Selection and Heterogeneity in the Returns to Migration*

Eduardo Cenci Marieke Kleemans† Emilia Tjernström‡

This version: November 15, 2024
Please click here for the latest version of the paper

Abstract

There is considerable debate on the returns to rural-urban migration in developing countries, and magnitudes differ depending on the empirical methods used. We aim to reconcile these divergent estimates by explicitly accounting for the role of heterogeneity in the returns to migration. We develop a correlated random coefficient model that allows for location-specific skills and heterogeneous returns, estimated using rich longitudinal data from Indonesia, China, and Tanzania. This model lets us extrapolate the returns identified from switcher sub-populations to non-switchers—a group of particular interest to policymakers deciding whether to encourage migration as a development strategy. Our results reveal considerable heterogeneity in the returns to migration and show a clear pattern in the relationship between absolute and comparative advantage across countries: those with the lowest productivity in rural areas stand the most to gain from migrating. This suggests that migration is a pro-poor strategy but that barriers to migration may prevent workers from realizing their potential. As such, individuals appear to be inefficiently sorted across space; therefore, encouraging migration could lead to large returns.

JEL Classification: J24, J61, C14, O15
Keywords: migration, labor mobility, panel data, heterogeneity, generalized method of moments

*Corresponding author: Emilia Tjernström (emilia.tjernstrom@monash.edu). We are grateful to David Buller for helping prepare the data. This project greatly benefited from conversations with David Albouy, Alex Bartik, Dan Bernhardt, Joshua Deutschmann, Paul Castañeda Dower, Andy Garin, Dalia Ghanem, Douglas Gollin, James Heckman, Robert Jensen, Brian Kovak, Thomas Lemieux, Nicholas Y. Li, Ana Paula Melo, Edward Miguel, Mushfiq Mobarak, Suresh Naidu, Charly Porcher, Vinicios Sant'Anna, Laura Schechter, Jeff Smith, Sergio Urzua, and seminar and conference participants at the Econometric Society Africa Conference, Midwest International Economic Development Conference, CDES Sustainable Development Conference, University of Colorado Denver, University of Illinois Urbana-Champaign, University of Nebraska at Omaha, University of Pittsburgh, and University of Wisconsin–Madison. All errors are our own.

†University of Illinois, Urbana-Champaign
‡Monash University

1 Introduction

As economies develop, labor typically migrates out of rural areas into higher-productivity sectors in cities. Yet, a puzzling phenomenon persists across the developing world: striking income and consumption gaps between rural and urban areas persist, even after decades of rapid urbanization. Urban residents consistently earn two to three times more than their rural counterparts, a gap that remains even after accounting for cost-of-living differences, educational attainment, and other observable characteristics (Young, 2013; Gollin, Lagakos and Waugh, 2014; Herrendorf and Schoellman, 2018).

With countries undergoing structural transformation and rapid urbanization, why do we still observe such striking income and consumption disparities? This puzzle lies at the heart of debates on labor allocation, productivity, and welfare in developing economies. One view suggests that labor is misallocated across space due to frictions or policies, implying that reducing these barriers could increase productivity and welfare. Another perspective posits that workers efficiently sort themselves based on unobserved characteristics, such as location-specific skills, which would make these wage gaps less amenable to policy intervention. A key input into this debate—and essential for informed policy decisions—is reliable estimates of the returns to migration. In other words, good policy decisions require robust estimates of the causal effect of migration on individual consumption and earnings.

In this paper, we estimate the consumption returns to rural-urban migration for both migrant and non-migrant populations. Different empirical methods rely on distinct subpopulations for identification, which can lead to divergent results in the presence of heterogeneous returns. By explicitly accounting for the heterogeneity in returns across different subpopulations, we provide a framework that helps reconcile the wide range of estimates of rural-urban migration returns found in the literature. Our results also contribute estimates of the returns to migration for non-migrants, a large and important sub-population whose returns are unidentified in standard models.

To study the returns to migration, we leverage rich longitudinal data from Indonesia, China, and Tanzania, totaling over 75 thousand individuals across the three countries, for whom we observe location choices and labor market outcomes for three to five periods spanning several years. We use these rich data to estimate the observational returns to rural-urban migration, meaning the returns for those observed to migrate in the data. We then turn to non-parametric approaches to examine the extent to which migration returns vary across migration histories, which sheds light on the plausibility of the assumptions that underlie standard panel data estimators. Further, we develop a

model that acknowledges that workers have location-specific skills, which are rewarded differently in rural and urban labor markets. Building on the Roy (1951) model and its extensions, our model accounts for unobserved heterogeneity and comparative advantage in the returns to migration.

Our methodological innovation is to cast this model as a group random coefficient model that imposes a specific restriction on the form of the heterogeneous returns, inspired by Suri (2011) and Tjernström, Ghanem, Barriga-Cabanillas, Lybbert, Michuda and Michler (2024). By restricting the returns to urban migration to be linear in comparative advantage, we are able to extrapolate returns to non-migrants—an important sub-population whose returns are typically unidentified in standard models.

Our analysis yields four main results. First, pooled OLS regressions reveal large average rural-urban consumption gaps, ranging from 40 log points in Indonesia to 74 log points in Tanzania, consistent with previous findings in the literature. Second, these gaps decrease to 20-57 log points when we include controls and further narrow to 6-15 log points with the addition of individual fixed effects. While individual fixed effects controls for time-invariant individual characteristics, this approach also effectively restricts identification to individuals who migrate between urban and rural areas. For comparison, we therefore re-estimate our OLS results using a sample of only switchers and obtain estimates similar to those from the fixed-effects models (3-12 log points with controls). These patterns suggest that selection into migration may play a larger role than time-invariant characteristics in explaining the estimated consumption gaps. The importance of selection motivates us to further investigate the returns to migration across different switcher sub-populations, based on migration histories.

Third, we examine the observational returns to urban location across different migration histories and document striking heterogeneity. This suggests that standard panel data estimators, like fixed-effects models, may fail to capture important features of the data. Fixed-effects models assume that the returns to urban migration are homogeneous regardless of a person's migration trajectory. As a result, these estimators overlook the important variation in consumption outcomes associated with different migration histories, potentially resulting in flawed estimates of the key policy-relevant parameters.

Fourth, to address this limitation, we turn to our group random coefficient model, which allows for heterogeneous returns across different migration histories. Our analysis reveals a consistent negative relationship between comparative and absolute advantage. When we extrapolate the estimates to non-migrants, we find significantly greater potential returns for individuals who have never migrated—particularly those remaining

in rural areas. This pattern is remarkably consistent across the three countries that we study, indicating a pattern of labor misallocation that may signal the existence of untapped opportunities for enhancing welfare through policy interventions that reduce migration barriers. One way to interpret our findings is that migration acts as a "pro-poor" mechanism: rural individuals with the lowest baseline consumption experience the most significant gains from moving to urban centers.

Our work contributes to two strands of the literature. First, we add to the evidence on the magnitudes and sources of sectoral labor productivity gaps in developing countries.[1] A central debate in this literature is whether earnings or consumption gaps imply that productivity and welfare would increase if workers were able to reallocate to more productive sectors, or whether they simply reflect self-selection of heterogeneous workers among sectors. Our findings provide a way to reconcile diverging estimates of returns to rural-urban migration and occupational mobility from agriculture to non-agriculture sectors.

Several recent studies using microdata to estimate the returns to rural-urban migration (Alvarez, 2020; Hamory, Kleemans, Li and Miguel, 2021; Herrendorf and Schoellman, 2018) find that the average rural-urban migration returns for many migrant populations are relatively small—especially when compared to earlier results that used national accounts data or cross-sectional analyses (see e.g., Gollin *et al.* 2014; Young 2013). One implication of these results obtained using microdata is that efficient sorting of workers based on comparative advantage can explain rural-urban earnings gaps, which suggests limited opportunities for welfare-enhancing policy interventions. While our panel data results are similar to these studies, we show that the average returns estimated using fixed effects hide substantial heterogeneity.

Another set of papers explores the role of misallocation and barriers to sectoral mobility and uncover evidence consistent with the notion that rural non-migrants face external constraints or costs that keep them from relocating to urban labor markets. Adamopoulos, Brandt, Leight and Restuccia (2022) use data from China to examine the interaction between misallocation and selection, finding that distortionary policies disproportionately affect more productive farmers, making them less likely to work in agriculture, which in turn leads to lower agricultural productivity overall. Gai, Guo, Li, Shi and Zhu (2024) focus in on migration costs as the chief barrier, also in China, finding that reduced migration barriers would increase productivity and GDP alike. Pulido

[1] See Donovan and Schoellman (2023) for a recent review of the role of labor market frictions in the agricultural productivity gap and Lagakos (2020) for a review of the evidence for urban-rural migration.

and Święcki (2021) explicitly incorporate barriers to mobility into a selection model and find that they act as significant productivity constraints on both sectors.[2] Similarly, the estimates in Bryan, Chowdhury and Mobarak (2014), based on experimentally-induced seasonal migration, suggest that urban migration increases consumption by 33 percent.

A key innovation in our paper is our ability to extrapolate the returns to migration to non-migrant subpopulations. When we impose additional assumptions on the form of heterogeneity, we show that returns to non-migrants are consistent with the evidence that finds evidence of important barriers to mobility. There are some similarities between our approach and that in Alvarez-Cuadrado, Amodio and Poschke (2023), who study the alignment of absolute and comparative advantage in African agriculture. Leveraging households who are involved in both the agricultural and non-agricultural sectors, they find that absolute and comparative advantage are negatively correlated in African agriculture, suggesting that self-selection may not be the key driver of agricultural productivity gaps between sub-Saharan African countries and higher-income economies. Like Alvarez-Cuadrado *et al.* (2023), we are interested in the relationship between absolute and comparative advantage, but our methodological approach differs from theirs in several key ways: first, rather than focusing on the sign of the relationship between absolute and comparative advantage, we use this relationship to extrapolate the returns to non-switcher subpopulations. Second, given our focus on rural-urban migration, we rely on sectoral switches rather than infra-marginal decisions for identification. Third, we track individuals' choices over time, as opposed to households.[3]

The second key literature that we contribute to is the vast literature on micro-economic policy evaluation, examining models with heterogeneous effects and endogenous regressors. We build on the generalized Roy models previously used by research in labor and development economics and adapt them to our context, the study of heterogeneity in the returns to rural-urban migration.[4] Our empirical strategy builds

[2]Relatedly, a counterfactual simulation in Bryan and Morten (2019), based on a structural model that incorporates sorting and agglomeration effects, suggests that removing barriers to mobility would increase productivity by 22 percent. They further argue that this is likely a lower bound on the gains from removing migration barriers, as their main analysis excludes self-employed individuals, who likely have an even greater variance in earnings.

[3]An extensive literature has shown that household decisions are often the result of complex negotiations or bargaining within the household, implying that household decisions may reflect this bargaining process rather than comparative advantage. As examples, see (Vermeulen, 2002) for a survey of household collective models and (Chen, 2013) for an application of such models in the context of migration in a developing country. Additionally, the aggregate nature of household-level data may mask individual heterogeneity if household members have a comparative advantage in different sectors.

[4]Lemieux (1998) develops a panel data estimator that accounts for two-sided non-random selection and differential skill rewards across sectors, and applies it to the study of unions' effects on wages. Suri (2011) uses a similar model to estimate the returns to hybrid seed adoption for different farmer

on recent work in the econometrics literature (Verdier, 2020; Tjernström *et al.*, 2024) aiming to flexibly estimate the returns to switcher subpopulations and extrapolating these returns to non-switchers.

Returns for groups like non-migrants in our study represent a potential parameter of interest for policymakers and a technical challenge for researchers. Heckman and Vytlacil (1999) propose a way to estimate returns for inframarginal groups by leveraging marginal treatment effects (MTE).[5] Estimates for these populations may differ sharply from those obtained based on marginal individuals, like switcher populations. The approaches used in the MTE literature approaches require exogenous policy variation or another type of instrument, while our approach instead relies on imposing structure on the form of comparative advantage.[6]

2 Model and Identification

Consider an economy where individuals decide whether to work in the rural or urban labor market, based on their skills and preferences. As is common in the literature, we motivate our empirical strategy with a generalized Roy (1951) model.[7] The classical Roy model assumes that people make choices entirely based on income or consumption, which is an important limitation in many economic applications. Our generalized Roy model instead allows migration decisions to also depend on a non-pecuniary component of utility. In the context of migration, we can think of the non-pecuniary component of utility as capturing the proximity to family and differences in local amenities.

2.1 Location Choice Model

The two labor markets, indexed by $l = R, U$ for rural and urban, demand distinct skills from its workers and compensate these skills differently. We can express the expected value of log consumption in each sector as

$$y_{it}^U = \beta_t^U + x_{it}'\gamma^U + \theta_i^U \tag{1}$$

$$y_{it}^R = \beta_t^R + x_{it}'\gamma^R + \theta_i^R \tag{2}$$

subpopulations.

[5] See also Heckman and Vytlacil (2005), Heckman and Urzua (2010), and Cornelissen, Dustmann, Raute and Schönberg (2016).

[6] See Gai *et al.* (2024) for a recent application that uses policy variation in the context of rural-urban migration in China.

[7] See for example Bazzi, Gaduh, Rothenberg and Wong 2016; Alvarez 2020; Lagakos and Waugh 2013; Pulido and Święcki 2021.

where x_{it} is a vector of observable characteristics, β_t^l is the average log consumption in sector l, and θ_i^l is a time-invariant unobservable characteristic that captures worker i's productivity in sector l. The difference between these two sector-specific unobservables, $\theta_i^U - \theta_i^R$, represents person i's comparative advantage.

Following Lemieux (1998) and Suri (2011), we can project the time-invariant, sector-specific unobservables onto this difference, $\theta_i^U - \theta_i^R$, which allows us to simplify the model by focusing on the worker's comparative advantage, and yields the following expressions:

$$\theta_i^U = b_U(\theta_i^U - \theta_i^R) + \tau_i \tag{3}$$

$$\theta_i^R = b_R(\theta_i^U - \theta_i^R) + \tau_i \tag{4}$$

We can think of the projection coefficients b_U and b_R as measuring how much of person i's productivity in sector l is explained by their comparative advantage, $\theta_i^U - \theta_i^R$. They are defined as follows, with σ_l^2 denoting the variance of the unobserved productivity in sector l and σ_{UR} representing the covariance between an individual's rural and urban productivity, $\text{Cov}(\theta_i^U, \theta_i^R)$:

$$b_U = \frac{(\sigma_U^2 - \sigma_{UR})}{(\sigma_U^2 + \sigma_R^2 - 2\sigma_{UR})} \tag{5}$$

$$b_R = \frac{(\sigma_{UR} - \sigma_R^2)}{(\sigma_U^2 + \sigma_R^2 - 2\sigma_{UR})} \tag{6}$$

Given these projections, the unobservable absolute advantage, τ_i, affects individuals' outcomes regardless of their location choices, and is by construction orthogonal to their comparative advantage, $\theta_i^U - \theta_i^R$.

We further introduce two additional parameters, θ_i and ϕ. The parameter $\theta_i = b_R\left(\theta_i^U - \theta_i^R\right)$ is simply a rescaled function of comparative advantage that captures both comparative advantage and how much the rural sector values it. The parameter $\phi \equiv \frac{(b_U - b_R)}{b_R}$ expresses how much more important comparative advantage is for urban productivity relative to rural productivity.

We can now rewrite the sector-specific unobservables as the following functions:

$$\theta_i^U = (1+\phi)\theta_i + \tau_i \tag{3'}$$

$$\theta_i^R = \theta_i + \tau_i \tag{4'}$$

Plugging equations (3′) and (4′) into our expressions for sector-specific consumption,

(1) and (2), we get the following expressions for urban and rural consumption:

$$y_{it}^U = \beta_t^U + x_{it}'\gamma^U + (1+\phi)\theta_i + \tau_i \tag{7}$$

$$y_{it}^R = \beta_t^R + x_{it}'\gamma^R + \theta_i + \tau_i \tag{8}$$

We allow for factors other than consumption to influence migration decisions by modeling location choice as a function of utility. In addition to consumption, utility depends on non-monetary aspects like proximity to family and local amenities, which we model as a time- and location-specific idiosyncratic shock, ν_{it}^l. We can therefore express person i's utility in market l at time t by $V_{it}^l = y_{it} + \nu_{it}^l$. This utility shock can lead workers to change labor market even in the absence of systematic changes to the observable factors that determine consumption.

We assume that the utility shocks, ν_{it}^l, are independently and identically distributed (i.i.d.) across individuals, time periods, and locations. The independence of ν_{it}^l is important for several reasons. First, it ensures that these shocks can induce movements across sectors without introducing persistence or correlation with past decisions or outcomes. This feature allows for period-to-period variations in location choices that are not driven by changes in observable characteristics or sector-specific productivity. Second, the independence assumption ensures that ν_{it}^l is orthogonal to the shocks in the outcome equation. This orthogonality is essential for identification. By introducing these independent shocks, we capture the idea that individuals may move between rural and urban areas due to transitory, idiosyncratic factors that are unrelated to their productive capabilities or to systematic differences between the sectors.

Workers make their location choices for the next period by comparing the expected utility they would obtain. At the end of each period, worker i observes her consumption and utility shocks and forms expectations about future shocks. She will decide to work in the urban market in period t if her expected utility of doing so exceeds that of working in the rural market. Like Lemieux (1998), we assume that the average rural-urban consumption gap, $\beta = \beta_t^U - \beta_t^R$ is constant over time. Abstracting away from covariates, we then obtain that workers choose to work in the urban market if

$$E\left(V_{it}^U\right) > E\left(V_{it}^R\right) \Leftrightarrow \beta + \phi\theta_i + E\left(\nu_{it}^U - \nu_{it}^R\right) > 0. \tag{9}$$

2.2 Empirical Model

Let the indicator variable $D_{it} = 1$ denote urban location. Combining Equations 7 and 8 and re-arranging terms, we can write the following generalized consumption equation:

$$y_{it} = \beta^R + \theta_i + \tau_i + (\beta + \phi\theta_i)D_{it} + x'_{it}\gamma^R + D_{it}x'_{it}(\gamma^U - \gamma^R) + \varepsilon_{it} \tag{10}$$

Simplifying further, we assume that covariates influence consumption in the same way across sectors for both rural and urban sectors, i.e. we let $\gamma^U = \gamma^R$ so that equation (10) simplifies to:

$$y_{it} = \beta^R + \theta_i + \tau_i + (\beta + \phi\theta_i)D_{it} + x'_{it}\gamma + \varepsilon_{it}. \tag{10'}$$

The LCA restriction can be written as follows:[8]

$$\Delta_i = \beta + \phi\theta_i \tag{11}$$

where we refer to the parameter ϕ as the LCA parameter, as it indicates the slope of the linear relationship between a person's returns to migration, Δ_i and their (scaled) comparative advantage, θ_i. Intuitively, the parameter ϕ expresses how much more important comparative advantage is for urban productivity relative to rural productivity.

2.2.1 Unrestricted Group Random Coefficient Model

As a first step towards our preferred model, we develop a group random coefficient (GRC) model (Tjernström *et al.*, 2024). This reduced-form model allows us to identify the average returns to migration for subpopulations that we observe changing locations in our data, i.e. the switcher subpopulations. Since our choice variable, D_{it}, is binary and we have a finite number of time periods, our data contain a finite number of migration histories, or trajectories, $\underline{d} \equiv (\underline{d}_1, \ldots, \underline{d}_T) \in \mathcal{D} = \{0, 1\}^T$.

This includes the set of switcher trajectories, $\mathcal{D}_S = \left\{\underline{d} \in \mathcal{D} : 0 < \sum_{t=1}^T D_{it} < T\right\}$ and its complement, the set of non-switcher trajectories, $\mathcal{D}_{NS} = \mathcal{D}_S^C = \mathcal{D} \setminus \mathcal{D}_S$. The set of non-switcher trajectories, in our context, comprises the non-migrant trajectories, i.e., the always-urban trajectory, $d_T = \left\{\underline{d} \in \mathcal{D} : \sum_{t=1}^T D_{it} = T\right\}$, and the always-rural trajectory, $d_N = \left\{\underline{d} \in \mathcal{D} : \sum_{t=1}^T D_{it} = 0\right\}$. These migration histories may entail different distributions of ability or comparative advantage, making it natural to define subpopulations in terms of migration trajectories.

[8] See Proposition 1 in Tjernström *et al.* (2024) for more detail.

Without any additional restrictions, we can identify the average returns to adoption for switcher sub-populations using the following unrestricted GRC model:

$$y_{it} = \sum_{\underline{d} \in \mathcal{D} \setminus d_T} \mu_{\underline{d}} \mathbb{1}\{d_i = \underline{d}\} + \sum_{\underline{d} \in \mathcal{D}_S} \Delta_{\underline{d}} d_{it} \mathbb{1}\{d_i = \underline{d}\} + \kappa_{d_T} d_{it} \mathbb{1}\{\underline{d} = d_T\} + \varepsilon_{it} \tag{12}$$

for any $t \geq 2$. All the parameters in this reduced-form equation have economic meaning: $\mu_{\underline{d}}$ captures the average rural consumption for subpopulation $\underline{d}$, $\Delta_{\underline{d}}$ is the average return to urban for switcher subpopulation $\underline{d}$, and κ_{d_T} is average consumption in urban for the always-urban subpopulation.[9]

Note that in this unrestricted model, $\Delta_{\underline{d}}$ is only nonparametrically identified for the switcher sub-populations, $\underline{d} \in \mathcal{D}_S$. For the non-migrant subpopulations, we can only identify their average consumption in the urban or in the rural market, for d_T and d_N, respectively. In Equation (12), we denote the former average by $\kappa_{d_T} = \mu_{d_T} + \Delta_{d_T}$. Without further restrictions, we cannot separately identify μ_{d_T} and Δ_{d_T}, and similarly for always-rural individuals we can only identify their average rural consumption, μ_{d_N}.

For identification, we must assume strict exogeneity of the error term, ε_{it}, with respect to the explanatory variables. Strict exogeneity implies that, for each worker i, the idiosyncratic error term ε_{it} is uncorrelated with the entire history of the explanatory variables. Formally, letting $d_i = (D_{i1}, D_{i2}, \ldots, D_{iT})$ denote the sequence of sector choices over time (with T equal to the length of the panel), this assumption requires $E[\varepsilon_{it} | d_i, \theta_i, \tau_i] = 0$.[10] In other words, once we condition on θ_i and τ_i, the remaining error term ε_i contains no systematic relationship with D_{it} or other explanatory variables. Intuitively, this rules out any feedback mechanism where past shocks (captured by ε_{it}) influence future decisions. For example, if a worker experiences a positive consumption shock in time t, this shock should not influence their decision to move to an urban area in subsequent periods.

Implicitly, we also assume that the non-pecuniary utility shocks, ν_i, are uncorrelated with ε_{it} in (10′). In our setting, it is reasonable to assume that non-pecuniary shocks are orthogonal to the error term in the consumption equation because they represent individual preferences or local amenities that do not directly influence a worker's productivity in a given sector.

Focusing on migration trajectories as an important dimension of heterogeneity is

[9] Technically, the average consumption for a subpopulation is by definition the expected consumption of this subpopulation. However, we prefer the more intuitive term "average consumption" to "expected consumption" and use it to refer to the latter.

[10] For simplicity, we consider here a model without covariates, but we can allow for exogenous, additively separable covariates.

appealing for several reasons. Trajectories, which can be seen as a person's revealed preference for locations over time, likely capture a substantial amount of information about an individual's unobserved ability and their efficient allocation. For example, if sorting is efficient, individuals with the lowest returns to working in the urban market will seldom be observed in that market and the opposite will be true for those with high returns. Further, the individuals that we observe moving back and forth between the rural and urban markets may be those with relatively similar returns to both markets. On the other hand, if sorting is inefficient, individuals with high returns may not be able to migrate due to market failures such as information barriers or missing credit or insurance markets that could be targeted for policy interventions.

2.2.2 Restricted Group Random Coefficient Model

To estimate the returns to non-migrants, we impose the LCA restriction in Eq. 11 on the unrestricted GRC model in Eq. (12).[11] We do not restrict the sign of the relationship between returns to migration and comparative advantage; rather, we allow this to be determined by the data. This yields the following restricted GRC model:

$$\begin{aligned} y_{it} = &\sum_{\underline{d} \in \mathcal{D} \setminus d_T} \mu_{\underline{d}} + \Delta_{\underline{d}_0} d_{it} + \sum_{\underline{d} \in \mathcal{D} \setminus \underline{d}_0} \phi \left(\mu_{\underline{d}} - \mu_{\underline{d}_0} \right) d_{it} \mathbb{1} \left\{ d_i = \underline{d} \right\} \\ &+ \left(\mu_{\underline{d}_T} + \phi \left(\mu_{\underline{d}_T} - \mu_{\underline{d}_0} \right) \right) d_{it} \mathbb{1} \left\{ \sum_{t=1}^{T} d_{it} = T \right\} + \varepsilon_{it} \end{aligned} \tag{13}$$

for some baseline trajectory $d_0 \in \mathcal{D}_S$.

The relationship in the LCA restriction forms the basis of how we identify returns to non-migrants in our model. Tjernström *et al.* (2024) show that for any two trajectory types $d \neq d'$ the following equality holds:[12]

$$\phi = \frac{\Delta_{\underline{d}} - \Delta_{\underline{d}'}}{\mu_d - \mu_{\underline{d}'}}, \quad \mu_{\underline{d}} \neq \mu_{\underline{d}'}. \tag{14}$$

If the LCA parameter, ϕ, is positive, then urban location has higher returns for the people with higher average consumption in the rural market. Conversely, a negative ϕ would suggest that those who have the lowest rural consumption would receive the highest incremental consumption gains from migrating to the urban market. We can refer to these two scenarios as migration being either "pro-rich" (positive ϕ) or "pro-poor"

[11] This assumption is analogous to that in Suri (2011), who uses it to study the returns to hybrid seed adoption. However, her estimation approach differs from ours.

[12] See Proposition 1 in Tjernström *et al.* (2024).

(negative ϕ).

We estimate the restricted model in Eq. (13) using Generalized Method of Moments (GMM), as in Tjernström *et al.* (2024). While the key identifying variation comes from the switchers in the balanced panel, we include observations that appear in a subset of panel waves in our estimation. We include an "unbalanced" dummy variable that takes a value of one for individuals with at least one missing round of data, as well as the interaction between the "unbalanced" dummy and the urban indicator. Appendix A shows results using the balanced panel.

3 Data

We use longitudinal data from three developing countries—Indonesia, China, and Tanzania—to understand selection and heterogeneity in the returns to migration and test the model's predictions. For each country, we draw from data collected in household surveys designed to collect information on the living standards of people in settings where informal employment, home production, and rural work are prevalent. Our data include detailed information on tens of thousands of individuals across multiple geographies and time periods, and provides rich information on demographic characteristics and place of residence (rural or urban), and comprehensive measures of consumption and income.[13] Sections 3.2, 3.3, and 3.4 provide more detail on each of these datasets.

Table 1: Overview of Data Sources

	Indonesia	China	Tanzania
Data source	Indonesia Family Life Survey (IFLS)	China Family Panel Study (CFPS)	National Panel Survey (NPS)
Number of waves	5	4	3
Years included	1993, 1997/98, 2000, 2007/08, 2014/15	2010, 2012, 2014, 2016	2008/09, 2010/11, 2012/13
Observations	77,744	129,466	34,527
Individuals	34,399	49,398	15,667
Rural/Urban Stayers	92.9%	93%	92%
Ag/Non-Ag Stayers	89%		

[13] We organized and cleaned the Indonesia data from the raw IFLS data files, following Kleemans and Magruder (2018), Hamory *et al.* (2021) and Kleemans (2023). For China and Tanzania, we use the same data that was used in Lagakos, Marshall, Mobarak, Vernot and Waugh (2020). We are grateful to David Lagakos, Samuel Marshall, Mushfiq Mobarak, Corey Vernot, and Michael Waugh for generously sharing their data.

Table 1 provides an overview of our data sources. For all countries, we restrict the data to individuals aged 16 and above with non-missing information on urban/rural status and total consumption.[14] The bottom of Table 1 highlights the proportion of individuals that never migrate between rural and urban areas. In all three countries, more than 90% of individuals always stay in their rural or urban area. We see similar numbers for sectoral switchers, i.e., people who switch between agriculture and non-agriculture, in Indonesia.[15] When we limit the sample to a balanced panel, the proportions of non-migrants fall to 59.6% in Indonesia, 91.8% in China, and 85.5% in Tanzania (see tables A.1, A.2, and A.3 for details).

Figure 1 examines these migration patterns in more detail. The figure shows the proportion of the sample, in each country, that we observe following different aggregated migration histories. In all three datasets, we observe that non-migrants—both always-urban and always-rural—constitute the largest subpopulations. For those who migrate, the most common pattern is a single move from rural to urban. In Indonesia, this is followed by multiple moves, with more of these occurring for individuals who we first observe in an urban location. In Tanzania, one-time urban-to-rural migration is relatively more common than in the other two countries.

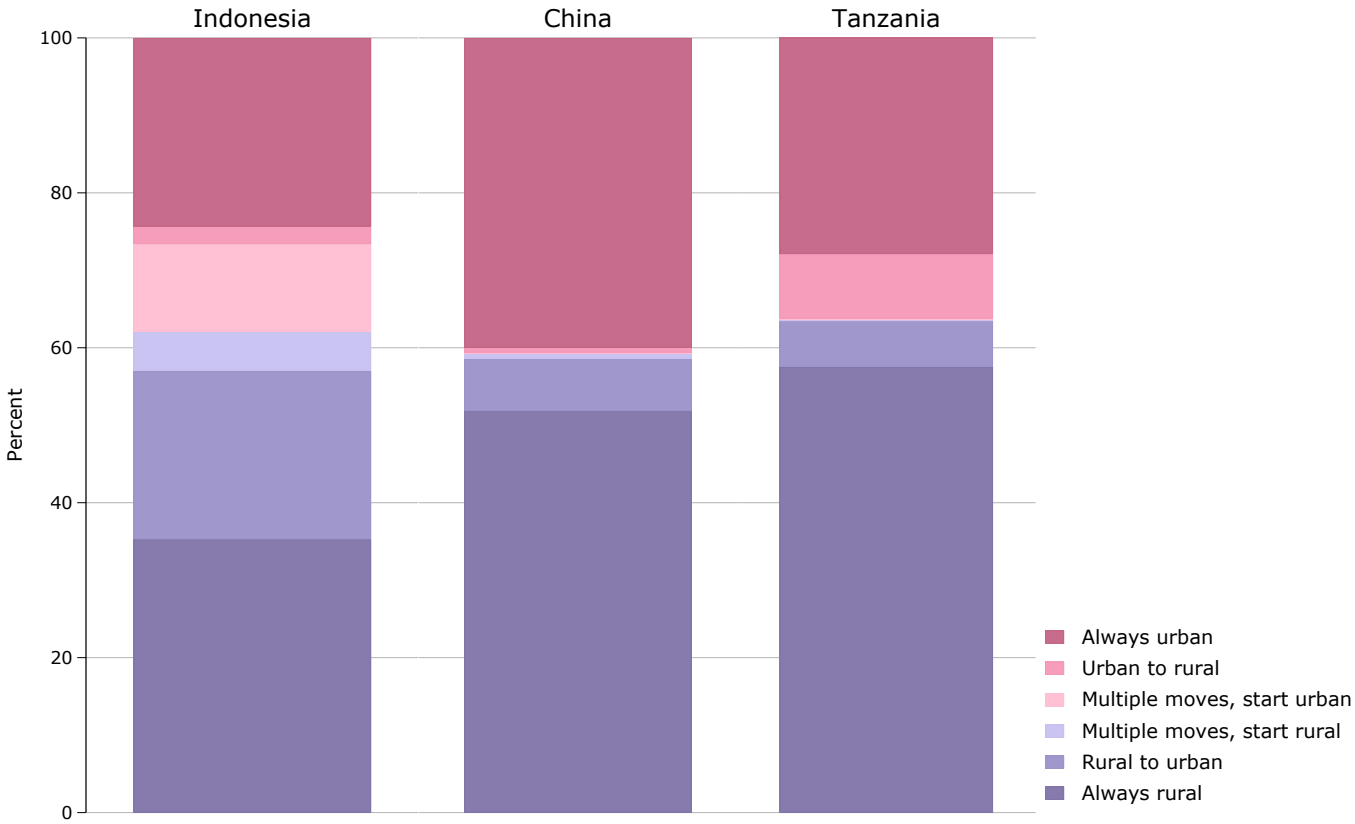
Despite spanning multiple years, these surveys have relatively low attrition rates and represent a significant portion of the population in each country, as detailed in Sections 3.2, 3.3, and 3.4. If individuals from a household were temporarily away or unwell, enumerators made an effort to get a proxy from that household to answer questions on behalf of the missing individual. Nonetheless, despite these efforts, the number of observations in each sample varies across the periods included in our study, meaning that limiting the sample to a balanced panel results in substantially fewer individuals than the unbalanced panel.[16] We acknowledge that entering or dropping out of a sample can be correlated with aging, employment, and migration trajectories and that the motivation of our model assumes individuals are observed in all periods. For that reason, while our main results use the full sample, we include robustness checks with the balanced panel in Section 5.1.

[14]When studying sector of employment or income, we adjust the sample restrictions accordingly, keeping only those with non-missing information on sector or income, respectively.

[15]We can only quantify agriculture/non-agriculture switchers in the Indonesian data as we do not have sectoral data for China and Tanzania.

[16]Despite relatively low attrition rates, Indonesia has the largest drop in observations when we go from the unbalanced to the balanced panel. This is, in part, due to the length of the panel, which spans a 23-year period. Combined with the fact that we restrict our sample to individuals over the age of 16, this results in some attrition due to death.

Figure 1: Migration Patterns



Migration patterns in Indonesia, China, and Tanzania, grouped by their migration histories as follows: those who we never observe moving in the data ("always urban" and "always rural"), those who move once ("urban to rural" and "rural to urban"), and those who move multiple times ("multiple moves, start urban" and "multiple moves, start rural").

3.1 Key Variable Definitions

We use consumption as our main outcome variable. For all surveys, this variable is constructed using detailed expenditure modules, in which respondents are presented with lists of items they consumed and asked to report their total expenditure for each good. In addition to asking for total expenditure purchased, these surveys ask for the approximate value of items that were consumed by the household and that were self-produced.

We prefer consumption over income for several reasons. First, in developing countries, consumption typically measures welfare better than earnings or wages, precisely because it also captures the value of home production and in-kind transfers. Second, computing earnings from wages requires information on hours worked, which tends to introduce measurement error that is correlated with sector and formality. Third, wages or income is missing for a substantial portion of observations in our data, namely those who are not engaged in income-generating activities. In addition to our main analysis on consumption, we report results for income in the Appendix. For all countries, income is the sum of earnings from formal and informal employment and self-employment.

We define an individual as being in an urban location if she reports living in a city or a town, rather than a village. In our data, an individual's current location correlates quite strongly with their current location the first time that we observe them. For example, in Indonesia, we find that individuals born in urban locations are 40-70 percent more likely to be observed in an urban labor market in the first survey wave. This helps lend credence to the notion that individuals may be "stuck" in a sub-optimal labor market, meaning that their skills are not efficiently located and they may have high potential returns to migration. For this reason, we believe our model assumptions are more likely to hold when studying returns to rural-urban migration than returns to sectoral shifts out of agriculture into non-agricultural sectors. Nevertheless, we repeat our analysis for sectoral choice in Indonesia, the only country that consistently provides these data. Sectoral choice is defined by whether or not an individual reports that their primary employment is in non-agriculture. Results for sectoral choice are reported in the Appendix. In the sections that follow, we provide more details on the data source and variable definitions by country.

3.2 Indonesia

Indonesia is the world's fourth largest country with a population size of 275 million. We use data from all five waves of the Indonesia Family Life Survey (henceforth, IFLS), which was collected between 1993 and 2015. During this study period, Indonesia was characterized as a lower middle income country (World Bank, 2024).

The IFLS is representative of 83 percent of the Indonesian population (Strauss et al., 2004). The IFLS is particularly suitable for studies involving migration because of intensive tracking efforts. As a result, attrition is low with re-response rates of over 90 percent between any two consecutive survey waves and 87 percent of households from the first wave were interviewed in all five waves (Strauss, Witoelar and Sikoki, 2016).[17]

In a detailed consumption module, respondents are presented with lists of items and asked to report the value they consumed during a reference period. For all 37 food items the reference period is the past week. For 9 non-food items, such as toiletries, utilities and transportation, respondents are asked about consumption in the past month. For 11 less frequently consumed items, such as clothing, furniture, education and health services, respondents are asked about the past year. For all items, respondents are asked for the value of both purchased and self-produced items.

Table 2 shows summary statistics from Indonesia. The units of observation in this table are individual-year pairs of which there are 93,038 in the data. These come from 29,716 unique individuals, 92.9 percent of whom never switch between rural and urban locations. Columns 2 and 3 show that consumption and income is substantially higher for observations in urban areas than in rural areas. There are also significant differences in the demographic characteristics shown, which we will use as our main control variables in our analysis. Individuals in urban areas are on average two years younger and have received an additional 2.7 years of education.

For Indonesia, we also have information on sector of employment—agriculture vs non-agriculture—which we use to estimate returns to sectoral switching, as documented in the Appendix.

3.3 China

We use Chinese data from four waves of the China Family Panel Study (henceforth, CFPS) that were collected biannually from 2010 to 2016 (Institute of Social Science Survey, 2015). China is currently the second most populous country in the world

[17]Please refer to Kleemans and Magruder (2018), Hamory *et al.* (2021), and Kleemans (2023) for more details on the IFLS data. Unlike Hamory *et al.* (2021), we do not use recall data between survey waves, because consumption is only collected at the time of the survey.

Table 2: Summary Statistics, Indonesia, Unbalanced Panel

	All	Rural	Urban	Difference t-test
Location		52.6%	47.4%	
Log Consumption	12.02	11.84	12.23	-0.39***
	(0.79)	(0.77)	(0.76)	
Log Income	14.88	14.64	15.15	-0.50***
	(1.15)	(1.15)	(1.08)	
Female	0.52	0.51	0.53	-0.01***
	(0.50)	(0.50)	(0.50)	
Age (years)	39.37	39.78	38.91	0.88***
	(14.95)	(15.36)	(14.48)	
Education (years)	8.05	6.79	9.45	-2.65***
	(4.65)	(4.45)	(4.46)	
Household Size	4.89	4.78	5.00	-0.22***
	(2.21)	(2.08)	(2.35)	
Observations	93,038	48,972	44,066	
Individuals	29,716			
Non-switchers	92.9%			

Summary statistics for Indonesia for the unbalanced panel across all five waves. Source: IFLS. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 61,300 observations.

with 1.4 billion inhabitants, and during the time period of our sample, China was characterized as an upper middle income country (World Bank, 2024). The survey is representative of 95 percent of the Chinese population. The initial survey included 16,000 households living in 25 provinces and re-contact rates were 85.3 percent in 2012 and and 89.7 percent in 2014. (Lagakos *et al.*, 2020).

The consumption variable in the CFPS is based on 31 non-food categories of items consumed. Food consumption is divided into two aggregate categories only: purchased items and self-produced food. We use cleaned data files that were generously made available by Lagakos *et al.* (2020).

Table 3 shows summary statistics based on 109,535 individual-year pairs, originating from 34,746 individuals. The share of individuals that we observe migrating between rural and urban areas is considerably lower in China than in Indonesia and Tanzania with only 4.3 percent of individuals engaging in such moves. This may partly be explained by China's Hukou system that effectively restricts internal migration during our study period, for example see Khanna et al. (2019). Similar to in Indonesia, consumption and income is higher in urban area. The education gap is similar to Indonesia with individuals in urban areas having received an additional 2.7 years of education. Unlike Indonesia, household size is smaller in urban areas, and the difference in age is modest.

3.4 Tanzania

Tanzania was a low-income country during our study period from 2008 to 2013 and currently has a population size of 65 million. Following Lagakos *et al.* (2020), we use three waves of the Tanzania National Panel Survey (henceforth, NPS), which were collected in 2008/2009, 2010/2011, and 2012/2013.

The initial sample of the NPS was nationally representative and attrition is the lowest of all three countries. Between waves 2 and 3 attrition was 3.5 percent and between waves 1 and 3, was 4.8 percent. The NPS consumption modules include 72 food items and 46 non-food items. We use the replication data provided by Lagakos *et al.* (2020) who use a spatial price deflator to account for different price levels between rural and urban areas.

We present summary statistics from Tanzania in Table 4. The NPS is smaller than our other surveys and includes 29,864 individual-year pairs from 11,012 individuals. There is more migration between rural-urban areas in Tanzania than in our samples of Indonesia and China. 11.4 percent of individuals are observed in both a rural and an

Table 3: Summary Statistics, China, Unbalanced Panel

	All	Rural	Urban	Difference t-test
Location		54.2%	45.8%	
Log Consumption	10.46	10.26	10.70	-0.44***
	(0.93)	(0.90)	(0.90)	
Log Income	8.74	8.25	9.20	-0.95***
	(1.93)	(2.08)	(1.64)	
Female	0.51	0.51	0.52	-0.01***
	(0.50)	(0.50)	(0.50)	
Age (years)	46.66	46.70	46.60	0.10
	(16.57)	(16.51)	(16.64)	
Education (years)	7.38	6.16	8.83	-2.68***
	(4.92)	(4.68)	(4.80)	
Household Size	4.26	4.56	3.91	0.65***
	(1.91)	(1.98)	(1.76)	
Observations	109,535	59,354	50,181	
Individuals	34,746			
Non-switchers	95.7%			

Summary statistics for China for the unbalanced panel across all waves. Source: China survey. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 49,191 observations.

Table 4: Summary Statistics, Tanzania, Unbalanced Panel

	All	Rural	Urban	Difference t-test
Location		64.6%	35.4%	
Log Consumption	14.89	14.65	15.31	-0.66***
	(0.81)	(0.72)	(0.78)	
Log Income	13.80	13.25	14.49	-1.24***
	(1.93)	(1.82)	(1.83)	
Female	0.52	0.52	0.53	-0.01*
	(0.50)	(0.50)	(0.50)	
Age (years)	36.21	37.22	34.38	2.84***
	(16.84)	(17.58)	(15.23)	
Education (years)	6.69	5.69	8.50	-2.80***
	(3.97)	(3.72)	(3.75)	
Household Size	6.38	6.66	5.86	0.80***
	(4.08)	(4.46)	(3.21)	
Observations	29,864	19,282	10,582	
Individuals	11,012			
Non-switchers	88.6%			

Summary statistics for Tanzania for the unbalanced panel across all waves. Source: Tanzania survey. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 12,052 observations.

urban area at some point during the panel. Similar to Indonesia and China, consumption and income is considerably higher in urban areas. Similar to Indonesia, individuals in urban areas are more than two years younger. Households are larger in Tanzania than in Indonesia and China and there is a starker difference in household size between rural and urban areas. Differences in education between rural and areas are remarkably similar between the three countries with urban residents having an additional 2.67, 2.66, and 2.81 years of education in Indonesia, China and Tanzania, respectively.

4 Results

4.1 Observational Returns

We start with showing results from pooled OLS regressions, in line with Alvarez (2020) and Hamory *et al.* (2021). Table 5 shows results from the three countries in our data, with the results by country in Panels A (Indonesia), Panel B (China), and Panel C (Tanzania). The dependent variable in all regressions is the log of total consumption. Because this variable is measured at the household level, we control for the log of household size (number of household members) in all regressions.

We report results from seven different specifications. Results in column (1) come from a regression of log consumption on an indicator for urban location. The coefficient on the urban dummy therefore reflects the raw consumption gap between rural and urban labor markets. This gap is 40 log points in Indonesia, 52 log points in China, and 74 log points in Tanzania, meaning that average differences range from 50% to 109% higher consumption in urban than in rural areas. In subsequent columns we gradually add controls. In column (2), we add a female indicator, and in column (3), we also add age and age square. For all three countries, the rural-urban consumption gap changes little with the addition of these controls as we move from columns (1) to (2) and (3).

The inclusion of education controls (years of education and years of education squared) in column (4) significantly reduces the rural-urban consumption gap. The rural-urban consumption gap decreases by 13-18 log points, consistent with the fact that urban residents are more educated than rural residents: across the three countries in our data, urban residents have almost three more years of schooling than rural ones. The addition of a linear time trend in column (5) has minimal effect on the estimated consumption gap.

In column (6), we repeat the specification in column (5), including all covariates and a time trend but for a sample of only migrants. In other words, this sample only

includes individuals who we observe switching between rural and urban at least once in the data. For all three countries, the estimated consumption gap reduces substantially. In Indonesia, the consumption gaps drops by 12 log points down to 8.4 log points. In China, restricting the sample to migrants results in a drop of 33 log points, down to an estimated consumption gap of 2.9 log points, which is statistically indistinguishable from 0. In Tanzania, restricting the sample to migrants reduces the consumption gap by 45 log points to 12 log points.

Finally, in column (7), we show the results of a specification that includes individual fixed effects. In the Indonesia and Tanzania data (Panels A and C), the estimated consumption gaps are similar across columns (6) and (7) are similar, suggesting that limiting the identifying variation to migrants-only has a similar effect on the estimated consumption gap as the inclusion of individual fixed effects, which control for time-invariant unobserved characteristics. In China, restricting the sample to migrants-only as in column (6) reduces the estimated consumption gap much more than the inclusion of individual fixed effects in column (7). Internal migration in China is unique due to its Hukou system whereby individuals are assigned Hukou status based on where they were born. When individuals migrate away from the area of their assigned Hukou, they often lose access to various government-provided institutions such as schooling and health care. This effectively increases the cost to migrating, differently so for different subgroups of the population, and leads to distinct selection into migration, as discussed in Section 3. Table A.4 repeats this analysis for the balanced sample. Sample sizes are naturally smaller, but the results are qualitatively similar.

Table 5: OLS Estimates of the Returns to Urban Location on log Consumption

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Indonesia							
Urban	0.402***	0.402***	0.399***	0.222***	0.210***	0.0863***	0.0668***
	(67.39)	(67.50)	(67.60)	(39.73)	(38.28)	(4.96)	(9.42)
Observations	93,026	93,026	93,026	93,026	93,026	6,635	93,026
Individuals	29,710	29,710	29,710	29,710	29,710	1,327	29,710
Adj. R^2	0.18	0.18	0.18	0.31	0.37	0.38	0.59
Panel B: China							
Urban	0.523***	0.523***	0.516***	0.385***	0.375***	0.0287	0.145***
	(73.91)	(73.91)	(74.22)	(55.23)	(54.36)	(1.02)	(9.61)
Observations	109,535	109,535	109,535	109,535	109,535	4,664	109,535
Individuals	34,746	34,746	34,746	34,746	34,746	1,166	34,746
Adj. R^2	0.13	0.13	0.14	0.18	0.27	0.23	0.55
Panel C: Tanzania							
Urban	0.738***	0.739***	0.729***	0.559***	0.572***	0.121***	0.121***
	(65.96)	(66.07)	(65.11)	(52.11)	(54.32)	(7.50)	(8.55)
Observations	29,862	29,862	29,862	29,862	29,862	3,414	29,862
Individuals	11,010	11,010	11,010	11,010	11,010	1,138	11,010
Adj. R^2	0.37	0.37	0.38	0.45	0.49	0.40	0.76
Covariates		Female	& Age^2	All	All	All	All
Time trend					Y	Y	Y
Individual FE							Y
Migrants only						Y	

The dependent variable is log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Column 6 restricts the sample to switchers, i.e. those who we observe switching between rural and urban at least once in our data. All regressions control for log of household size. Other covariates include female, age squared, education (years of schooling), and education squared. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

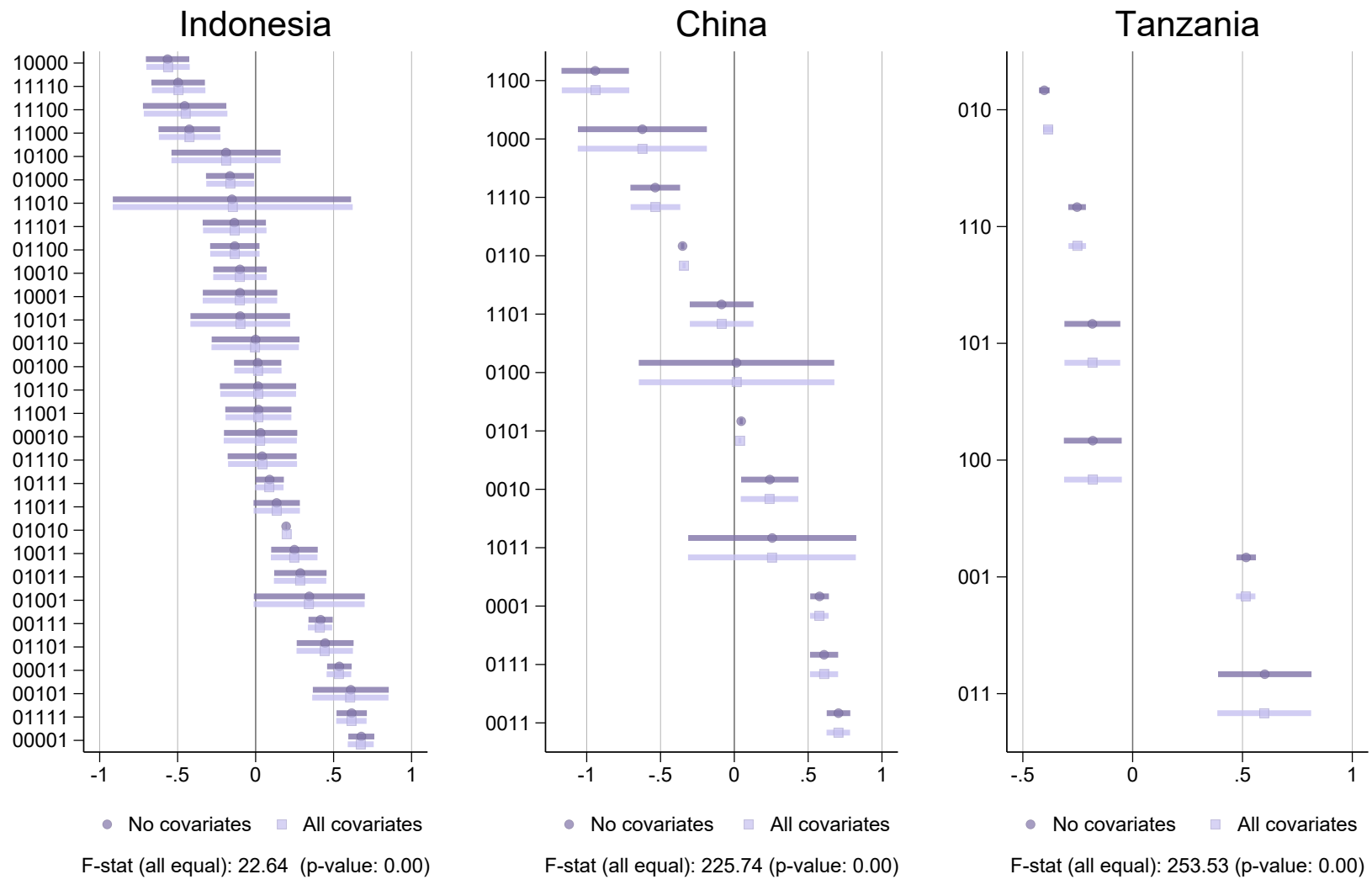
4.2 Heterogeneity by Trajectory Type

Consumption gaps estimated using individual fixed effects in Table 5 represent a weighted average of the returns for different switcher types. The weights will be determined by the extent of within-individual variation in migration status and the variance of returns across individuals. Migrants with greater within-person variance and/or more frequent sectoral changes will influence the overall estimate more. In this section we start fully non-parametrically and test for heterogeneity in the returns for various switcher types as defined by their trajectories, i.e. their migration histories. If returns differ substantially between trajectory types, the FE results could mask this variation in the returns across different migrant types, making it less informative about the returns to migration for any single switcher group, and likely less informative about the returns to migration for non-switchers.

Figure 2 examines the observational returns for switcher trajectories in the data. More precisely, the figure plots the Δ-coefficients from the following regression equation: $y_{it} = \sum_{\underline{d} \in \mathcal{D} \setminus d_T} \mu_{\underline{d}} \mathbb{1}\{d_i = \underline{d}\} + \sum_{\underline{d} \in \mathcal{D}_s} \Delta_{\underline{d}} D_{it} \mathbb{1}\{d_i = \underline{d}\} + X'\delta + \varepsilon_{it}$, where $\mathcal{D} \setminus \mathcal{D}_T$ denotes all trajectories less always-adopters, $\mathcal{D}_S$ denotes the set of switcher trajectories, and X is a vector of covariates. In other words, $\Delta_{\underline{d}}$ is the coefficient on the interaction between indicator variable for urban location and a dummy for each switcher-trajectory, $\underline{d} \in \mathcal{D}_{\mathcal{S}}$.

The Δ-coefficients are sorted by size and the figure additionally shows their 95-percent confidence intervals. For all countries, coefficient values cover a wide range, from negative to positive values. The F-statistic below each graph tests the null hypothesis that the Δ-coefficients for all trajectories in $\mathcal{D}_S$ are equal, which we reject for all three countries. This suggests that relying on a single estimated return for all switcher types is likely to mask important heterogeneity. We now turn to our analysis of a model that adds more structure in order to leverage this heterogeneity in order to obtain the estimated returns for non-movers.

Figure 2: Consumption Returns to Urban Location by Country



Plot of coefficients representing the observational returns to urban location for each switcher trajectory in the data. More precisely, the figure plots the Δ-coefficients from the following regression equation: $y_{it} = \sum_{\underline{d} \in \mathcal{D} \setminus \mathcal{D}_T} \mu_{\underline{d}} \mathbb{1}\{d_i = \underline{d}\} + \sum_{\underline{d} \in \mathcal{D}_s} \Delta_{\underline{d}} D_{it} \mathbb{1}\{d_i = \underline{d}\} + X'\delta + \varepsilon_{it}$, where $\mathcal{D} \setminus \mathcal{D}_T$ denotes the full set of trajectories, excluding always-adopters, $\mathcal{D}_T$; $\mathcal{D}_S$ denotes the set of switcher trajectories, and X is a vector of covariates. In other words, $\Delta_{\underline{d}}$ is the coefficient on the interaction between indicator variable for urban location and a dummy for switcher-trajectories $\underline{d}$. The test statistics below each graph come from an F-test of the equality of the Δ-coefficients for all trajectories in $\mathcal{D}_S$.

4.3 Estimates from Restricted GRC Model

We present the results from the restricted GRC model for Indonesia in Table 6, for China in Table 7, and for Tanzania in Table 8. In the first row, we show the returns for non-migrants, extrapolated based on returns from all switcher trajectories using the linear in comparative advantage (LCA) assumption, as detailed in Section 2.2.2. For Indonesia, the estimated returns for non-migrants are fairly large in column (1) at 422 log points, and gradually reduce as we add covariates in subsequent columns. When we add controls for education and education squared in column (4), the estimated coefficient drops to 208 log points. Further, when we additionally control for a linear time trend in column (5), we estimate the returns to non-migrants to be 119 log points.

Comparing these results to the observational returns in Panel A of Table 5, we note that the estimated returns for non-migrants are 77 percent larger than the 67 log points estimated using individual fixed effects. Our results suggest that the average consumption gap between rural and urban workers controlling for observables and individual fixed effects is considerably smaller than the returns we estimate for non-migrants when we explicitly account for heterogeneity in the returns to migration.

The slope of the extrapolation line, shown as ϕ in the second row, is consistently negative in all specifications in Indonesia. This indicates that those who stand most to gain from migrating to an urban area are those with the lowest initial consumption in rural areas. This group may face constraints to migrate, for example liquidity or information constraints, and alleviating these constraints would allow them to increase consumption and benefit from migrating. We conclude that for Indonesia, migration can be seen as a pro-poor strategy.

Turning to the results from China, shown in Table 7, the estimated returns to non-migrants are strikingly similar to those of Indonesia. The returns to never-movers estimated without any controls in column (1) are 428 log points and decrease with inclusion of additional covariates. Once we include all controls and a linear time trends, the returns are 107 log points. Unlike for Indonesia, however, this is lower than the observational returns of 145 log points that we estimated in Panel B of Table 5. As explained in Section 4.1 the higher observational returns for China may be due to the Hukou system that effectively restricts rural-urban migration.

As is the case for Indonesia, the slope of the extrapolation line, ϕ, for China is consistently negative across all specifications, indicating that rural-to-urban migration is most beneficial for those with the lowest baseline consumption in rural. Therefore, migration appears to act as a pro-poor technology in China as well.

In our Tanzania data, the estimated returns to non-migrants are much larger, as

Table 6: Restricted GRC Estimates of the Returns to Urban Location on log Consumption, Indonesia

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)
Δ_{never}	0.422***	0.423***	0.443***	0.206***	0.119***
	(6.77)	(6.82)	(6.89)	(3.28)	(3.76)
ϕ	-1.421***	-1.407***	-1.519***	-1.607***	-0.598***
	(-14.51)	(-14.62)	(-14.78)	(-13.29)	(-7.52)
Individuals	29,716	29,711	29,711	29,711	29,711
Observations	93,038	93,027	93,027	93,027	93,027
J-stat	111.4	111.5	113.8	101.4	59.4
J-stat (p-value)	3.19e-12	3.08e-12	1.28e-12	1.52e-10	0.000311
Covariates		Female	& Age^2	All	All
Time trend					Y

The dependent variable is the log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Individuals are assigned to trajectories based on their location history across the survey waves. This table reports the extrapolated returns to migrating to an urban location for individuals who are never observed in urban location in the data. Estimates of returns for all trajectories is in Appendix table **XXX**. All columns control for log(household size). Column (2) controls for female, column (3) adds controls for age squared, column (4) adds education (years of schooling) and education squared, and column (5) adds a time trend. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 7: Restricted GRC Estimates of the Returns to Urban Location on log Consumption, China

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)
Δ_{never}	0.428***	0.428***	0.408***	0.337***	0.107***
	(20.85)	(20.85)	(19.63)	(15.32)	(5.28)
ϕ	-0.851***	-0.851***	-0.898***	-0.993***	-0.231*
	(-15.91)	(-15.92)	(-17.13)	(-17.79)	(-1.90)
Individuals	34,746	34,746	34,746	34,746	34,746
Observations	109,535	109,535	109,535	109,535	109,535
J-stat	110.0	110.0	112.5	99.1	19.4
J-stat (p-value)	3.88e-20	3.88e-20	1.18e-20	6.38e-18	0.0131
Covariates		Female	& Age^2	All	All
Time trend					Y

The dependent variable is the log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Individuals are assigned to trajectories based on their location history across the survey waves. This table reports the extrapolated returns to migrating to an urban location for individuals who are never observed in urban location in the data. Estimates of returns for all trajectories is in Appendix table **XXX**. All columns control for log(household size). Column (2) controls for female, column (3) adds controls for age squared, column (4) adds education (years of schooling) and education squared, and column (5) adds a time trend. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 8: Restricted GRC Estimates of the Returns to Urban Location on log Consumption, Tanzania

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)
Δ_{never}	0.784***	0.785***	0.773***	0.805***	0.368***
	(15.50)	(15.55)	(15.84)	(11.38)	(11.37)
ϕ	-1.488***	-1.492***	-1.488***	-2.252***	-0.978***
	(-16.34)	(-16.37)	(-16.74)	(-11.93)	(-6.27)
Individuals	11,012	11,012	11,012	11,012	11,012
Observations	29,864	29,864	29,864	29,864	29,864
J-stat	9.82	9.62	9.26	10.2	24.1
J-stat (p-value)	0.0202	0.0220	0.0260	0.0171	0.0000244
Covariates		Female	& Age^2	All	All
Time trend					Y

The dependent variable is the log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Individuals are assigned to trajectories based on their location history across the survey waves. This table reports the extrapolated returns to migrating to an urban location for individuals who are never observed in urban location in the data. Estimates of returns for all trajectories is in Appendix table **XXX**. All columns control for log(household size). Column (2) controls for female, column (3) adds controls for age squared, column (4) adds education (years of schooling) and education squared, and column (5) adds a time trend. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

shown in 8. Without covariates, the returns are 784 log points and unlike for Indonesia and China, our estimates are not very sensitive to adding covariates in columns (2) to (4). Once we control for all our covariates and a time trend in column (5), the returns for non-migrants remain high at 368 log points, around three times the average observational returns that we estimate with fixed effects in Table 5 . As is the case in the two other countries, we estimate a consistently negative slope coefficient, ϕ, which again is consistent with rural-urban migration having greater benefits for the poorest individuals in rural areas.

5 Robustness

5.1 Balanced panel

In this section, we present results for the restricted GRC model, estimated using the balanced panel. Since sample attrition can be correlated with aging, employment, and migration trajectories, we believe that it is important to examine whether our results are robust to restriction the sample to a balanced panel. Tables 9, 10, and 11 show the results for Indonesia, China, and Tanzania, respectively. The results for China and Tanzania, shown in Tables 10 and 11, are nearly identical to those obtained in the full sample.

The results for Indonesia, shown in Table 9, are also similar to those estimated with the unbalanced panel, albeit slightly attenuated. In particular, the estimated returns to urban migration for non-migrants fall to 69 log points once we include all controls and a time trend. This estimate is very similar to the one that we obtain using fixed effects in the unbalanced panel (column 7 in Table 5). However, a more appropriate comparison would be the fixed-effects model estimated on a balanced panel (see column 7 in Table A.4), which are much smaller at 29 log points.

The length of the time period over which the Indonesian data were collected leads to a drastic reduction in the available sample size, which is why our preferred estimates are those with the full sample. Comparing the summary statistics across the full and balanced samples, with the latter reported in Table A.1, we can see that the age profile of respondents differs quite substantially across the two samples. In the full sample, rural respondents are on average around one year younger than their urban counterparts. In contrast, this difference is reversed in the full sample, with urban respondents being nearly two years older. This may reflect different mortality patterns, which would also skew attrition patterns, reinforcing our prefernce for the full sample.

Table 9: Restricted GRC Estimates of the Returns to Urban Location on log Consumption, Indonesia, Balanced Panel

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)
Δ_{never}	0.418***	0.419***	0.333***	0.196***	0.0698***
	(6.78)	(6.82)	(6.77)	(3.83)	(2.67)
ϕ	-1.375***	-1.364***	-1.000***	-1.000***	-0.429***
	(-14.55)	(-14.63)	(-18.42)	(-13.20)	(-4.82)
Individuals	3,284	3,284	3,284	3,284	3,284
Observations	16,420	16,420	16,420	16,420	16,420
J-stat	110.9	110.9	105.8	54.4	54.9
J-stat (p-value)	3.87e-12	3.90e-12	2.83e-11	0.00202	0.00117
Covariates		Female	& Age^2	All	All
Time trend					Y

The dependent variable is the log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Individuals are assigned to trajectories based on their location history across the survey waves. This table reports the extrapolated returns to migrating to an urban location for individuals who are never observed in urban location in the data. Estimates of returns for all trajectories is in Appendix table **XXX**. All columns control for log(household size). Column (2) controls for female, column (3) adds controls for age squared, column (4) adds education (years of schooling) and education squared, and column (5) adds a time trend. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 10: Restricted GRC Estimates of the Returns to Urban Location on log Consumption, China, Balanced Panel

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)
Δ_{never}	0.429***	0.429***	0.412***	0.333***	0.108***
	(20.88)	(20.88)	(19.89)	(15.25)	(5.45)
ϕ	-0.843***	-0.843***	-0.883***	-0.993***	-0.252**
	(-15.84)	(-15.84)	(-16.97)	(-18.91)	(-2.11)
Individuals	14,214	14,214	14,214	14,214	14,214
Observations	56,855	56,855	56,855	56,855	56,855
J-stat	112.2	112.2	114.4	100.1	19.8
J-stat (p-value)	1.38e-20	1.37e-20	4.81e-21	4.14e-18	0.0111
Covariates		Female	& Age^2	All	All
Time trend					Y

The dependent variable is the log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Individuals are assigned to trajectories based on their location history across the survey waves. This table reports the extrapolated returns to migrating to an urban location for individuals who are never observed in urban location in the data. Estimates of returns for all trajectories is in Appendix table **XXX**. All columns control for log(household size). Column (2) controls for female, column (3) adds controls for age squared, column (4) adds education (years of schooling) and education squared, and column (5) adds a time trend. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 11: Restricted GRC Estimates of the Returns to Urban Location on log Consumption, Tanzania, Balanced Panel

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)
Δ_{never}	0.780***	0.782***	0.770***	0.800***	0.364***
	(15.49)	(15.54)	(15.79)	(11.33)	(8.14)
ϕ	-1.483***	-1.487***	-1.483***	-2.240***	-0.966***
	(-16.34)	(-16.37)	(-16.71)	(-11.88)	(-5.13)
Individuals	7,842	7,842	7,842	7,842	7,842
Observations	23,526	23,526	23,526	23,526	23,526
J-stat	9.85	9.64	9.32	10.2	22.9
J-stat (p-value)	0.0199	0.0219	0.0253	0.0168	0.0000424
Covariates		Female	& Age^2	All	All
Time trend					Y

The dependent variable is the log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Individuals are assigned to trajectories based on their location history across the survey waves. This table reports the extrapolated returns to migrating to an urban location for individuals who are never observed in urban location in the data. Estimates of returns for all trajectories is in Appendix table **XXX**. All columns control for log(household size). Column (2) controls for female, column (3) adds controls for age squared, column (4) adds education (years of schooling) and education squared, and column (5) adds a time trend. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

6 Conclusion

In this paper, we aim to reconcile divergent estimates on the returns to rural-urban migration in developing countries by accounting for the role of selection and heterogeneity in returns to migration. Motivated by a multi-period Roy model, we formulate a correlated random coefficient model that allows for location-specific skills and heterogeneous returns. We use a person's migration history as the relevant dimension of heterogeneity, which we interpret as a person's revealed preference for location and that is conveniently observable in the data. Then, we build on Suri (2011), Lemieux (1998), and Tjernström *et al.* (2024) to leverage a linear relationship between comparative and absolute advantage that allows us to extrapolate the returns identified from migrant sub-populations to non-migrants. The group of non-migrants plays a central role in debates on misallocation, whereby non-migrants may reside in areas where they do not live up to their economic potential. Moreover, this group is of specific interest to policymakers determining whether to promote migration as a development strategy.

We test our model using detailed survey data from three developing countries, Indonesia, China and Tanzania. In line with the existing literature, we confirm the existence of large cross-sectional consumption gaps between rural and urban areas. The gap narrows when we use the panel component of the data and include demographic controls, a time trends, and especially when we include individual fixed effects. We then estimate our Group Random Coefficient model and test its assumptions. Results show a distinct pattern of the relationship between absolute and comparative advantage that is remarkably consistent across the three countries: individuals with the lowest consumption in rural areas stand most to gain from migrating to urban areas. As such, migration can be seen as a pro-poor technology but individuals with the lowest consumption may face barriers to migrate, which might be due to borrowing, liquidity, or information constraints. We conclude that individuals are inefficiently sorted across space and that promoting migration would reduce misallocation and, thereby, increase overall growth.

References

Adamopoulos, T., Brandt, L., Leight, J. and Restuccia, D. (2022). Misallocation, Selection, and Productivity: A Quantitative Analysis With Panel Data From China. *Econometrica*, **90** (3), 1261–1282.

Alvarez, J. A. (2020). The Agricultural Wage Gap: Evidence from Brazilian Microdata. *American Economic Journal: Macroeconomics*, **12** (1), 153–73.

Alvarez-Cuadrado, F., Amodio, F. and Poschke, M. (2023). Selection, Absolute Advantage, and the Agricultural Productivity Gap. Working paper.

Bazzi, S., Gaduh, A., Rothenberg, A. D. and Wong, M. (2016). Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia. *American Economic Review*, **106** (9), 2658–98.

Bryan, G., Chowdhury, S. and Mobarak, A. M. (2014). Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. *Econometrica*, **82** (5), 1671–1748.

— and Morten, M. (2019). The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia. *Journal of Political Economy*, **127** (5), 2229–2268.

Chen, J. J. (2013). Identifying non-cooperative behavior among spouses: Child outcomes in migrant-sending households. *Journal of Development Economics*, **100** (1), 1–18.

Cornelissen, T., Dustmann, C., Raute, A. and Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, **41**, 47–60.

Donovan, K. and Schoellman, T. (2023). The Role of Labor Market Frictions in Structural Transformation. *Oxford Development Studies*, **51** (4), 362–374.

Gai, Q., Guo, N., Li, B., Shi, Q. and Zhu, X. (2024). Migration Costs, Sorting, and Agricultural Productivity Gap.

Gollin, D., Lagakos, D. and Waugh, M. E. (2014). The Agricultural Productivity Gap. *The Quarterly Journal of Economics*, **129** (2), 939–993.

Hamory, J., Kleemans, M., Li, N. Y. and Miguel, E. (2021). Reevaluating Agricultural Productivity Gaps with Longitudinal Microdata. *Journal of the European Economic Association*, **19** (3), 1522–1555.

Heckman, J. J. and Urzua, S. (2010). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, **156** (1), 27–37.

— and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, **73** (3), 669–738.

— and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, **96** (8), 4730–4734.

Herrendorf, B. and Schoellman, T. (2018). Wages, Human Capital, and Barriers to Structural Transformation. *American Economic Journal: Macroeconomics*, **10** (2), 1–23.

Kleemans, M. (2023). Migration Choice under Risk and Liquidity Constraints. Working paper.

— and Magruder, J. (2018). Labour market responses to immigration: Evidence from internal migration driven by weather shocks. *The Economic Journal*, **128** (613), 2032–2065.

Lagakos, D. (2020). Urban-Rural Gaps in the Developing World: Does Internal Migration Offer Opportunities? *Journal of Economic Perspectives*, **34** (3), 174–192.

—, Marshall, S., Mobarak, A. M., Vernot, C. and Waugh, M. E. (2020). Migration Costs and Observational Returns to Migration in the Developing World. *Journal of Monetary Economics*, **113**, 138–154.

— and Waugh, M. E. (2013). Selection, Agriculture, and Cross-Country Productivity Differences. *American Economic Review*, **103** (2), 948–980.

Lemieux, T. (1998). Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection. *Journal of Labor Economics*, **16** (2), 261–291.

Pulido, J. and Święcki, T. (2021). Barriers to Mobility or Sorting? Sources and Aggregate Implications of Income Gaps across Sectors in Indonesia.

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, **3** (2), 135–146.

Strauss, J., Witoelar, F. and Sikoki, B. (2016). *The Fifth Wave of the Indonesia Family Life Survey: Overview and Field Report: Volume 1*. RAND Corporation.

Suri, T. (2011). Selection and Comparative Advantage in Technology Adoption. *Econometrica*, **79** (1), 159–209.

Tjernström, E., Ghanem, D., Barriga-Cabanillas, O., Lybbert, T. J., Michuda, A. and Michler, J. D. (2024). Comment on Suri (2011) "Selection and Comparative Advantage in Technology Adoption". Working paper.

Verdier, V. (2020). Average Treatment Effects for Stayers with Correlated Random Coefficient Models of Panel Data. *Journal of Applied Econometrics*, **35** (7), 917–939.

Vermeulen, F. (2002). Collective household models: principles and main results. *Journal of Economic Surveys*, **16** (4), 533–564.

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

Appendix

A Additional Results: Balanced Panel

Table A.1: Summary Statistics, Indonesia, Balanced Panel

	All	Rural	Urban	Difference t-test
Location		55.2%	44.8%	
Log Consumption	12.00	11.82	12.21	-0.39***
	(0.80)	(0.80)	(0.76)	
Log Income	14.92	14.73	15.16	-0.43***
	(1.13)	(1.14)	(1.07)	
Female	0.50	0.48	0.52	-0.04***
	(0.50)	(0.50)	(0.50)	
Age (years)	43.45	42.59	44.52	-1.93***
	(12.16)	(12.19)	(12.03)	
Education (years)	7.57	6.63	8.72	-2.09***
	(4.40)	(4.27)	(4.28)	
Household Size	4.89	4.82	4.99	-0.17***
	(2.09)	(1.98)	(2.21)	
Observations	16,420	9,059	7,361	
Individuals	3,284			
Non-switchers	59.6%			

Summary statistics for Indonesia for the balanced panel across all five waves. Source: IFLS. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 12,510 observations.

Table A.2: Summary Statistics, China, Balanced Panel

	All	Rural	Urban	Difference t-test
Location		56.2%	43.8%	
Log Consumption	10.39	10.21	10.63	-0.43***
	(0.92)	(0.89)	(0.89)	
Log Income	8.60	8.08	9.13	-1.05***
	(1.91)	(2.06)	(1.58)	
Female	0.52	0.51	0.53	-0.01***
	(0.50)	(0.50)	(0.50)	
Age (years)	49.28	49.34	49.20	0.13
	(14.61)	(14.38)	(14.91)	
Education (years)	6.92	5.70	8.49	-2.79***
	(4.78)	(4.46)	(4.73)	
Household Size	4.12	4.39	3.77	0.62***
	(1.81)	(1.87)	(1.67)	
Observations	56,855	31,968	24,887	
Individuals	14,214			
Non-switchers	91.8%			

Summary statistics for China for the balanced panel across all waves. Source: China survey. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 25,530 observations.

Table A.3: Summary Statistics, Tanzania, Balanced Panel

	All	Rural	Urban	Difference t-test
Location		64.5%	35.5%	
Log Consumption	14.85	14.61	15.29	-0.68***
	(0.81)	(0.71)	(0.79)	
Log Income	13.86	13.29	14.55	-1.26***
	(1.92)	(1.81)	(1.83)	
Female	0.52	0.52	0.52	-0.00
	(0.50)	(0.50)	(0.50)	
Age (years)	37.88	39.07	35.72	3.35***
	(16.58)	(17.16)	(15.24)	
Education (years)	6.59	5.55	8.48	-2.93***
	(4.03)	(3.75)	(3.83)	
Household Size	6.27	6.49	5.86	0.63***
	(4.06)	(4.43)	(3.22)	
Observations	23,526	15,165	8,361	
Individuals	7,842			
Non-switchers	85.5%			

Summary statistics for Tanzania for the balanced panel across all waves. Source: Tanzania survey. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 9,760 observations.

Table A.4: OLS Estimates of the Returns to Urban Location on log Consumption, Balanced Sample

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Indonesia							
Urban	0.405***	0.406***	0.386***	0.240***	0.196***	0.0863***	0.0298*
	(26.49)	(26.53)	(24.44)	(16.49)	(13.84)	(4.96)	(1.92)
Observations	16,420	16,420	16,420	16,420	16,420	6,635	16,420
Individuals	3,284	3,284	3,284	3,284	3,284	1,327	3,284
Adj. R^2	0.17	0.17	0.20	0.31	0.38	0.38	0.60
Panel B: China							
Urban	0.508***	0.509***	0.503***	0.358***	0.342***	0.0287	0.0958***
	(48.91)	(48.91)	(49.07)	(34.85)	(33.37)	(1.02)	(4.51)
Observations	56,855	56,855	56,855	56,855	56,855	4,664	56,855
Individuals	14,214	14,214	14,214	14,214	14,214	1,166	14,214
Adj. R^2	0.13	0.13	0.14	0.19	0.28	0.23	0.55
Panel C: Tanzania							
Urban	0.748***	0.748***	0.740***	0.562***	0.572***	0.121***	0.112***
	(57.93)	(57.99)	(57.18)	(45.57)	(47.28)	(7.50)	(7.70)
Observations	23,526	23,526	23,526	23,526	23,526	3,414	23,526
Individuals	7,842	7,842	7,842	7,842	7,842	1,138	7,842
Adj. R^2	0.37	0.38	0.38	0.46	0.50	0.40	0.76
Covariates		Female	& Age^2	All	All	All	All
Time trend					Y	Y	Y
Individual FE							Y
Migrants only						Y	

The dependent variable is log of total consumption. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. The sample in this table is restricted to individuals with observations in all waves of the data. Column 6 restricts the sample to switchers, i.e. those who we observe switching between rural and urban at least once in our data. All regressions control for log of household size. Other covariates include female, age squared, education (years of schooling), and education squared. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

B Additional Results: Non-agricultural

Table B.5: Summary Statistics, Indonesia, Unbalanced Panel

	All	Rural	Urban	Difference t-test
Non-Agricultural		33.9%	66.1%	
Log Consumption	12.05	11.75	12.20	-0.46***
	(0.79)	(0.74)	(0.78)	
Log Income	14.89	14.41	15.09	-0.68***
	(1.14)	(1.15)	(1.08)	
Female	0.44	0.43	0.45	-0.02***
	(0.50)	(0.49)	(0.50)	
Age (years)	39.94	43.40	38.16	5.24***
	(13.61)	(15.02)	(12.45)	
Education (years)	7.97	5.54	9.22	-3.67***
	(4.63)	(3.92)	(4.47)	
Household Size	4.80	4.69	4.86	-0.16***
	(2.16)	(2.04)	(2.22)	
Observations	93,038	46,797	69,289	
Individuals	29,716			
Non-switchers	92.9%			

Summary statistics for Indonesia for the unbalanced panel across all five waves. Source: IFLS. The table reports means and standard deviations (in parentheses) based on individual-year pairs. See section 3 for further details. All variables have the same number of observations, except for income, which is missing for some observations. Income has 61,300 observations.

Table B.6: OLS Estimates of the Returns to Non-Agricultural Sector on log Consumption, Unbalanced Sample

Dep. var: log(consumption)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Indonesia							
Non-agricultural	0.402***	0.402***	0.399***	0.222***	0.210***	0.0863***	0.0668***
	(67.39)	(67.50)	(67.60)	(39.73)	(38.28)	(4.96)	(9.42)
Observations	93,026	93,026	93,026	93,026	93,026	6,635	93,026
Individuals	29,710	29,710	29,710	29,710	29,710	1,327	29,710
Adj. R^2	0.18	0.18	0.18	0.31	0.37	0.38	0.59
Covariates		Female	& Age^2	All	All	All	All
Time trend					Y	Y	Y
Individual FE							Y
Migrants only						Y	

The dependent variable is log of total consumption. Non-agricultural is an indicator equal to one for individuals who report working in the non-agricultural sector. Column 6 restricts the sample to switchers, i.e. those who we observe switching between rural and urban at least once in our data. All regressions control for log of household size. Other covariates include female, age squared, education (years of schooling), and education squared. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

C Additional Results: Income

Table C.7: OLS Estimates of the Returns to Urban Location on log Income

Dep. var: log(income)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Indonesia							
Urban	0.516***	0.540***	0.524***	0.284***	0.274***	0.0896**	0.0577***
	(42.74)	(46.53)	(45.94)	(26.23)	(25.17)	(2.55)	(4.12)
Observations	57,468	57,468	57,468	57,468	57,468	3,513	57,468
Individuals	19,299	19,299	19,299	19,299	19,299	704	19,299
Adj. R^2	0.052	0.11	0.12	0.24	0.25	0.26	0.54
Panel B: China							
Urban	0.891***	0.950***	0.896***	0.695***	0.696***	-0.0143	0.216***
	(39.89)	(43.43)	(48.66)	(36.56)	(36.53)	(-0.11)	(3.24)
Observations	41,107	41,107	41,107	41,107	41,107	305	41,107
Individuals	17,199	17,199	17,199	17,199	17,199	77	17,199
Adj. R^2	0.057	0.090	0.27	0.31	0.31	0.22	0.55
Panel C: Tanzania							
Urban	1.198***	1.208***	1.215***	0.881***	0.865***	0.136	0.0495
	(26.06)	(26.57)	(26.96)	(19.76)	(19.53)	(1.20)	(0.59)
Observations	9,339	9,339	9,339	9,339	9,339	777	9,339
Individuals	3,874	3,874	3,874	3,874	3,874	259	3,874
Adj. R^2	0.098	0.11	0.12	0.19	0.20	0.10	0.47
Covariates		Female	& Age^2	All	All	All	All
Time trend					Y	Y	Y
Individual FE							Y
Migrants only						Y	

The dependent variable is log of income. Urban is an indicator equal to one for individuals who report living in a city or town, as opposed to a village. Column 6 restricts the sample to switchers, i.e. those who we observe switching between rural and urban at least once in our data. All regressions control for log of household size. Other covariates include female, age squared, education (years of schooling), and education squared. We report robust standard errors, clustered at the individual level, in parentheses. Stars denote: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.