataset to the coast, the nearest river, the nearest road, and major cities using the Euclidean distance tools from ArcView.

Slope, Aspect, and Elevation Data. We construct the topographical variables using raster data from the *Harmonized World Soil Database* (HWSD), Version 2.0 (Fischer et al., 2008).[5] We use the raster data to compute the average elevation, slope, and aspect over the entire polygon for each village. For the slope variables, we the average share of each village corresponding to each slope class (0-2 percent, 2-4 percent, etc.) using ArcView.

Ruggedness. A 30 arc-second ruggedness raster was computed for Indonesia according to the methodology described by Sappington et al. (2007), and village-level ruggedness was recorded as the average raster value. The authors propose a Vector Ruggedness Measure (VRM), which captures the distance or dispersion between a vector orthogonal to a topographical plane and the orthogonal vectors in a neighborhood of surrounding elevation planes.

Soil Quality Covariates. HWSD provides detailed information on different soil types across the world. For Indonesia, the data come from the FAO-UNESCO Soil Map of the World (FAO 1971-1981). We created for each village the following measures of soil types: percentage of land covered by coarse, medium, and fine soils, percentage of land covered by soils with poor or excessive drainage, average organic carbon percentage, average soil salinity, average soil sodicity, and average topsoil pH.

Rainfall and Temperature, 1948–1978. The database of Matsuura and Wilmott (2012a,b) at the Department of Geography, University of Delaware compiles monthly temperature and rainfall data across the globe. The monthly data for Southeast Asia come from the Global Historical Climatology Network v2 (GHCN2) database, which were interpolated to estimate monthly precipitation and temperature to a 0.5×0.5 degree (or 55 km) resolution grid (Matsuura and Wilmott, 2012a,b). For the districts in our dataset, we averaged the numbers provide by the database for the period of 1948–1978 to obtain the predetermined measures of rainfall and temperature.

Measuring Agroclimatic Similarity. We use a measure of agroclimatic similarity from Bazzi et al. (2016) to construct the inequality indices used in Table 7. The agroclimatic similarity measure captures the similarity in the agroclimatic environments between migrant origins and destinations. As in Bazzi

[4] The DMSP-OLS Nighttime Lights Time Series Version 4 datasets can be downloaded here: http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html.

[5] Data from the HWSD project are publicly available at http://www.iiasa.ac.at/Research/LUC/luc07/External-World-soil-database/HTML/index.html?sb=1

et al. (2016), we construct this variable using the spatial, topographical, and agroclimatic variables described above. All land attributes are either time-invariant or measured before the villages we study were created, and hence do not reflect settler activities.

The agroclimatic similarity between an individual's origin i and her destination j is defined as:

$$agroclimatic\ similarity_{ij} \equiv \mathcal{A}_{ij} = (-1) \times \mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{D.2}$$

where $\mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_j\right)$ is the agroclimatic distance between locations i and j, using a metric defined on the space of agroclimatic characteristics. We observe origins at the district-level and hence construct the index based on measures of $\mathbf{x}$ in the destinations at that same spatial frequency. We use the sum of absolute deviations as the distance metric, converting each characteristic to z-scores before taking the absolute difference between origins and destinations. Then, $\mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_g |x_{ig} - x_{jg}|$ projects these differences in G dimensions onto the real line. We multiply by (-1) so that larger differences correspond to lower values of agroclimatic similarity.

An agroclimatic similarity index for location j is then calculated by aggregating the individual $\mathcal{A}_{ij}$ across i using population weights:

$$agroclimatic\ similarity_{j} \equiv \mathcal{A}_{j} = (-1) \times \sum_{i=1}^{I} \pi_{ij}\, \mathrm{d}\left(\mathbf{x}_i, \mathbf{x}_j\right), \tag{D.3}$$

where π_{ij} is the share of migrants residing in Transmigration village j who were born in district i (calculated using the 2000 Population Census microdata). We use π_{ij} terms based on all individuals born in Java/Bali.

D.4 Linguistic Distance: World Language Mapping System (WLMS) and *Ethnologue*

We use the *World Language Mapping System* (WLMS) to construct our measure of linguistic distance. WLMS uses the sixteenth edition of the *Ethnologue* database and maps each of 6,909 living languages recorded in the database to its relevant location. There are more than 700 ethnolinguistic groups in its entries for Indonesia, including eight ethnolinguistic groups indigenous to Java/Bali.

We then map each unique ethnicity in the 2000 and 2010 Population Censuses to corresponding groups in the *Ethnologue*. For 2000, we use WLMS's language-to-location mapping to define the native local language at each settlement. For 2010, we use the individual-level information on the home language available in the 2010 Census to define the native language for each ethnic group. We assign the modal (non-Indonesian) language spoken by an ethnic-group in a province to be its native language.[6]

We then match that language with those recorded in *Ethnologue*, using ISO language codes. In cases with duplicates, we pick the match associated with by largest population. In some cases, we are unable to match a daily language to a corresponding ISO code, in which case we would impute linguistic distances based on the average of the non-missing languages. This affects under 4 percent of the entire population age $5+$ in Transmigration villages (63,741 out of 1.8 million people).

We use the linguistic classifications in *Ethnologue* to construct the distance, δ_{ij}, between ethnic groups i and j, which is used to construct the exogenous linguistic-distance-adjusted polarization index in Table 7 and Figure 6. As elaborated in Section 7.1, $\delta_{ij} = 1 - \tau_{ij}^{\kappa}$ where $\tau_{ij} = \left(\frac{\text{branch}_{ij}}{\max(\text{branch}_i,\text{branch}_j)}\right)$, branch$_{ij}$ is the number of shared language tree branches, and $\max(\text{branch}_i, \text{branch}_j)$ is the maximum number of among the two. We set κ to 0.5 or 0.05 as in prior literature. We also use the WLMS shapefiles to identify the ethnolinguistic homeland covering each Transmigration village (see Appendix Table A.11).

[6] We assign the province-specific modal language for each ethnic group to allow for regional variations in the spoken language for the larger ethnic groups. For example, people who identified themselves as Malays in certain parts of South Sumatra and South Kalimantan could often be recorded as speaking different languages (e.g., Palembang and Banjar respectively).

Effective Linguistic Distance, $\tilde{\delta}_{ij}$. In Figure 6, we calculated polarization based on endogenous language choices rather than exogenous native language based on *Ethnologue* as above. The endogenous language choice is simply the one reported by each individual in the 2010 Census. To account for the endogenous linguistic distance between ethnicity i and ethnicity j, we adjust the linguistic distance between them to be a weighted average of linguistic distances across all possible combinations of endogenous language choices ℓ_i and ℓ_j spoken by individuals in ethnic groups i and j: $\tilde{\delta}_{ij} = \sum_{\ell_i \ell_j} w_{\ell_i \ell_j} \min(\delta_{\ell_i \ell_j}, \delta_{ij})$.

We apply a population-based weight $w_{\ell_i \ell_j}$ to each language pair, where the sum of the weights across all language pairs equals 1. We assume that speaking a common language serves to reduce the linguistic distance between the two groups (otherwise, they can always revert to their own ethnic language): $\min(\tau_{\ell_i \ell_j}, \tau_{ij})$ is the minimum of the linguistic distance between the two languages and the linguistic distance between the native languages of the two groups. If everyone decides to speak their own native language, then, $\min(\delta_{\ell_i \ell_j}, \delta_{ij}) = \delta_{ij}$, and we would not see a reduction in the endogenous linguistic distance.

To better illustrate this calculation, consider a village whose entire population belongs to only two ethnicities. In this village, there are 10 Javanese and 15 Sundanese. Among the 10 Javanese, 4 speak Javanese at home, and 6 speak Indonesian. Among the 15 Sundanese, 7 speak Sundanese, and 8 speak Indonesian. In this case, there are four possible language pairs: Javanese and Sundanese, Javanese and Indonesian, Indonesian and Sundanese, Indonesian and Indonesian. To find the appropriate weight $w_{\ell_i \ell_j}$ for the Javanese-Sundanese language pair, we multiply the share of ethnically Javanese people who speak Javanese, $\frac{4}{10}$, by the share of ethnically Sundanese people who speak Sundanese, $\frac{7}{15}$. Repeating the calculation for all language pairs, we obtain these weights:

	Sundanese	Indonesian
Javanese	$\frac{28}{150}$	$\frac{32}{150}$
Indonesian	$\frac{42}{150}$	$\frac{48}{150}$

Intuitively, given 10 Javanese and 15 Sundanese people in the village, there are $10 \times 15 = 150$ possible inter-ethnic interactions between individuals in the village. This is captured by the denominator in each of the weights. Among the 4 ethnically Javanese people who speak Javanese and the 7 ethnically Sundanese people who speak Sundanese, there are $4 \times 7 = 28$ possible interactions between them, captured in the numerator. Similar calculations apply to the other groups. Taking the sum of these weights, we see that $\frac{28}{150} + \frac{32}{150} + \frac{42}{150} + \frac{48}{150} = 1$. For any ethnicity i and ethnicity j, we can extend this example to include any number of languages.

References

Alesina, A. and E. Zhuravskaya, "Segregation and the Quality of Government in a Cross-Section of Countries," *American Economic Review*, 2011, *101*, 1872–1911.

Arnall, A., D. S. G. Thomas, C. Twyman, and D. Liverman, "Flooding, resettlement, and change in livelihoods: evidence from rural Mozambique," *Disasters*, 2013, *37* (3), 468–488.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein, "Improving refugee integration through data-driven algorithmic assignment," *Science*, 2018, *359* (6373), 325–329.

Barnhardt, S., E. Field, and R. Pande, "Moving to Opportunity or Isolation? Network Effects of Randomized Housing Lottery in Urban India," *American Economic Journal: Applied Economics*, 2017, *9* (1), 1–32.

Baswedan, A. R., "Political Islam in Indonesia: present and future trajectory," *Asian Survey*, 2004, *44*, 669–690.

Bazzi, S., A. Gaduh, A. Rothenberg, and M. Wong, "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia," *American Economic Review*, 2016, *106* (9), 2658–2698.

Becker, S. O., I. Grosfeld, P. Grosjean, N. Voigtländer, and E. Zhuravskaya, "DP12975 Forced Migration and Human Capital: Evidence from Post-WWII Population Transfers," 2018.

Boustan, L., "Racial Residential Segregation in American Cities," in Nancy Brooks and Gerrit-Jan Knaap, eds., *Oxford Handbook of Urban Economics and Planning*, Oxford University Press: UK, 2011, pp. 318–339.

Cameron, A. C., J. B. Gelbach, and D. L. Miller, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

_ , _ , and _ , "Robust Inference with Multiway Clustering," *Journal of Business & Economic Statistics*, 2011, *29* (2).

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

d. Sherbinin, A., M. Castro, F. Gemenne, M. Cernea, S. Adamo, P. M. Fearnside, G. Krieger, S. Lahmani, A. Oliver-Smith, A. Pankhurst, T. Scudder, B. Singer, Y. Tan, G. Wannier, P. Boncour, C. Ehrhart, G. Hugo, P. Balaji, and G. Shi, "Preparing for Resettlement Associated with Climate Change," *Science*, 2011, *344*, 456–457.

de Wet, Chris, "The Experience with Dams and Resettlement in Africa," 2000. Consultancy report written for the World Commission on Dams.

Dutch Refugee Council, "Housing for Refugees in the European Union," Technical Report, Dutch Refugee Council 1999.

Fan, J. and I. Gijbels, *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Vol. 66, CRC Press, 1996.

Fischer, G., F. Nachtergaele, S. Prieler, H.T. van Velthuizen, L. Verelst, and D. Wiberg, "Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008)," 2008.

Fryer, R. G. and S. D. Levitt, "The Causes and Consequences of Distinctively Black Names," *Quarterly Journal of Economics*, 2004, *119* (3), 767–805.

Glitz, A., "The Labor Market Impact of Immigration: A Quasi-Experiment Exploiting Immigrant Location Rules in Germany," *Journal of Labor Economics*, 2012, *30* (1), 175–213.

Hall, A., "Making People Matter: Development and the Environment in Brazilian Amazonia," *International Journal of Contemporary Sociology*, 1993, *30* (1), 63–80.

Harari, M. and E. LaFerrara, "Conflict, climate, and cells: a disaggregated analysis," *Review of Economics* *and Statistics*, 2018, *100* (4), 594–608.

Henderson, J. V., A. Storeygard, and D. N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, 2012, *102* (2), 994–1028.

Hsiang, S. M., M. Burke, and E. Miguel, "Quantifying the influence of climate on human conflict," *Science*, 2013, *341* (6151), 1235367.

Matsuura, K. and C. J. Wilmott, "Terrestrial Air Temperature: 1900-2010 Gridded Monthly Time Series (V 3.01)," 2012.

_ **and** _ , "Terrestrial Precipitation: 1900-2010 Gridded Monthly Time Series (V 3.02)," 2012.

Philips, L., "The double metaphone search algorithm," *C/C++ users journal*, 2000, *18* (6), 38–43.

Polikoff, A., *Sustainable Intergration or Inevitable Resegregation*, University of North Carolina Press,

Robinson, P. M., "Root-N-consistent Semiparametric Regression," *Econometrica*, 1988, *56* (4), 931–954.

Rubin, D. B., "The Bayesian Bootstrap," *The Annals of Statistics*, 1981, *9* (1), 130–134.

Sandholm, W. H., *Population games and evolutionary dynamics*, MIT press, 2010.

Sappington, J. M., K. Longshore, and D. Thompson, "Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study using Bighorn Sheep in the Mojave Desert," *Journal of Wildlife Management*, 2007, *71* (5), 1419–1426.

Schlag, Karl H, "Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits," *Journal of economic theory*, 1998, *78* (1), 130–156.

Strauss, J., F. Witoelar, and B. Sikoki, "The Fifth Wave of the Indonesia Family Life Survey (IFLS5): Overview and Field Report," RAND Working Paper WR-1143/1-NIA/NICHD, RAND March 2016.

Surbakti, S., R. L. Praptoprijoko, and S. Darmesto, "Indonesia's 2000 Population Census: A Recent National Statistics Activity," Technical Report, United Nations Economic and Social Commission on Asia and Pacific 2000.

Taye, M. and I. Mberengwa, "Resettlement: A Way to Achieve Food Security? A Case Study of Chewaka Resettlement Scheme, Oromia National Regional State, Ethiopia," *Journal of Sustainable Development in Africa*, 2013, *15* (1), 141–154.

Wong, M., "Estimating Ethnic Preferences Using Ethnic Housing Quotas in Singapre," *Review of Economic Studies*, 2013, *80* (3), 1178–1214.

World Bank, *Resettlement and Development: The Bankwide Review of Projects Involving Involuntary Resettlement, 1986–1993*, Washington, DC: International Bank for Reconstruction and Development, 1999.

Young, A., "Inequality, the Urban-Rural Gap, and Migration," *The Quarterly Journal of Economics*, 2013, *128* (4), 1727–1785.

_ , "Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections," *Unpublished Manuscript*, 2016.

Young, H. P., "The evolution of social norms," *Annual Review of Economics*, 2015, *7* (1), 359–387.