ONLINE APPENDIX
Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia

By Samuel Bazzi, Arya Gaduh, Alexander D. Rothenberg and Maisy Wong *

This supplementary appendix comprises three sections. Section A describes in detail the data sources and variable construction. Section B provides additional details, including figures and tables on results noted briefly in the paper. Section C provides more information on the optimal assignment exercise implemented in the paper.

Appendix A: Data Appendix

For the analysis, we combine information on transmigrant placements with demographic, economic, linguistic, and spatial and agroclimatic characteristics. Table A.1 summarizes the datasets used for this analysis. Each of the datasets is described in detail in the following sections.

1. *Transmigration Census and Maps*

The main source of information on where transmigrants are placed is the census of Transmigration sites established between 1952 and 1998 produced by the Ministry of Manpower and Transmigration (MOMT). The census contains information on the physical locations and names of the Transmigration sites, the years they were established, and the number of people and households sent to each site at the time of the initial settlement. From the physical location information, we can identify the villages where these sites reside.

A total of 2,625 sites are listed in the dataset. We manually matched the village names listed there with those listed in the 2000 Indonesian Population Census data and identified 1,702 Transmigration villages.[1] This paper analyzes the Transmigration sites established during Indonesia's Third and Fourth Development Plan (1979-1988) in the Outer islands excluding Papua, which leaves us

* Bazzi: Department of Economics, Boston University, 270 Bay State Rd., Boston, MA 02215 (e-mail: sbazzi@bu.edu); Gaduh: Department of Economics, Sam M. Walton College of Business, University of Arkansas, Business Building 402, Fayetteville, AR 72701-1201 (e-mail: agaduh@walton.uark.edu); Rothenberg: RAND Corporation, 1200 South Hayes St., Arlington, VA 22202-5050 (e-mail: arothenb@rand.org); Wong: Wharton Real Estate, University of Pennsylvania, 3620 Locust Walk, 1464 SHDH, Philadelphia, PA 19104-6302 (e-mail: maisy@wharton.upenn.edu).

[1] In most cases, each site corresponds to a village; however, there are some cases where a village comprises multiple sites.

with 911 Transmigration villages.[2]

Table A.1—: Summary of Datasets

Dataset	**Description**	**Obs. Unit**
Transmigrant placement		
Transmigration census	Location of Transmigration sites; the number of households and individuals, and years placed in each site.	Transmigration site
Reppprot Maps	1:250,000 Regional Physical Planning Program for Transmigration (Reppprot) maps that include "recommended development area" or RDAs.	
Demographic Variables		
Population Census, 2000	Full dataset: Highest level of schooling, ethnicity, sectoral employment, birth information (year and month, district), district of residence in 1995.	Individual
Intercensal Survey (*SUPAS*), 1985	Intercensal survey with information on migration between districts.	Individual
Population Census, 1980	IPUMS subset (5 percent of population). Housing characteristics, highest level of schooling, sectoral employment, birth information (year and month, province), province of residence in 1975.	Individual
Economic Variables		
Podes 2003	Crop types and yield; share of farmland; share of land by legal status, i.e., public, private with certificate, private under Islamic trust (*waqf*).	Village
Agricultural Census 2003, 1963	Household-level land ownership	Household: full (2003); table by district (1963)
FAO/PriceSTAT	Crop-specific prices.	National
NOAA Light Intensity	Light intensity data, 2010.	30-arc-second grid
Susenas, 2004	Household-level rice productivity, HH-head education level	Household
Linguistic Variable		
World Language Mapping System (WLMS), Ethnologue	Language by ethnic groups and the shared branches of different languages.	
Agroclimatic Variables		
FAO-GAEZ	Potential output (tons/ha) for major crops (see Costinot, Donaldson and Smith, 2016; Nunn and Qian, 2011).	5-arc-minute grid
GIS Map - Dept. Public Works	Village area, distance to coast, roads and others.	
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, 2012*b*) and temperature (Matsuura and Wilmott, 2012*a*), 1948-1978.	Monthly, interpolated to 0.5 × 0.5 degree grid

[2] A small part of the program involved involuntary resettlement of households displaced by disasters and

The ministry also produces a set of maps that identify the locations of planned Transmigration sites that were ultimately not selected for the program. Outlines of these planned sites—also known as "recommended development areas" (RDAs)—were digitally traced using GIS software. We overlay the traced RDA maps onto digital maps of Indonesia's administrative boundaries that were produced by BPS for use in fielding the 2000 and 2010 Household Censuses.[3] We identified a total of 907 "RDA villages" in the administrative boundary file that shared any area with the RDA polygons. These villages act as the control villages in our analysis of the average treatment effect of the program.

2. Demographic and Economic Variables

Population Census Data, 2000

Indonesia's 2000 Population Census is a dataset issued by Indonesia's Central Statistical Agency, *BPS Statistics* (hereafter, BPS), that was designed as a complete enumeration of the individual members of every household in Indonesia with 100 percent coverage. However, due to riots and communal violence following the political transition, the population numbers for the provinces of Aceh, Maluku, Papua, and Central Sulawesi had to be estimated (instead of enumerated) by the provincial statistical offices (Surbakti, Praptoprijoko and Darmesto, 2000).

The census contains information on the respondents' religion, ethnicity, birth information (year, month, and district), as well as the sector of their employment (if they were working) and their district of residence in 1995. It also includes questions on the respondents' sex, marital status, education, and main activities in the past week. We aggregate the individual-level observations to construct village-level demographic variables and population weights for the similarity indices, and use the individual-level observations to examine occupation choice among the transmigrants.

Village Potential (Podes), 2003

We construct the agricultural outcome variables using Indonesia's Village Potential (*Podes*) survey, also issued by BPS. *Podes* is an administrative census of all villages in Indonesia that collects a rich set of village-level data, and has been conducted approximately every three to four years since 1976. We use, among other variables, the area planted and output by crop during the 2001-2002 agricultural season which is available in the 2003 *Podes.* All crops besides rice and *palawija* (maize, cassava, sweet potato, soy, and groundnut) are classified as cash crops in keeping with agronomic literature and policy in Indonesia.

infrastructure development (Kebschull, 1986). We exclude strategic settlements in Maluku and Papua associated with the Indonesian military as part of its territorial management system (Fearnside, 1997). We also omit Papua entirely due to concerns about data quality.

[3] These administrative boundary maps are extremely detailed, with the 2010 map containing over 75,000 polygons identifying the locations of different villages, their names, and the names of the provinces, districts, and subdistricts to which they belong.

Agricultural Census, 1963, 2003

To construct measures of landholdings inequality, we use the 2003 and 1963 Agricultural Census data issued by BPS. In the 2003 census, BPS collected detailed data on various farming activities from a universal census of farm households and enterprises across Indonesia. We use measures of landlessness (i.e., households with agricultural landholdings < 0.1 hectares) as well as the Pareto dispersion parameter to capture inequality among landholders. Meanwhile, for 1963, the district-level Pareto dispersion parameter was estimated from tabulations of several landholding size bins (from 0.1 to 5 ha and > 5 ha) by district using a maximum likelihood procedure. Details on these measures can be found in Bazzi (2015).

FAO/PriceSTAT

We calculate the revenue-weighted average productivity measures using yields data from *Podes* and crop prices from the FAO/ PriceStat database. Only crops whose 2001 and 2002 prices are available are included, namely avocados; bananas; beans, dry; beans, green; cabbages and other brassicas; carrots and turnips; cashew nuts, with shell; cassava; chillies and peppers, green; cinnamon (canella); cloves; cocoa beans; coconuts; coffee, green; cotton lint; cucumbers and gherkins; eggplants (aubergines); garlic; groundnuts, with shell; maize; maize, green; mangoes, mangosteens, guavas; natural rubber; nutmeg, mace and cardamoms; oil palm fruit; onions, dry; oranges; other bird eggs, in shell; palm oil; papayas; pepper (piper spp.); pineapples; potatoes; soybeans; spinach; sweet potatoes; tea; tobacco, unmanufactured; tomatoes; vanilla.

Susenas 2004, *SUPAS* 1985, and Population Census Data 1980

We also included three additional national-level datasets published by BPS. First, we use the 2004 *Susenas*, which is a household survey, to estimate household-level rice productivity regressions. Second, we use the 1985 Intercensal Survey, or *SUPAS* to calculate the inter-district migration flows in the early 1980s. Finally, we use the 5 percent population subset of the 1980 Indonesia Population Census that are available from IPUMS to construct the pre-1980 district-level characteristics that are included as control variables. In particular, we estimate district-level characteristics using the population that had been living in each district prior to 1979 when the transmigrant influx began. This ensures the exclusion of all potential transmigrants and the population of non-transmigrant immigrants that may have arrived in response to the program.

FAO GAEZ

We downloaded grid-cell level data of potential yield from FAO's Global Agro-Ecological Zones (GAEZ) for wetland rice, dryland rice, cocoa, coffee, palmoil, cassava and maize. To convert grid-level data to village level data, we aggregated across grids using area weights, calculated as the total area of the grid overlapping with the village, divided by the total village area. We use the same method to convert grid-level data to district-level data at the transmigrants' origin districts.

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, Storeygard and Weil (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, Storeygard and Weil (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]

3. Linguistic Similarity: World Language Mapping System (WLMS) and Ethnologue

To construct the linguistic distance measure, we use the *World Language Mapping System* (WLMS) which maps the languages documented in the *Ethnologue* database (Lewis, 2009) to the relevant locations across the world. WLMS maps the languages to locations using the sixteenth edition of the *Ethnologue* database, which contains 6,909 living languages around the world. Its entries for Indonesia contain more than 700 ethnolinguistic groups, including eight ethnolinguistic groups indigenous to Java/Bali.[5] We map the groups in Ethnologue and WLMS to those recorded in the 2000 Population Census.

A critical feature of the *Ethnologue* database is the linguistic trees for each of the 6,909 languages that are available. This linguistic tree, which shows how different languages and dialects are related among the different language families, enables the calculation of the linguistic distance, which is based on the number of shared branches between two languages. For each village j, we deem the native language to be the linguistic homeland polygon with maximum coverage of village area.

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

[5] The indigenous Java/Bali ethnicities include, in descending order of population shares in the Outer Islands: Javanese, Sundanese, Balinese, Madurese, Betawi, Tengger, Badui, and Osing.

4. Spatial, Topographical, and Agroclimatic Variables

We use agricultural and climatic variables to construct the agroclimatic similarity measures and control for natural endowments. These variables were created from a variety of sources and often were calculated with the assistance of GIS software (ArcView). This section describes those data in detail and how each of the variables were constructed.

Distances and Map Projection

To compute distances correctly, using linear units of measurement (e.g. kilometers), we made use of the Batavia Transverse Mercator (TM) 109 SE projected coordinate system in all of the GIS work. For each village polygon from the administrative boundary shapefile, we constructed the distance to the coast, the nearest river, the nearest road, and major cities using the Euclidean distance tools from ArcView. 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*).

Slope, Aspect, and Elevation Data

Topographical variables were created using raster data from the *Harmonized World Soil Database* (HWSD), Version 2.0 (Fischer et al., 2008).[6] The raster files are compiled from high-resolution source data and aggregated to 30 arc-second grids. These data are more detailed than other similar datasets used in the literature, such as the *Atlas of the Biosphere* data (used by Michalopoulos, 2012, among others), which is available at a 55 km resolution (0.5 degree grids), or the FAO's Global Agro-Ecological Zones (GAEZ) dataset (used by Costinot, Donaldson and Smith, 2016, among others), which is available at a 10 km resolution (5 arc-minute grids)

Elevation data were computed for each village as the average elevation over the entire village polygon, using raster data from HWSD.[7] Slope and aspect data were also recorded for each village and calculated similarly. Variables equal to the average share of each village corresponding to each slope class (0-2 percent, 2-4 percent, etc.) were constructed using ArcView.

[6] Data from the HWSD project are publicly available and can be downloaded here: http://www.iiasa.ac.at/Research/LUC/luc07/External-World-soil-database/HTML/index.html?sb=1. The terrain, slope, and aspect database provided by HWSD researchers was compiled from a high-resolution digital elevation map constructed by the Shuttle Radar Topography Mission (SRTM). SRTM data is also publicly available as 3 arc-second digital elevation maps (DEM) (approximately 90 meters resolution at the equator), available here: ftp://e0srp01u.ecs.nasa.gov/srtm/.

[7] The HWSD elevation raster file records the median elevation (in meters) for each 30 arc-second grid of the Earth's surface. The median is computed across space, from the values of all 3 arc-second cells in the SRTM database.

Ruggedness

A 30 arc-second ruggedness raster was computed for Indonesia according to the methodology described by Sappington, Longshore and Thompson (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.

To calculate the measure, one first calculates the x, y, and z coordinates of vectors that are orthogonal to each 30-arc second grid of the Earth's surface. These coordinates are computed using a digital elevation model and standard trigonometric techniques. Given this, a resultant vector is computed by adding a given cell's vector to each of the vectors in the surrounding cells; the neighborhood or window is supplied by the researcher. Finally, the magnitude of this resultant vector is divided by the size of the cell window and subtracted from 1. This results in a dimensionless number that ranges from 0 (least rugged) to 1 (most rugged).[8]

For example: on a flat (3×3) surface, all orthogonal vectors point straight up, and each vector can be represented by $(0, 0, 1)$ in the Cartesian coordinate system. The resultant vector obtained from adding all vectors is equal to $(0, 0, 9)$, and the VRM is equal to $1 - (9/9) = 0$. As the (3×3) surface deviates from a perfect plane, the length of the resultant vector gets smaller, and the VRM increases to 1.

Soil Quality Covariates

We also make use of the HWSD data for soil quality measures. HWSD provides detailed information on different soil types across the world. The HWSD data for Indonesia is taken from information printed in the FAO-UNESCO Soil Map of the World (FAO 1971-1981), a map printed at a 1:5,000,000 scale. For each village, we created 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 rainfall and temperature data are based on the database of Matsuura and Wilmott (2012*a*,*b*) at the Department of Geography, University of Delaware. They were compiled from a number of sources; for Southeast Asia, the monthly data come from the Global Historical Climatology Network v2 (GHCN2) database.

[8] The authors have generously provided a Python script for computing their Vector Ruggedness Measure (VRM) in ArcView. The script and detailed instructions for installation can be found here: `http://arcscripts.esri.com/details.asp?dbid=15423`.

Matsuura and Wilmott (2012*a*,*b*) interpolated these weather data stations to estimate monthly precipitation and temperature to a 0.5 × 0.5 degree (or 55 km) resolution grid. Then, for the districts in the dataset, we averaged these numbers for the period of 1948-1978 to obtain the predetermined measures of rainfall and temperature.

Task-based Grouping of the Agroclimatic Variables

The agroclimatic similarity index, $\mathcal{A}_j$, is constructed using the full set of agroclimatic variables. In addition, we decompose $\mathcal{A}_j$ into three measures that proxy for skills associated with managing *topography*, *water* and *soil* conditions. This decomposition corresponds to a set of critical tasks for growing rice (De Datta, 1981) and other relevant crops (e.g., Espinoza and Ross, 2015). *Topography* groups together agroclimatic attributes related to the preparation of land, including the slope, ruggedness, and elevation variables; *water* groups together variables related to water management and soil moisture, including rainfall and temperature (which affects the evapotranspiration of water), drainage, and the village's distance-to-river variables; and *soil condition* includes the soil texture, organic carbon content, topsoil pH, sodicity, and the village's distance-to-coast variables (indicating whether the soil is sandy). Each of these similarity variables is constructed in the same manner as $\mathcal{A}_j$.

5. *Constructing Key Variables*

Table A.2 summarizes the data sources for our key outcome variables and regressors. The specific formulas to calculate each of these variables are provided in the main text.

6. *On Sample Sizes and Missing Data*

Our main dataset comprises 814 out of the total 1,021 Transmigration villages that we were able to merge to shapefiles according to the procedure described in Section A.1. The 207 Transmigration villages not in our analysis (i) have missing data in one or more of our many administrative, Census, and geospatial data sources, or (ii) could not be merged with existing spatial identifiers. Importantly, though, we can show that agroclimatic similarity is uncorrelated with the probability of being in our main sample of 814 villages (results available upon request).[9]

It is also possible that some of the villages bordering (or nearby) our main 814 villages were part of the initial Transmigration settlement. We address this concern by rerunning our main regressions for all villages located within d kilometers of the 814 Transmigration villages. Doing so for $d \leq 10$ leaves our main conclusions unchanged (results available upon request).

[9] There are 911 Transmigration villages for which we are able to construct agroclimatic similarity. The remaining 110 villages are largely in Papua where data constraints preclude inclusion.

Table A.2—: Definitions and Data Sources for Key Variables

Variables	**Description**	**Data source***
Log population density	Log of population/total village area (m^2).	2000-PopCen (pop.), GIS (total area)
Economic outcomes		
Sectoral employment choice	Individual choice of sectoral employment	2000-PopCen
Log crop productivity	Log of crop tonnage per hectare in the village	2003-Podes/AgCen (crop yield, farm area)
Revenue-weighted average (cash crop) yield	Average village-level yield of all (cash) crops produced weighted by the share of each crop's revenue in the village	2003-Podes/AgCen (crop yield, farm area), FAO (crop prices)
Nighttime light intensity, 2010	The level of light intensity in the village, 2010; percent of village with light coverage in 2010	NOAA
Similarity indices		
Agroclimatic similarity indices	Agroclimatic similarity between two locations, aggregated to the village-level using population weights.	2000-PopCen (pop. weights), Section A.4 for agroclimatic chars.
Linguistic similarity	Linguistic similarity between ethnoliguistic groups aggregated to the village-level using ethnic population weights	2000-PopCen (pop. weights), *Ethnologue* (linguistic distance)

Notes: *FAO = FAO/PriceSTAT; GIS = GIS Map; NOAA = National Oceanic and Atmospheric Administration's (NOAA) night lights data; PopCen = Population Census; Podes = Village Potential; AgCen = Agricultural Census.

This also has implications for the sample sizes for the individual-level regressions that use the Population Census. According to the Population Census, there are 627,667 Java/Bali-born migrants in our 814 Transmigration villages in 2000, while the MOT Census reports 1,534,264 individuals placed in the original settlements to which we match these contemporary villages. This is why our occupation choice regressions in Table 6 only include 566,956 Java/Bali-born individuals between the ages of 15–65. Looking at villages that lie within 10 km of the borders of our 814 villages, we find an additional 989,249 Java/Bali-born migrants. As argued in the paper, there was not systematic large-scale, ex post migration. However, it is plausible that village boundaries changed over time so that many of the original settlers ended up belonging to villages adjacent to the nucleus of the settlement as defined by our 814 villages.[10] That the main results are robust to including a wider radius around the main settlement villages is consistent with this possibility. For the main analysis, we only use the 814 villages whose village names and locations were high quality matches to the Transmigration site names and locations reported in the 1998 Transmigration census.

[10] We chose to use village boundaries in 2000 because village boundary files are not available for the 1980s.

Appendix B: Additional Results

This section discusses additional results mentioned in the paper including tables and figures.

1. *Further Evidence of Crop Adjustment*

In Table B.4, we provide a second piece of evidence on the crop adjustment mechanism. Adapting an approach developed by Michalopoulos (2012), we identify the extent to which transmigrants bring their preferences for growing rice with them to the Outer Islands. In particular, we focus on the three main staple crops (rice, maize, cassava) and estimate the following equation for Transmigration villages,

$$\frac{rice_j}{staples_j} = \alpha + \rho_1 \left(\frac{rice_{-j}}{staples_{-j}} \right) + \rho_2 \left(\frac{rice_{j(i)}}{staples_{j(i)}} \right) + \mathbf{x}_j' \boldsymbol{\phi} + \nu_j,$$

where $rice_j/staples_j$ is the fraction of rice paddy in total staples planted in 2001; $rice_{-j}/staples_{-j}$ is the corresponding measure in neighboring villages (measured as the average share in the district, excluding Transmigration villages); and $rice_{j(i)}/staples_{j(i)}$ is the corresponding measure for Java/Bali-born migrants' origin districts weighted by the usual π_{ij} term capturing the share of migrants from different origins represented in j. After conditioning on $\mathbf{x}_j$, ρ_1 captures the correlation in cropping patterns across nearby villages subject to the same unobservable ecological constraints, and ρ_2 captures the persistence of migrants' growing preferences beyond these constraints. If $\rho_2 = 0$, then transmigrants fully adapted their cropping patterns to such constraints.

While $\rho_1 > 0$ across all specifications in Table B.4, columns 2 and 4 show that origin region cropping patterns explain about 15-20 percent of the patterns accounted for by spatial autocorrelation across nearby villages. Consistent with Michalopoulos (2012), these results indicate that Java/Bali migrants appear to have preferences for growing (and consuming) rice and replicating the basket of goods grown in their origin regions. While the estimates are not directly comparable, ρ_1 and ρ_2 are larger in our context, with relatively less weight on origin cropping patterns and more weight on destination patterns, suggesting some crop adjustments by individual farmers.

2. *Accounting for Selection in Rice Farming*

We provide further details on our claim in Section IV.B that selection into or out of rice farming does not affect our main results. We deal with village-level selection by running Poisson and Tobit regressions with rice productivity in levels instead of logs—villages that do not produce rice have zero productivity—and find similarly large productivity effects.[1] However, the lower rice productivity in low similarity villages could still be driven by the selection of unobservably higher ability individuals out of rice farming.

We show here why the selection of high ability individuals out of rice farming (and into cash crops) cannot explain the main 20 percent effect of agroclimatic similarity on rice productivity reported in Table 3. The main concern is that this effect is driven by the productivity gap between unobservably high and low ability farmers rather than the gap between farmers with high versus low agroclimatic similarity. In Figure B.3, we show that individual agroclimatic similarity is uncorrelated with years of schooling among transmigrants educated prior to leaving Java/Bali. Here, we calculate and discuss the degree of selection on *unobservable* ability needed to explain our results.

We begin with equation (4) in the paper, but abstract from natural advantages, without loss of generality. Therefore, $y_j = \gamma \mathcal{A}_j + \eta_j^u + \omega_j$. Let the selection margin be represented by I_{ij}, an indicator of whether farmer i chose to farm rice in village j, and I_j be the set of individuals for whom $I_{ij} = 1$. Let η_i^u be an index of farmer $i's$ unobserved productivity (ability). Let η_j^u be mean ability in the village, averaged over the set of individuals who selected into rice farming,$\sum_{i \in I_j} \eta_i^u$.

Consider two villages, one with high agroclimatic similarity (H) and one with agroclimatic similarity that is one standard deviation lower (L). Then, the productivity differential between the two villages depends on γ and the selection effect:

$$E(y_j|H) - E(y_j|L) = \gamma + \underbrace{E(\eta_j^u|H) - E(\eta_j^u|L)}_{\text{selection}}$$

The main concern is that higher ability farmers selected out of rice farming into farming cash crops and more of them did so in low similarity villages (average unobserved ability is lower in L). In the absence of this selection margin, the rice productivity would be similar between high and low similarity villages. Suppose the true $\gamma = 0$, how large would the ability differential have to be to explain the entire productivity differential between the two villages?

We assume the rice productivity for these high ability farmers would have been 4 tons per ha if they did not select out of rice farming. This is the 90th percentile in the distribution for all rice farmers in Transmigration villages, according to the 2004 *Susenas*. We then proceed to calculate what the village-level rice productivity would have been if all farmers were rice farmers.

We first consider the average village, where the rice productivity is 2.5 tons per ha, 65 percent of Java/Bali-born farmers are food crop farmers and 35 percent are cash crop farmers (according to the Population Census). Then, if all farm-

[1] Appendix Table B.6 shows that a one standard deviation increase in agroclimatic similarity increases the likelihood that the village produces any rice by 8.8 percentage points relative to a mean of 74 percent. However, formal Tobit decompositions suggest that the majority of the rice productivity effects in levels are due to an increase in the intensive margin of productivity (i.e., among villages growing any rice).

ers were food crop farmers, the rice productivity for the village would instead be $0.65\times2.5+0.35\times4=2.5+0.35\times1.5=3.025$, assuming that 65 percent of farmers have an average productivity of 2.5 tons per ha and 35 percent of high ability farmers have an average productivity of 4 tons per ha.

Next, we consider a village where $\mathcal{A}_j$ is one standard deviation lower. The rice productivity in this village is 2 tons per ha, if we only consider the productivity of farmers who selected into rice farming (this is 20 percent lower than the average village, using our coefficient estimate in column 1 of Table 3). Let p be the share of high ability farmers who selected out of rice farming. Then, if all farmers were rice farmers, the rice productivity for the village would be $(1-p)\times2+p\times(2+2) = 2+2p$.

To solve for how large p would have to be to explain the entire productivity differential between the two villages, we set $3.025 = 2+2p$, which implies that $p = 0.51$. This means that 16 percentage points (p.p.) more farmers would have to have selected out of rice farming in the low similarity village to explain the entire 0.5 ton per ha productivity differential. This effect size is implausibly large given the observed effects of agroclimatic similarity on crop choice in Table 7.

Instead of comparing Java/Bali-born food versus cash crop farmers only, we also compared food crop farmers to other occupations (for Java/Bali born working age individuals only, and also for all individuals of working age in the village), and the conclusions are similar. In all cases, the selection effect implied by the exercise above is one order of magnitude larger than the estimated effect of $\mathcal{A}_j$ on occupation choice.

3. Additional Robustness Checks

We discuss in detail several robustness tests noted in Section IV.C.

Random Similarity Index. We also compare our agroclimatic similarity index to indices that arise from purely random assignment of individuals from origin i across potential destinations j. We simulate this random matching 10,000 times and compare the resulting indices to our actual index. In doing so, we cannot reject that the means and standard deviations of the random and actual distributions of $\mathcal{A}_j$ across villages are equal.

Baseline Rice Productivity Regressions. Table B.5 reports the results of the robustness checks for our main rice productivity result. Each row introduces a single change to the baseline specification for rice productivity in column 1 of Table 3.

The results are unchanged from the baseline in row 1 when controlling for the log number of total transmigrants placed in the initial year of settlement (row 2) or indicators for the year of settlement (row 3). Row 4 controls for 124 fixed effects (FEs) for province × year of settlement. This very demanding specification cuts the estimate in half, but we still find a statistically and economically meaningful positive effect. This is reassuring given that these FEs account for, among other things, time-varying differences in the quality of local institutional support for the transmigrants in those important first months after arrival.

Next, rows 5 to 8 show robustness to different controls. We partially address aggregation bias in row 5 by controlling for the share of the population from Java/Bali as well as overall log population density. We also find no substantial change when controlling for a third-order polynomial in the latitude and longitude of the village in row 6. This offers a demanding and flexible way to account for spatial heterogeneity in unobservable determinants of productivity.

In row 7, we include several proxies for the nature of local land markets including the share of village land publicly held, the share of private land under certification, the share of land under Islamic trust (*waqf*), the share of households with agricultural landholdings < 0.1 hectares, and the Pareto dispersion parameter λ for landholdings ≥ 0.1 ha (higher λ implies lower inequality). These controls account for ex-post differences in land markets that may have been correlated with agroclimatic similarity. Although these are “bad controls” in that they can be affected our key regressor of interest, their inclusion does not significantly affect the main result.

Additionally, in row 8, we attempt to capture any confounding variation in transmigrants’ experience with different farm size and production scale in their origin districts. In particular, we add two additional controls to the baseline specification. First, we draw upon the 1963 Agricultural Census, which reports district-specific tabulations of the number of farmers with landholding sizes in different bins (see Bazzi, 2015). Then, we estimate the Pareto dispersion parameter λ_i for landholdings ≥ 0.1 ha and take the origin π_{ij}-weighted average of λ_i for each Transmigration village j. Second, we account for the possibility that the inverse-size productivity relationship varies across origin districts as a result of different scale effects. Specifically, we use household-level data from *Susenas* 2004 to estimate origin district-specific elasticities (ζ_i) of rice yield per hectare with respect to area planted and take the π_{ij}-weighted average of ζ_i for each Transmigration village.[2] Adding these controls, we find no change compared to the baseline estimate in row 1.

Rows 9-12 show robustness to different ways of constructing the similarity index. Rows 9 and 10 use different distance metrics and functional forms. Row 11 uses population weights restricted to migrants arriving before 1995, and row 12 restricts to Java/Bali migrants who were at least 30 years old in the year 2000 (and hence eligible to be relocated through the program). These changes address concerns that our baseline π_{ij} terms that include all Java/Bali migrants (calculated using birth locations in the 2000 Census) may not be capturing the original transmigrants.

[2] We only include those ζ_i for which we have at least 30 households in the survey data. This leaves 91 out of 118 origin districts.

Testing for Selection on Unobservables. We follow Altonji, Elder and Taber (2005) in testing for the extent of selection on unobservables required to explain our main findings in Table 3. Altonji, Elder and Taber consider an empirical model with a bivariate normal structure while Bellows and Miguel (2009) develop the same test for a linear model relaxing the joint normality assumption. We implement this approach by dividing the estimate with the most controls (column 5) by the difference between the estimate with island fixed effects and no other controls (column 2) and the estimate with controls. The larger the magnitude of this ratio, the more unlikely that the effect is driven by selection on unobservables. This implementation follows Nunn and Wantchekon (2011), and we find ratios that are similar or larger in magnitude than these three papers. The ratios range from 4.87 to 10.93.

Agroclimatic Similarity and Crop Choice. We can also rule out endogeneity concerns associated with the fact that not all villages produce rice. OLS and Tobit regressions with rice productivity in levels instead of logs—villages that do not produce rice have zero productivity—yield similarly large productivity effects of agroclimatic similarity. Meanwhile, in Appendix Table B.6, we show that a one standard deviation increase in similarity increases the likelihood that the village has any rice production by 8.8 percentage points relative to a mean of 74 percent. However, formal Tobit decompositions (available upon request) suggest that the majority of the rice productivity effects in levels are due to an increase in the intensive margin of productivity (i.e., among villages growing any rice).

Individual-Level Regressions. In Table B.7, we address additional concerns about aggregation bias using auxiliary microdata. By regressing village-level outcomes on key similarity regressors that only apply to a subset of villagers (transmigrants), we risk misinterpreting the relationship between similarity and productivity.[3] We use the best available household-level survey data from a small sample of 74 Transmigration villages. Unfortunately, this dataset does not include migration data, but it does report the ethnicity of the household head. We therefore construct an agroclimatic similarity index using ethnic weights instead of birth locations. Estimating a household-level analogue to our main village-level estimating equation, we find that agroclimatic similarity has a positive effect on farm-level rice yields that is qualitatively and quantitatively very similar to our main estimate of γ. This is supportive evidence that the main productivity effects we identify in the paper are driven by transmigrants rather than natives.

[3] It is also important to note that the estimates of γ are largely unchanged when multiplying our village-level agricultural productivity outcomes by the share of the population working in the given sector as reported in the 2000 Population Census (i.e., multiplying total agricultural productivity by the share working as farmers, rice and food crop productivity by the share working as food crop farmers, and cash crop productivity by the share working in cash crops).

4. Developing the ATE Identification Strategy

This subsection describes the reweighting procedure used in estimating average treatment effects on the treated (ATT) in Table 10 of the paper. First, we predict the probability of being a Transmigration village

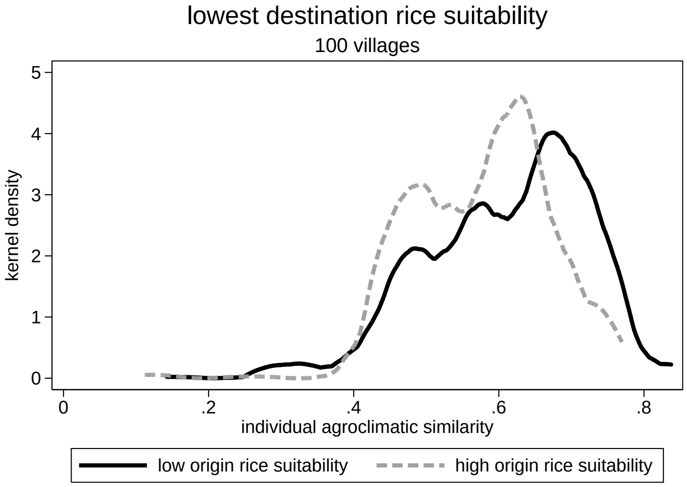
$$\mathbb{P}(T_j = 1) = \Lambda(\mathbf{x}_j' \widehat{\boldsymbol{\zeta}}) \tag{B.1}$$

where $\Lambda(\cdot)$ is a logit function. The estimates $\widehat{\boldsymbol{\zeta}}$ reported in Table B.8 are indicative of sequential site selection among eligible settlement areas. In particular, treated villages (i) are smaller in area, (ii) are at lower altitude, (iii) have more acidic soils, (iv) have better drainage, and (v) are closer to major roads. The covariates explain about one-third of the variation in site selection, and the estimated propensity scores, $\widehat{P}_j$, for treated and control villages exhibit considerable overlap (see Figure B.5).[4] Using the estimates of $\boldsymbol{\zeta}$, we reweight control village j according to its estimated odds of having been a Transmigration site, $\widehat{\kappa} = \widehat{P}_j/(1 - \widehat{P}_j)$, where $\widehat{P}_j$ is the predicted probability from equation (B.1). This reweighting of control villages effectively rebalances the sample as if planners in 1979 randomly chose half of the initial potential settlements. Without the $\widehat{\kappa}$ weights, more than half of the site selection variables exhibit large and statistically significant differences across treated and control villages. With the weights, that share falls to less than 10 percent. By removing observable site selection differentials, we get closer to a causal interpretation of the ATE.

[4] Column 2, which extends the sample to include villages that are within 10 kilometers of treated and almost treated villages, yields similar estimates. However, we explain less of the variation in site selection, and it becomes more difficult to distinguish treated from control villages when including those areas adjacent to but not within the original boundaries of the eligible settlements. This suggests that our model captures much of the same local, small-area variation identified by planners.

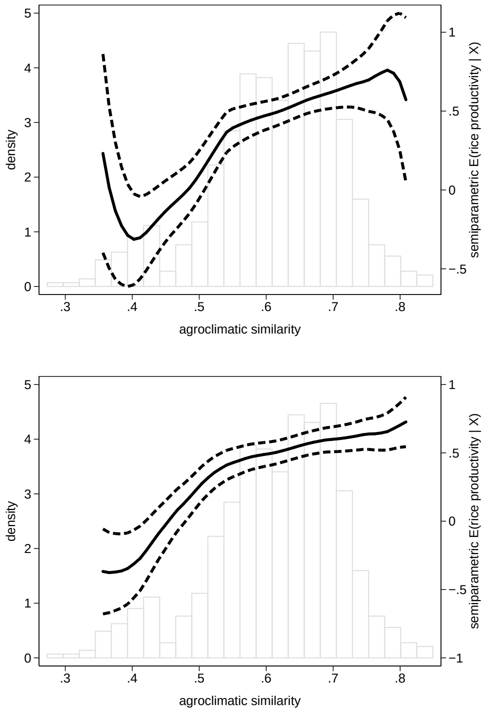
Figures

Figure B.1. : Agroclimatic Similarity and Rice-Specific Natural Advantage



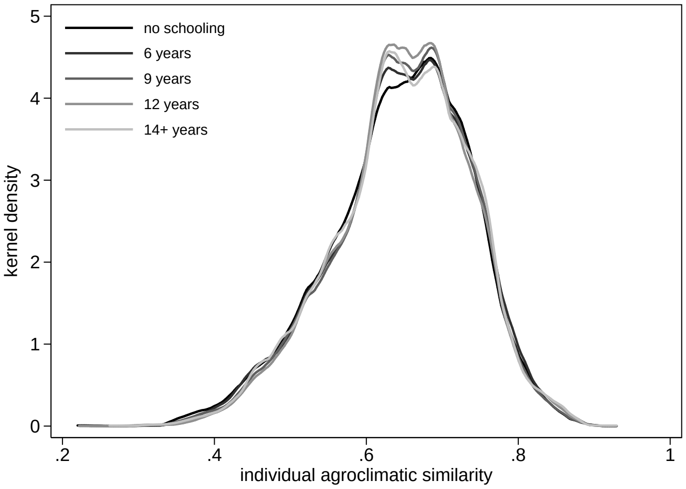
Notes: Individual-level agroclimatic similarity compared across migrants from the 20 out of 119 districts of Java/Bali with the lowest potential rice productivity versus those from the top 20 districts. The four districts of urban Jakarta, which produce no rice in 2001, are excluded from the analysis. The sample is restricted to the 100 Transmigration villages with the lowest potential rice productivity. A formal Kolmogorov-Smirnov test strongly rejects the null of equal distributions (p-value$<$ 0.001). Adopting less stringent tests, we also strongly reject that the means and standard deviations of the two distributions are the same with p-values < 0.001 in both cases. The results are unchanged when restricted to the 66 settlement villages that produce rice among those 100 identified in step 1. All results are robust to alternative rank cutoffs for high and low origins and an alternative cutoff for low natural advantages on the destination side.

Figure B.2. : Alternative Semiparametric Regressions



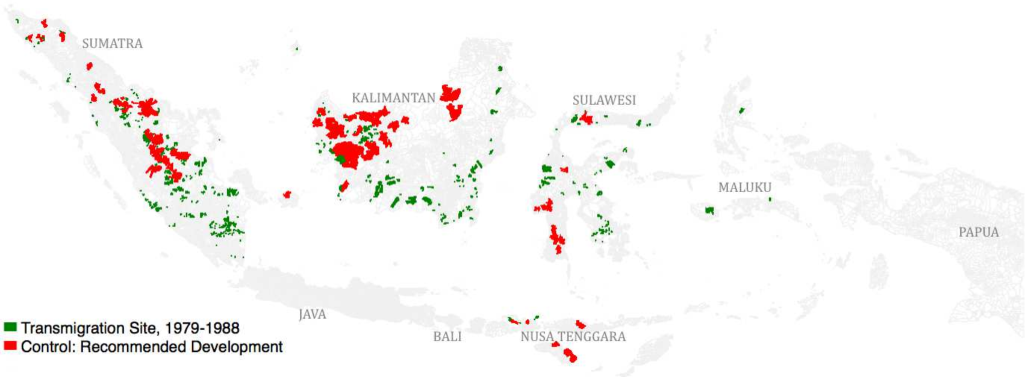
Notes: These graphs consider alternative semiparametric specifications corresponding to our Figure 4. Top Left: local linear regression with bandwidth = 0.05. Top Right: kernel regression with bandwidth = 0.05. Bottom Left: local linear regression with rule-of-thumb bandwidth(≈ 0.075) based on Fan and Gijbels (1996). Bottom Right: kernel regression with rule-of-thumb bandwidth.

Figure B.3. : Individual Agroclimatic Similarity by Schooling



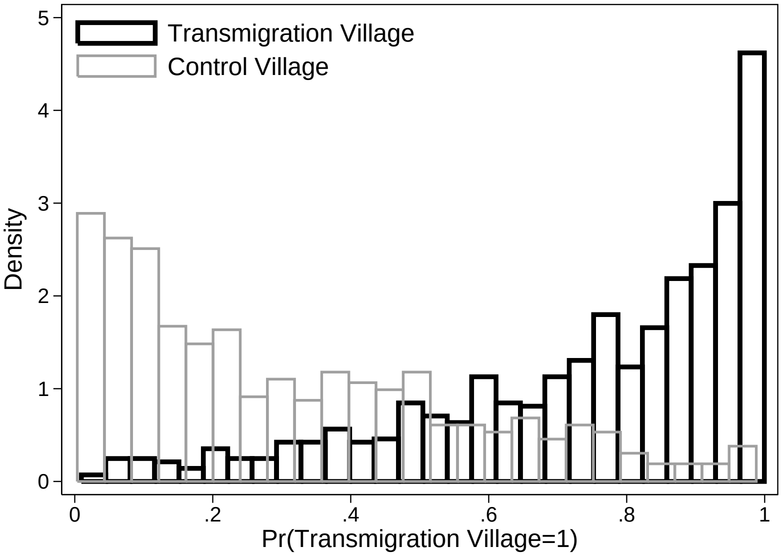
Notes: This figure shows the kernel densities of standardized individual-level agroclimatic similarity, $\mathcal{A}_{ij}$, by level of schooling for all Java/Bali-born individuals living in Transmigration villages and who are between the ages of 15 and 65 and were older than 10 years old in the initial year of settlement. The schooling levels are as reported in the 2000 Population Census. The lack of correlation holds up to the inclusion of additional (Mincerian) individual-level covariates including age, gender, marital status, and district of birth.

Figure B.4. : Map of Transmigration and Control Villages


SUMATRA
KALIMANTAN
SULAWESI
MALUKU
PAPUA
JAVA
BALI
NUSA TENGGARA
🟩 Transmigration Site, 1979-1988
🟥 Control: Recommended Development

Notes: Each colored location on the map corresponds to a Transmigration village or a control/RDA village outside of Papua. The white areas outlined in grey are neither Transmigration nor control villages.

Figure B.5. : Estimated Propensity Scores



Notes: This figure plots the distribution of estimated probabilities of site selection based on the estimates in column 1 of Table B.8.

Tables

Table B.1—: Summary Statistics for Agroclimatic Variables in Java/Bali and the Outer Islands

	Villages in [...]			
	Java/Bali		Outer Islands	
	Mean	Standard Deviation	Mean	Standard Deviation
ruggedness index	0.167	(0.169)	0.273	(0.159)
elevation (meters)	241.0	(316.8)	271.8	(376.9)
% land with slope between 0-2%	0.391	(0.358)	0.268	(0.296)
% land with slope between 2-8%	0.394	(0.270)	0.373	(0.245)
% land with slope between 8-30%	0.170	(0.237)	0.238	(0.238)
organic carbon (%)	0.021	(0.017)	0.033	(0.043)
topsoil sodicity (esp, %)	0.014	(0.003)	0.015	(0.005)
topsoil pH (-log(H+))	6.256	(0.686)	5.446	(0.748)
coarse texture soils (%)	0.045	(0.139)	0.060	(0.160)
medium texture soils (%)	0.528	(0.258)	0.699	(0.227)
poor or very poor drainage soils (%)	0.285	(0.315)	0.275	(0.335)
imperfect drainage soils (%)	0.076	(0.181)	0.135	(0.262)
average annual rainfall (mm), 1948-1978	198.8	(56.1)	205.2	(49.3)
average annual temperature (Celsius), 1948-1978	24.8	(2.8)	25.3	(2.8)
distance to nearest sea coast (km)	27.3	(20.0)	37.2	(39.6)
distance to nearest river (km)	2.5	(5.6)	5.4	(12.0)

Notes: This table reports summary statistics for each of the variables included in our agroclimatic similarity index. The mean and standard deviation for the given variable are computed over all villages in Java/Bali (Outer Islands) in columns 2-3 (4-5). Sample sizes vary slightly across measures, but the full coverage includes 40,518 villages in the Outer Islands and 25,756 in Java/Bali. See Appendix A for details on data sources and construction.

Table B.2—: Top 5 Crops by Potential Revenue in 2002: Transmigration Villages

	Ranking				
Crop	(1)	(2)	(3)	(4)	(5)
rice	0.292	0.217	0.122	0.064	0.036
palm oil	0.261	0.062	0.029	0.019	0.008
rubber	0.131	0.162	0.070	0.023	0.018
cassava	0.069	0.073	0.071	0.043	0.047
cocoa	0.058	0.053	0.019	0.017	0.008
coffee	0.030	0.022	0.021	0.032	0.031
pepper	0.025	0.006	0.014	0.005	0.006
maize	0.021	0.069	0.062	0.062	0.073
banana	0.017	0.049	0.064	0.047	0.021
soybean	0.016	0.014	0.032	0.031	0.025
groundnut	0.016	0.031	0.066	0.055	0.045
coconut	0.012	0.022	0.030	0.053	0.038
chili pepper	0.010	0.017	0.039	0.045	0.035
orange	0.009	0.027	0.027	0.019	0.018
mango, guava, mangosteen	0.009	0.017	0.026	0.027	0.025
cloves	0.006	0.006	0.005	0.012	0.004
cashew	0.004	0.005	0.004	0.003	0.006
sweet potato	0.004	0.005	0.013	0.017	0.016
cinnamon	0.003	0.003	0.004	0.004	
avocado	0.001	0.001		0.001	0.001
garlic	0.001				
cucumber	0.001	0.010	0.021	0.016	0.017
spinach	0.001	0.006	0.008	0.006	0.010
greenbean	0.001	0.012	0.022	0.044	0.042
nutmeg, cardamom	0.001				0.001
papaya		0.001		0.006	0.003
cotton		0.001			
tomato			0.001	0.009	0.001
onion		0.004		0.001	0.004
eggplant			0.008	0.013	0.018
pineapple				0.001	0.003

Notes: This table shows the percentage of Transmigration villages with the given crop as their first, second, third, fourth, or fifth most valuable product as measured by the total output in the 2001 growing season valued at the national price given in FAO PriceSTAT series. The figures are only defined over villages reporting nonmissing and nonzero agricultural output in any crop for which we have price data.

Table B.3—: Heterogeneous Effects of Land Quality on Occupational Choices

Age	All	Young	Old	All	Young	Old
	$Pr(Occupation = \dots)$					
Dependent Variable		Farming			Trading/Services	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Panel A: Baseline with Potential Rice Yield					
individual agroclimatic similarity	0.0090	0.0121	0.0079	-0.0037	-0.0050	-0.0032
	(0.0052)	(0.0057)	(0.0053)	(0.0027)	(0.0028)	(0.0026)
individual linguistic similarity	-0.0168	-0.0182	-0.0162	0.0185	0.0165	0.0193
	(0.0158)	(0.0176)	(0.0153)	(0.0068)	(0.0069)	(0.0067)
rice potential yield	0.0560	0.0536	0.0569	-0.0193	-0.0208	-0.0192
	(0.0273)	(0.0322)	(0.0269)	(0.0125)	(0.0137)	(0.0128)
Number of Individuals	566,956	175,546	391,410	566,956	175,546	391,410
Dependent Variable Mean	0.622	0.489	0.682	0.099	0.089	0.103
	Panel B: Heterogeneous Effects					
individual agroclimatic similarity	0.0550	0.0426	0.0589	-0.0199	-0.0272	-0.0170
	(0.0228)	(0.0242)	(0.0234)	(0.0108)	(0.0127)	(0.0105)
individual linguistic similarity	-0.0260	-0.0263	-0.0261	0.0298	0.0212	0.0339
	(0.0259)	(0.0267)	(0.0264)	(0.0095)	(0.0096)	(0.0097)
rice potential yield	0.0622	0.0581	0.0635	-0.0212	-0.0240	-0.0205
	(0.0261)	(0.0314)	(0.0260)	(0.0129)	(0.0141)	(0.0132)
agroclimatic similarity × potential yield	-0.0301	-0.0201	-0.0333	0.0107	0.0146	0.0090
	(0.0146)	(0.0158)	(0.0150)	(0.0065)	(0.0077)	(0.0064)
linguistic similarity × potential yield	0.0072	0.0061	0.0077	-0.0078	-0.0037	-0.0099
	(0.0126)	(0.0131)	(0.0127)	(0.0047)	(0.0058)	(0.0043)
Number of Individuals	566,956	175,546	391,410	566,956	175,546	391,410
Dependent Variable Mean	0.622	0.489	0.682	0.099	0.089	0.103
Island Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Year of Settlement Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Predetermined Village Controls ($\mathbf{x}_j$)	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table estimates by the OLS the probability that a Java/Bali-born individual living in a Transmigration village as recorded in the 2000 Population Census works in farming (columns 1-3) or trading/services (columns 4-6). Columns 1 and 4 include all Java/Bali-born individuals between the ages of 15 and 65. Columns 2 and 5 restrict to individuals who were less than 10 years old at the time of the initial settlement in their village. Columns 3 and 6 restrict to individuals aged 10 years and greater at the time of the initial resettlement. Both similarity measures are normalized to have mean zero and a standard deviation of one. All regressions include: (i) fixed effects for the year of settlement, (ii) predetermined village-level controls used in previous tables, and (iii) individual-level controls, including age interacted with a male dummy, married dummy, indicators for seven schooling levels, Java/Bali indigenous ethnic group dummy, immigrant from Java/Bali within the last five years, immigrant from another Outer Islands province within the last five years, immigrant from district within the same (Outer Islands) province within the last five years, and indicators for seven religious groups. The measure of potential rice productivity is from the FAO-GAEZ. Results are similar omitting the individual-level controls. Standard errors are clustered at the district level.

Table B.4—: Against the Grain: Neighborhood vs. Origin Effects in Rice Land Allocation

Dependent Variable	*Rice/Staples*		$Pr(Rice/Staples > 0.5)$	
	(1)	**(2)**	**(3)**	**(4)**
share of rice Ha in main staple Ha, neighbors	0.157	0.158	0.164	0.166
	(0.023)	(0.023)	(0.025)	(0.025)
share of rice Ha in main staple Ha, Java/Bali origin		0.021		0.036
		(0.008)		(0.012)
Number of villages	694	694	694	694
Dep. Var. Mean	0.684	0.684	0.707	0.707

Notes: The dependent variable is farmland area planted with rice as a fraction of area planted with the three main staples of rice, maize, and cassava. In columns 3-4, the share is transformed into a binary outcome equal to one if the share of rice is greater than 50%. The "share of rice hectares (Ha) in main staple Ha, neighbors" is the average share across all villages in the given district excluding Transmigration villages. The "share of rice hectares (Ha) in main staple Ha, Java/Bali origin" is a weighted average of the shares prevailing in the origin districts of Java/Bali with the weights being the share of Java/Bali-born immigrants in the given village from the given origin district. Both variables have been normalized to have mean zero and standard deviation one. All regressions include the usual predetermined village-level control variables and island fixed effects. Standard errors in parentheses allow for unrestricted spatial correlation between all villages within 150 kilometers of each other (Conley, 1999).

Table B.5—: Robustness: Rice Productivity

	Agroclimatic Similarity
1. Baseline Specification	0.204 (0.064)
2. Total Transmigrants Placed in Initial Year	0.205 (0.064)
3. Year of Settlement Fixed Effects	0.200 (0.063)
4. Province × Year of Settlement Fixed Effects	0.114 (0.065)
5. Controlling for Java/Bali-born Pop. Share and Overall Pop. Density	0.211 (0.063)
6. 3rd Degree Polynomial in Latitude/Longitude	0.193 (0.077)
7. Controlling for Land Markets and Distribution	0.255 (0.066)
8. Controlling for Origin Land Distribution and Size-Prod. Elasticity	0.205 (0.063)
9. Alternative Normalization of Agroclimatic Similarity Index	0.192 (0.060)
10. Euclidean Distance in Agroclimatic Similarity Index	0.161 (0.086)
11. Only pre-1995 Java/Bali Immigrants in Agroclimatic Similarity Index	0.206 (0.067)
12. Only Java/Bali-born age >30 in Agroclimatic Similarity Index	0.212 (0.060)

Notes: Each row corresponds to a separate regression of log rice productivity on agroclimatic similarity, predetermined village-level control variables, and island fixed effects unless noted otherwise. Agroclimatic similarity is normalized to have mean zero and a standard deviation of one. Row 1 is our baseline specification from column 1 of Table 3 in the paper. In each subsequent row, we make the single change in specification noted in the row description. Row 9 normalizes the difference between characteristic g in origin i and destination j based on the difference observed in all other ij pairs whereas our baseline index normalizes each characteristic before taking the difference. Otherwise, all other aspects of the estimating equation are as in the baseline. Standard errors in parentheses allow for unrestricted spatial correlation between all villages within 150 kilometers of each other (Conley, 1999).

Table B.6—: Agroclimatic Similarity and Crop Choice

Dependent Variable	Agroclimatic Similarity	Mean of Dependent Variable
	Panel A: Food Crops	
any food crop production	0.074	0.901
	(0.022)	
any rice production	0.088	0.737
	(0.019)	
any cassava production	0.037	0.442
	(0.037)	
any maize production	0.037	0.441
	(0.038)	
	Panel B: Cash Crops	
any cash crop production	-0.030	0.903
	(0.019)	
any rubber production	0.010	0.410
	(0.031)	
any palm oil production	0.018	0.382
	(0.022)	
any coffee production	0.055	0.184
	(0.017)	
any cocoa production	0.020	0.154
	(0.009)	

Notes: Each row corresponds to a separate regression for the given dependent variable using the same baseline specification applied in column 1 Table 3 in the paper. All dependent variables are as observed in the 2001-2 growing season. For each crop we report the extensive margin relationship between agroclimatic similarity and the (linear) probability that the crop is grown in the village. *any food* and *any cash* equal one if the village grows any of a large list of crops of the given type (see the notes to Table 8 in the paper). Agroclimatic similarity is normalized to have mean zero and a standard deviation of one. All regressions include predetermined village-level control variables and island fixed effects. Standard errors in parentheses allow for unrestricted spatial correlation between all villages within 150 kilometers of each other (Conley, 1999).

Table B.7—: Individual-Level Rice Productivity Regression

	All	Ethnicity	
		Java/Bali	Outer Islands
	(1)	(2)	(3)
agroclimatic similarity	0.069 (0.053)	0.169 (0.077)	0.133 (0.193)
Number of Households	546	449	97

Notes: Individual-level regressions of log rice output per hectare for individuals (household heads) living in a random sample of 74 Transmigration villages in a nationally representative household survey (*Susenas*) conducted in 2004. Although we do not observe migration histories (i.e., district of birth), we can use the observable individual (household head) ethnicity to construct a proxy for the individual's agroclimatic similarity based on the prevailing origins of the given ethnic population in the village. Agroclimatic similarity is defined at the individual-level based on an origin-weighted average of the ethnicity-specific agroclimatic similarity prevailing across individuals in the village as observed using the full 2000 Population Census. All regressions include predetermined village-level controls.

Table B.8—: Determinants of Site Selection

	Treated/Control Radius	
	0 km (1)	10 km (2)
log village area, Ha	-0.103	-0.028
	(0.019)	(0.014)
% w/ slope between 0-2%	0.006	0.002
	(0.002)	(0.001)
Vector Ruggedness Measure	-0.164	-0.027
	(0.115)	(0.076)
log altitude, m^2	-0.026	-0.018
	(0.009)	(0.008)
Organic Carbon (%)	-0.020	-0.010
	(0.006)	(0.007)
Topsoil Sodicity (ESP) %	0.086	0.006
	(0.093)	(0.065)
Topsoil pH (-log(H+))	-0.141	-0.155
	(0.051)	(0.041)
Coarse texture soils (%)	-0.033	0.108
	(0.226)	(0.214)
Very poor or poor drainage (%)	0.073	-0.032
	(0.085)	(0.081)
Imperfect drainage soils (%)	-0.231	-0.132
	(0.138)	(0.100)
Avg. rainfall, 1948-1978	-0.001	-0.001
	(0.001)	(0.001)
Avg. temp (Celcius), 1948-1978	-0.022	0.002
	(0.014)	(0.012)
Distance to Nearest Major Road	-0.300	-0.255
	(0.157)	(0.165)
Distance to Nearest Coast	-0.060	-0.065
	(0.038)	(0.029)
Distance to Nearest River	-0.011	-0.023
	(0.022)	(0.013)
Distance to District Capital	0.025	0.014
	(0.028)	(0.017)
N	1470	5032
Pseudo R^2	0.366	0.284
Log Likelihood	-641.9	-2109.1
LR χ^2	365.1	143.9

Notes: This table reports average marginal effects. The dependent variable is a binary indicator equal to one if the village is located within 0 or 10 kilometers of a Transmigration site. Standard errors clustered by district in parentheses.

Appendix C: Reassigning Transmigrants to Maximize Aggregate Production

The problem of finding the optimal allocation of transmigrants to destination villages is an example of the generalized assignment problem (GAP), a problem in computer science and operations research. This problem has been shown to be NP-Hard (Fischer, Jaikumar and van Wassenhove, 1986).[1] Intuitively, solutions to the problem require a search over the space of partitions of the set of transmigrant individuals, which we can represent by $\{1, 2, ..., N\}$. A partition would divide this set into subsets consisting of individuals assigned to each destination location. Without capacity constraints, a brute force algorithm for assigning people to places would have to check all possible different partitions for optimality. It has been shown that the number of partitions of a collection of N objects is given by B_N, the Bell number for a set of size N. The number of partitions of a set of N increases faster than e^N; the first few Bell numbers are $B_1 = 1$, $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, $B_6 = 203$, $B_7 = 877$, $B_8 = 4140$, $B_9 = 21147$, and $B_{10} = 115975$.

However, we can obtain a fast approximation to the optimal solution using a "greedy" assignment algorithm. The algorithm first orders villages by the total number of spaces for Transmigrants, from smallest to largest. This gives us an ordering of $v \in \{1, 2, ..., N\}$.[2] Next, we maximize similarity in village 1 by choosing people from origin locations with the smallest agricultural distance to village 1. Using the remaining unassigned people, we repeat the procedure to maximize similarity in subsequent villages. Our computationally-efficient program provides us with a new set of agroclimatic similarity indices, $\widetilde{\mathcal{A}}_1$, $\widetilde{\mathcal{A}}_2$, ..., $\widetilde{\mathcal{A}}_N$.

Maximizing similarity across villages is of first order importance to rice production because, for this exercise, we assume that log rice output per hectare in village v is given by:

$$\log y_v = x_v'\beta + \theta\mathcal{A}_v$$

where β' and θ are known parameters, obtained from Table 3, Column 1. Fixing x_v, we can predict the new rice output per hectare in village v as follows:

$$\log \widetilde{y}_v - \log y_v = \theta\left(\widetilde{\mathcal{A}}_v - \mathcal{A}_v\right)$$

where $\widetilde{\mathcal{A}}_v$ is village v's optimized similarity index. We construct the new output

[1] GAP has many applications, including vehicle routing, assignment of software development tasks to programmers, assigning jobs to computers in computer networks, and designing communication networks

[2] Note that our greedy assignment algorithm starts by maximizing similarity in the smallest village. We tried randomizing the order of villages in which similarity was maximized, and found that we could do no better than this ordering choice. However, total aggregate output always improved; in 100 different randomly assigned orderings, the range of output improvements was 16.3 percent to 27.4 percent.

per hectare by rearranging:

$$\widetilde{y}_v = \exp\left\{\widehat{\theta}\left(\widetilde{\mathcal{A}}_v - \mathcal{A}_v\right)\right\} y_v$$

Multiplying this result by H_v, village v's total hectares of cultivation, gives us the new tons of rice per hectare that would have been produced.

We can compare the total rice production across all villages to our counterfactual rice production if villages had been assigned transmigrants in a way that maximized similarity. This gives us:

$$\Delta = \sum_{v=1}^{V} \widetilde{y}_v H_v - \sum_{v=1}^{V} y_v H_v$$

From this calculation, we found that total aggregate rice production would have been 27 percent higher with our more optimal individual assignment.

REFERENCES

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–184.

Bazzi, Samuel. 2015. "Wealth Heterogeneity and the Income Elasticity of Migration." *Unpublished manuscript.*

Bellows, John, and Edward Miguel. 2009. "War and Local Collective Action in Sierra Leone." *Journal of Public Economics*, 93(11): 1144–1157.

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

Costinot, Arnaud, Dave Donaldson, and Cory Smith. 2016. "Evolving Comparative Advantage and the Impact of Climate Change in Agricultural Markets: Evidence from 1.7 Million Fields around the World." *Journal of Political Economy*, 124(1): 205–248.

De Datta, Surajit K. 1981. *Principles and Practices of Rice Production.* Hoboken, NJ:John Wiley & Sons.

Espinoza, Leo, and Jeremy Ross, ed. 2015. *Corn Production Handbook.* Little Rock, Arkansas:University of Arkansas Cooperative Extension Service.

Fan, Jianqing, and Irene Gijbels. 1996. *Local Polynomial Modelling and Its Applications.* Boca Raton, FL:CRC Press.

Fearnside, Philip M. 1997. "Transmigration in Indonesia: Lessons from Its Environmental and Social Impacts." *Environmental Management*, 21(4): 553–570.

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

Fischer, Marshall L., R. Jaikumar, and Luk N. van Wassenhove. 1986. "A Multiplier Adjustment Method for the Generalized Assignment Problem." *Journal of Management Science*, 32(9): 1095–1103.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review*, 102(2): 994–1028.

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

Lewis, M. Paul, ed. 2009. *Ethnologue: Languages of the World, Sixteenth edition.* Dallas, Texas:SIL International.

Matsuura, Kenji, and Cort J. Wilmott. 2012*a*. "Terrestrial Air Temperature: 1900-2010 Gridded Monthly Time Series (V 3.01)."

Matsuura, Kenji, and Cort J. Wilmott. 2012*b*. "Terrestrial Precipitation: 1900-2010 Gridded Monthly Time Series (V 3.02)."

Michalopoulos, Stelios. 2012. "The Origins of Ethnolinguistic Diversity." *American Economic Review*, 102: 1508–1539.

Nunn, Nathan, and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review*, 101(7): 3212–3252.

Nunn, Nathan, and Nancy Qian. 2011. "The Potato's Contribution to Population and Urbanization: Evidence From A Historical Experiment." *Quarterly Journal of Economics*, 126(2): 593–650.

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

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