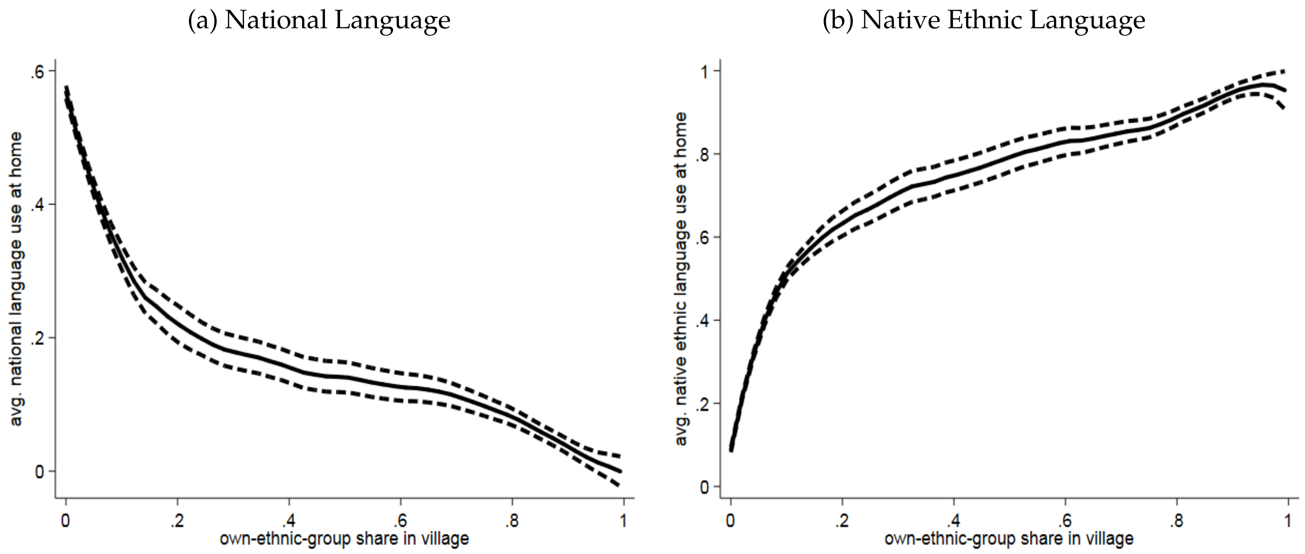
Online Appendix

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



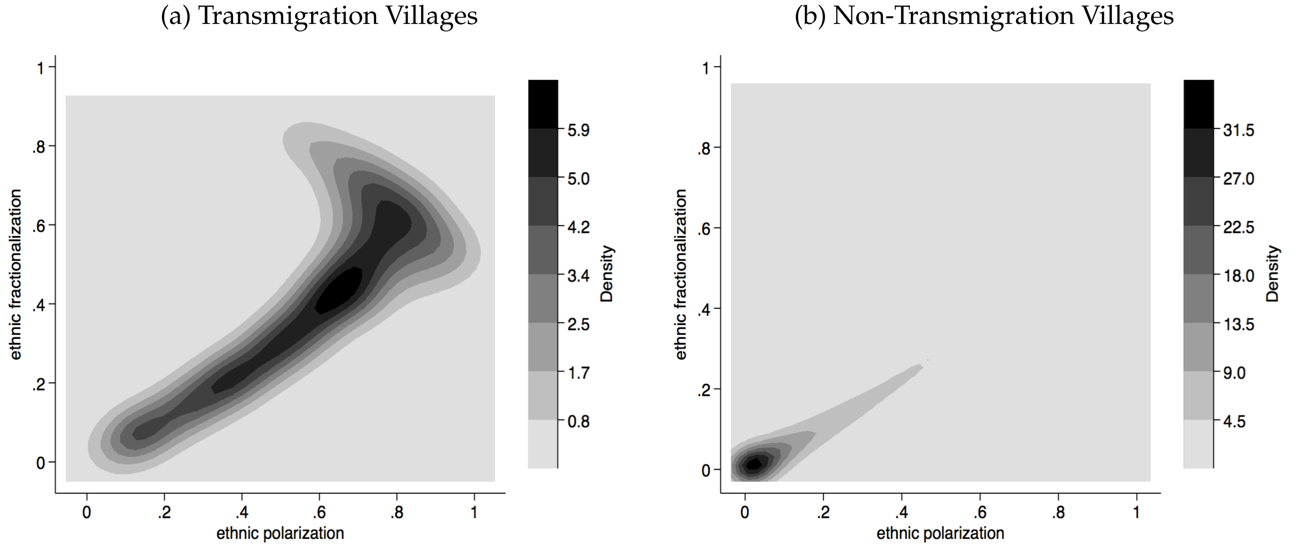
(a) National Language

(b) Native Ethnic Language

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



(a) Transmigration Villages

(b) Non-Transmigration Villages

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

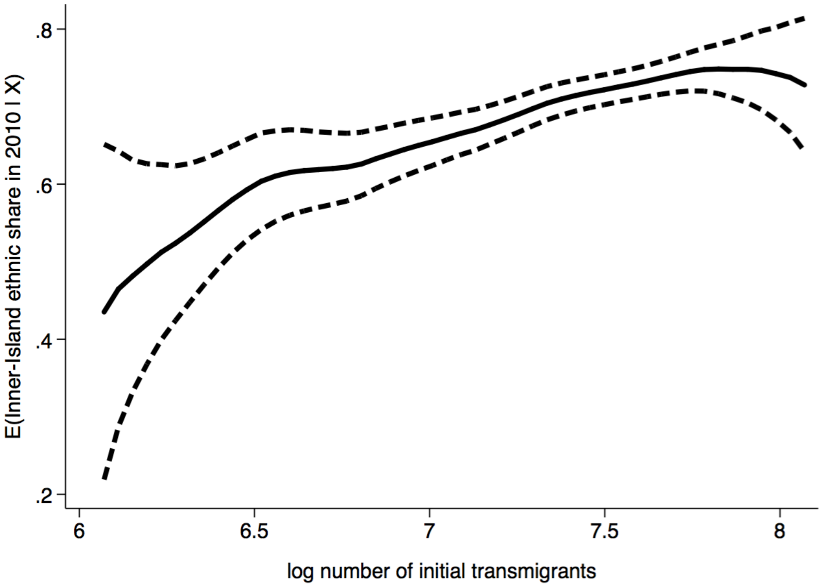
Notes: Standard errors are clustered by district.

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.

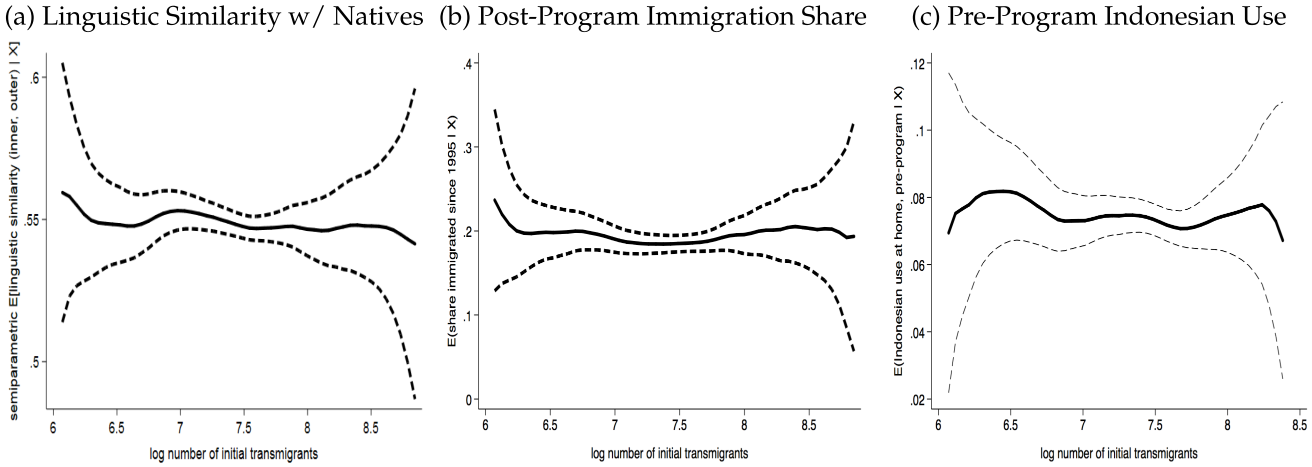
Figure A.3: Initial Transmigrant Assignment and Long-Run Inner-Island Ethnic Share



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 $\mathbf{x}$). We find no systematic evidence of such patterns looking across this large set of outcomes.

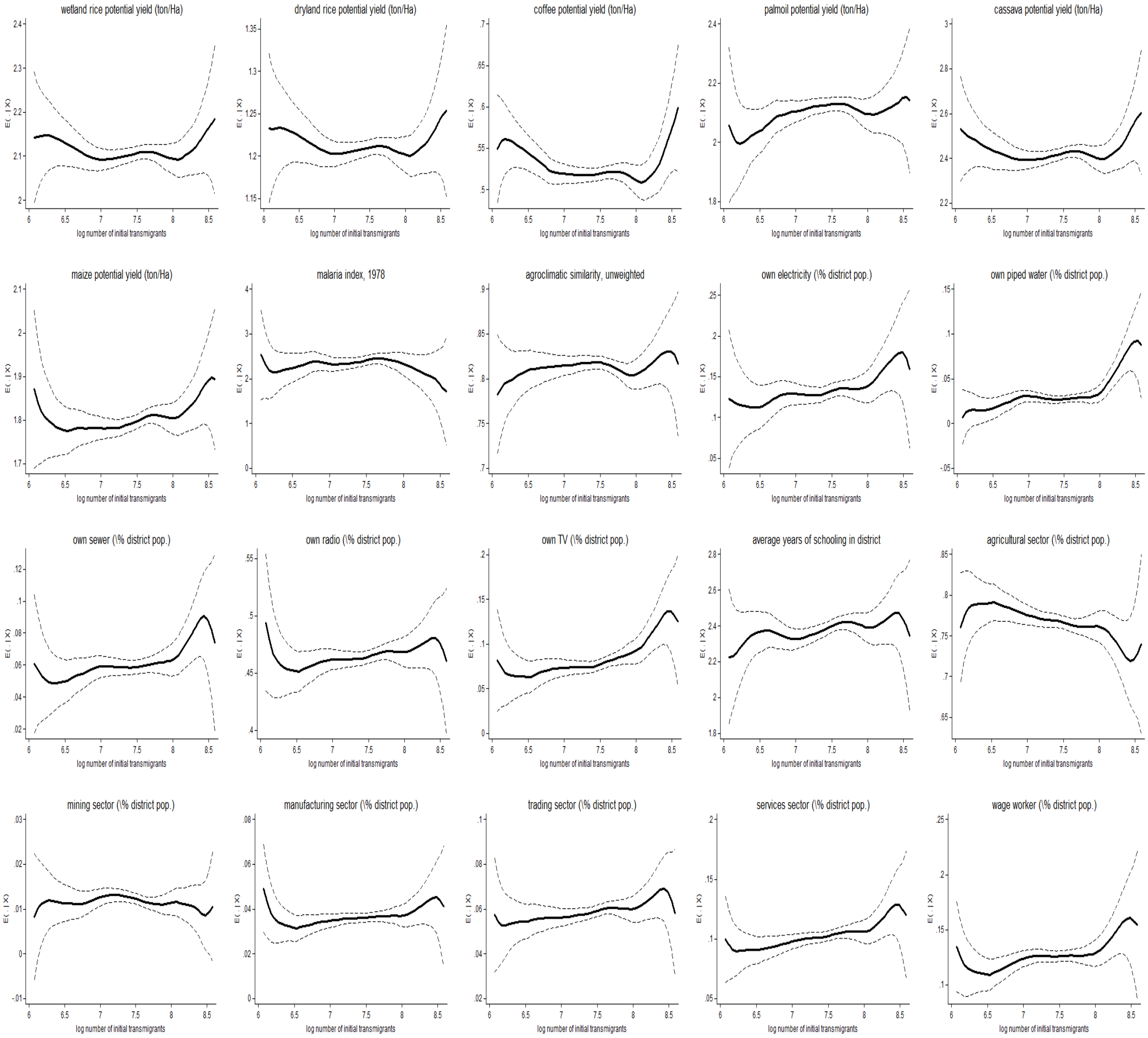
Figure A.4: Initial Transmigrant Allocation Uncorrelated with Proxies for Sorting


(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 $\mathbf{x}$ 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:

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

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.

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

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:

$$
\begin{aligned}
\dot{G}^N = \sum_j p_j \dot{g}_j^N &= \sum_j p_j A_j \left(1-\sigma\right)\theta (\bar{\pi} - p_j \pi_j) - \left(1-\sigma\right) \sum_j \sum_{k\neq j} [p_j p_j A_j (1-\pi_k) D_j] - \sum_i p_j A_j \gamma \\
&= \left(1-\sigma\right)\theta \Big(\bar{A}\overline{\pi} - \sum_i A_j \pi_j p_j^2 \Big) - \left(1-\sigma\right) \sum_i \sum_{j\neq i} [p_j p_j D_j A_j (1-\pi_k)] - \bar{A}\gamma \\
&= \left(1-\sigma\right)\theta \Phi \Big(1 - \sum_j \phi_j p_j^2 \Big) - \left(1-\sigma\right) \sum_j \sum_{k\neq j} p_j p_k T_{jk} - \bar{A}\gamma
\end{aligned}
\tag{B.3}
$$

where $\bar{A} = \sum_j p_j A_j$, $\overline{\pi}$ 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:

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

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

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*

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

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:

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

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

$$
\vdots
$$

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

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

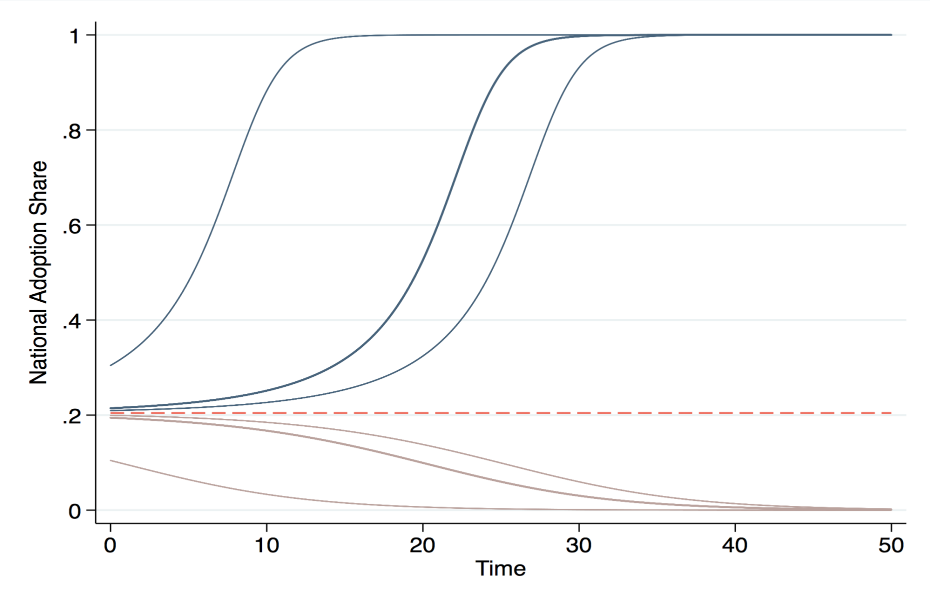
As an example, consider a 5-group village with groups of equal size, so that $p_1 = p_2 = p_3 = p_4 = p_5 = 0.2$. Fix $\theta = 1$, $D_j = 0.5p_j$, $\gamma = 0.1$, and $\sigma = 0$. Because of symmetry, we can just focus on the evolution of one group's shares. In this example, $\pi_j^* \approx 0.205$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

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

, 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:

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

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:

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

From the properties of geometric series, we can write:

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

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:

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

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:

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

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.