How Much Sorting Matters for Sectoral Productivity Gaps: Evidence from Indonesia

Isabella Germinario

October 25, 2025

Abstract

This paper examines how individual sorting contributes to sectoral productivity gaps in Indonesia between 1993 and 2000, using data from the Indonesian Family Life Survey (IFLS). I find that sector-wide efficiency differences, not sorting based on unobserved comparative advantage, account for most of the gap. This result contrasts sharply with prior literature and stems from an empirical strategy that separates fixed effects unrelated to sectoral choice from latent abilities, uses information from both switchers and stayers, and avoids strong distributional assumptions on comparative advantage. The paper makes two key contributions. Methodologically, it provides a novel application of a correlated random coefficients (CRC) framework to sectoral productivity gaps, with modifications that offer a more flexible characterization of how unobserved comparative advantage influences sectoral choice. Conceptually, it links the aggregate selection effects from this framework to the classic Roy model, allowing direct comparison with standard selection-correction approaches. The findings suggest that for Indonesia in the 1990s, policies aimed at reducing sector-wide inefficiencies were key to narrowing productivity gaps, highlighting a significant potential role for government in improving economic performance.

1 Introduction

Across countries, agriculture in poor economies exhibits a much larger productivity shortfall than non-agriculture. Given that most workers in low-income countries are employed in this less productive sector, improvements in agricultural productivity are pivotal for narrowing cross-country income gaps (Gollin et al., 2002; Caselli, 2005; Chanda and Dalgaard, 2008; Restuccia et al., 2008). Poor countries exhibit a pronounced and persistent productivity disparity between the agriculture and non-agriculture sectors, posing a fundamental development challenge (McMillan and Rodrik, 2014). Implementing effective pro-growth policies requires a deep understanding of the mechanisms underlying the substantial productivity gap between the agricultural and non-agricultural sectors in developing countries.

The structural transformation literature, pioneered by Lewis' seminal analysis (1954) of the dual economy and inter-sectoral labour movements, has since adopted the term, Agricultural Productivity Gap (APG)—the ratio of value-added per worker in non-agriculture relative to agriculture, to quantify sectoral productivity gap (Gollin et al., 2014). In what follows, I use the terms "sectoral productivity gap" and "APG" interchangeably.

In this body of literature, two prominent potential explanations for substantial APG in developing countries are: One is misallocation, where distortions and frictions that keep land, capital, inputs, and labour from flowing to their highest-productivity uses—both within agriculture and across sectors (Restuccia et al., 2008; Bryan et al., 2014; Munshi and Rosenzweig, 2016; Alvarez-Cuadrado et al., 2017; Pulido and Świecki, 2019; Lagakos, 2020; Gai et al., 2021). The alternative hypothesis emphasizes self-selection: a subsistence food requirement keeps workers with low agricultural productivity in farming, depressing agriculture's average productivity and, by composition, widening the observed cross-sector productivity gap in poorer countries (Lagakos and Waugh, 2013).

While both explanations can generate significant agricultural productivity gaps, they differ in mechanism and policy implications: misallocation implies inefficiencies that policy could correct, whereas sorting reflects productivity differences arising from individual heterogeneity. Distinguishing between these forces is crucial for understanding what drives observed gaps and how policy should respond.

This paper examines how much individual selection based on unobserved comparative advantage contributes to sectoral productivity gaps in Indonesia, a low-income country that underwent considerable structural transformation between 1993 and 2014. The answer to this question has far-reaching policy implications. Understanding how much individual sorting contributes to APG is therefore central to designing effective strategies for structural transformation and sustainable growth in low-income countries.

Since Lagakos and Waugh's (2013) self-selection hypothesis, a growing literature has sought to measure how much selection explains productivity gaps in developing countries (Lagakos and Waugh, 2013; Pulido and Świecki, 2019; Alvarez, 2020; Lagakos et al., 2020; Alvarez-Cuadrado et al., 2020; Adamopoulos et al., 2022). These studies consistently report that selection is important, but the estimated magnitudes vary widely. This variation reflects the difficulty of the task: selection stems from latent, sector-specific heterogeneity that workers internalize when choosing sectors. These unobserved abilities affect both sector choice and earnings, inducing endogeneity. The stakes are high, because mismeasuring selection risks misdirects policy—leading governments to focus on moving workers across sectors when the real constraint may be technology, or vice versa.

In the APG literature, there are two prevailing approaches. The first applies two-way fixed effects (TWFE) to panel data, leveraging switchers and controlling for time-invariant individual heterogeneity. It identifies the average within-individual change in earnings associated with switching sectors. This has been interpreted either as low barriers to mobility (when gains are small) or as evidence of selection by comparing results before and after including individual fixed effects. While TWFE can difference out time-invariant heterogeneity, it cannot isolate comparative advantage, because comparative advantage requires comparing an individual's abilities across sectors, whereas the fixed effect is at the individual level, not the individual–sector level. As a result, TWFE is not designed to deliver a measure of sorting on unobserved comparative advantage.

The second approach imposes a distribution for unobserved abilities (e.g., joint normal or Fréchet) and recovers a closed-form parameter in the spirit of the Roy (1951) framework. While tractable, these strong assumptions rarely have empirical support (Heckman and Honore, 1990) and can heavily influence estimated magnitudes. These challenges motivate an

alternative approach that targets unobserved comparative advantage directly while avoiding restrictive distributional assumptions.

To address these limitations, this paper adopts a Correlated Random Coefficient (CRC) framework, following Suri (2011), and applies it to the context of sectoral choice and productivity gaps. The CRC approach models unobserved abilities as time-invariant and sector-specific, which distinguishes individual fixed effects that matter for sectoral choice from those that do not. By exploiting the information embedded in individuals' sectoral choice histories, this method reveals their latent comparative advantages and recovers the selection effect without imposing distributional assumptions on unobserved heterogeneity.

Building on the original framework, this paper makes three adjustments. First, I redefine absolute advantage to reflect its relationship with comparative advantage and how comparative advantage affects sector choice in this setting. Second, I provide an economic interpretation in which sectoral choice histories contain information and reveal unobserved comparative advantage; choice trajectories are therefore informative about latent sector-specific abilities. Third, upon aggregation from individual earnings at the sector level, the adapted framework shows that sorting contributes to the observed APG through two components: a selection effect—the extra return to unobserved abilities where they are more rewarded in non-agriculture relative to agriculture—and the average difference in unobserved abilities between agricultural and non-agricultural workers. In this aggregated formulation, the first component is directly connected to the effect of the selection-correction term in the classic Roy model.

To make the selection term in the CRC approach transparent, I map it to the classical Roy framework. In Roy, "positive selection into a sector" means that those who choose it tend to have higher sector-specific ability and therefore higher expected earnings than the sector's unconditional average. The CRC selection term captures this same idea in a relative, comparative-advantage sense—how strongly non-agriculture rewards the relevant ability compared with agriculture. The mapping shows that the CRC object aligns with the Roy notion of selection while avoiding strong distributional assumptions, and it provides a common language to compare CRC estimates with standard selection-correction approaches.

To examine the role of individual sorting on the sectoral productivity gap, I use the

Indonesia Family Life Survey (IFLS) (Frankerberg et al., 1995), a nationally representative panel dataset spanning five waves from 1993 to 2014 and covering about 80 percent of the population. During this period, Indonesia moved from low- to lower-middle-income status, with per capita incomes roughly doubling and the share of agricultural employment falling from approximately 46% in 1993 to 34% in 2014 Group (2025). The IFLS is particularly well-suited for studying sorting and productivity gaps because of its rich panel structure, which tracks both sector choices and earnings. In this paper, I focus on the first three waves (1993–2000), before Indonesia's sweeping political and institutional changes after 2000, to provide a clean setting for analyzing sectoral productivity gaps.

Applying this framework to the first three waves of the IFLS (1993–2000), I find that Indonesia exhibited large sectoral productivity gaps, but individual sorting played only a limited role in explaining them. Roughly 80 percent of workers remained in their initial sectors, and among those who switched, the estimated selection effect is 0.39 log points (SE = 0.613), statistically indistinguishable from zero. The average earnings gap between sectors is about 1.07 log points. Of this gap, the average difference in unobserved abilities between farmers and non-farmers accounts for about 2%, while sector-wide productivity differences account for approximately 39%; the remainder reflects observed covariates and residual components captured by the empirical specification. These findings stand in contrast to other studies using the same data that report large selection effects, underscoring how methodological choices shape conclusions about the sources of APG.

This contrast arises for two reasons. First, unlike TWFE, which absorbs all unobserved time-invariant individual heterogeneity, the empirical framework here separates latent abilities that are irrelevant to sectoral choice from those that are, modelling unobserved comparative advantage as the sector-specific component of ability rather than a single person-level fixed effect. Second, it avoids distributional assumptions about latent abilities by using individuals' choice trajectories to recover the empirical distribution of unobserved comparative advantage; this reveals much greater dispersion than a normal distribution implies. Together, these features explain why sorting plays only a minor role in this setting.

This paper makes two main contributions. Methodologically, it extends a correlated random coefficient (CRC) framework to study sectoral productivity gaps and estimate self-

selection without relying on strong distributional assumptions. The approach redefines absolute advantage, uses individuals' choice trajectories to infer unobserved comparative advantage, and aggregates selection effects to the sector level. Conceptually, it links the estimated selection terms to the classic Roy-model structure, allowing direct comparison with standard selection-correction approaches. Together, these contributions show that sector-wide productivity differences account for most of the observed gap in Indonesia during the 1990s, while sorting based on unobserved comparative advantage plays only a minor role.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature on sectoral productivity gaps and self-selection. Section 3 defines the measurement of APG and develops the empirical framework based on Suri's (2011) correlated random coefficient model, detailing the adaptations for this setting and the mapping to the Roy model. Section 4 sets out the identification strategy and estimation procedures. Section 5 introduces the dataset, provides descriptive evidence, and presents baseline results for the first three IFLS waves. Section 6 relates the results to the existing TWFE and Roy-model evidence, reconciling differences in estimated selection effects and drawing policy implications. Section 7 concludes.

2 Literature Review

This section situates the paper within the literature on sectoral productivity gaps (APG), with particular attention to studies examining the role of individual self-selection. Two dominant hypotheses explain persistent productivity gaps in low-income countries: misallocation of resources and sorting based on comparative advantage. While not mutually exclusive, distinguishing their relative importance is critical for interpreting observed gaps. If misallocation dominates, then the APG signals inefficiencies that policy can alleviate. If sorting dominates, much of the gap reflects workers' latent abilities, while welfare-enhancing interventions, such as building irrigation systems and improving marketplace access for agricultural products, play a smaller role.

The review proceeds in three steps. First, it discusses the origins of the APG literature and the theoretical motivation for focusing on selection. Second, it synthesizes empirical

evidence on the magnitude of selection effects, highlighting the wide range of reported estimates. Third, it identifies the limitations of prevailing empirical approaches, motivating the need for the alternative framework developed in this paper.

2.1 APG and the Selection Hypothesis

The study of sectoral productivity gaps is rooted in classic theories of structural transformation and growth (Lewis, 1954; Kuznets, 1971). A consistent empirical finding is that agricultural productivity lags far behind non-agricultural productivity in poor countries, with gaps far larger than those observed in rich economies (Gollin et al., 2002; Restuccia et al., 2008; Gollin et al., 2014; McMillan and Rodrik, 2014). Using national accounts for seventy-two countries, Gollin et al. (2014) show that the ratio of value added per worker (non-agriculture relative to agriculture) is about 1.3 in rich economies (10th percentile) but roughly 6.4 in poor ones (90th percentile), underscoring the steep income gradient in APGs. Because low-income countries employ a high share of workers in agriculture, this productivity disadvantage contributes directly to cross-country income inequality.

Early critics questioned whether such large gaps merely reflected measurement error in national accounts. Yet Gollin et al. (2014) show that, after adjusting for hours worked and human capital on the same sample, the gap remains substantial—ranging from about 1.0 to 4.3 across the distribution—indicating that measurement refinements are not first-order. Herrendorf and Schoellman (2018) further show, across forty-two censuses in thirteen countries over seven decades, that poor countries consistently exhibit large wage differences between sectors. Together, these studies confirm that APGs are real, persistent, and central to understanding income disparities, especially in poor countries.

The literature offers two leading explanations for large APGs. The first is misallocation (Restuccia and Rogerson, 2017; Gollin and Kaboski, 2023). It arises when policy and market frictions prevent resources and workers from moving to their most productive uses, keeping economies below their potential frontier (Restuccia et al., 2008). In agriculture, farm-level distortions and land-market failures compress farm size and weaken the link between inputs and output, leading to large productivity losses (Adamopoulos and Restuccia, 2014; Chen et al., 2023). Mobility and institutional barriers also hinder efficient allocations: credit

and risk constraints curb migration from rural areas, while social or policy institutions limit occupational mobility (Bryan et al., 2014; Munshi and Rosenzweig, 2016; Pulido and Świecki, 2019; Gai et al., 2021). Finally, differences in capital–labour substitutability across sectors point to underlying technology gaps, suggesting that agricultural productivity could rise substantially if such technologies diffused more widely (Alvarez-Cuadrado et al., 2017).

By contrast, selection highlights that individuals choose sectors based on comparative advantage. Building on Roy's (1951) framework, Lagakos and Waugh (2013) formalize this hypothesis by combining comparative advantage with subsistence food requirements: in low-productivity economies, many workers must remain in agriculture, creating large dispersion in farm productivity, whereas in high-productivity economies only highly skilled farmers remain, raising average productivity. When comparative and absolute advantage are positively correlated, sorting amplifies the APG. Later work notes that selection and misallocation likely coexist Lagakos (2020), and that the correlation between comparative and absolute advantage may be weak or even negative in poor countries (Alvarez-Cuadrado et al., 2020).

Thus, the selection hypothesis provides a compelling microfoundation for observed APGs; the remaining challenge is to measure its magnitude—an issue I take up next.

2.2 Empirical Analysis of Selection on APG

Following the theoretical foundation of the selection hypothesis proposed by Lagakos and Waugh (2013), a growing empirical literature has sought to quantify how individual sorting across sectors contributes to the Agricultural Productivity Gap (APG). Two main approaches have emerged in addressing the challenge of unobserved comparative advantage, which drives sectoral choice. One relies on panel fixed-effects estimators using sector switchers to control for time-invariant unobserved heterogeneity. The other assumes a specific distribution for unobserved sectoral abilities within a structural framework. Across studies, the estimated contribution of selection varies widely—from roughly one-third to over ninety percent—leaving unsettled debate on the relative importance of selection and misallocation.

Structural Approach

A group of studies—Lagakos and Waugh (2013); Pulido and Świecki (2019); Herrendorf

and Schoellman (2018); Alvarez (2020)—combine micro evidence with structural models to assess how selection and misallocation contribute to sectoral productivity gaps. These papers converge methodologically, using reduced-form estimates or panel moments as inputs for structural analysis rather than relying solely on macro data.

Some studies assume functional forms for unobserved abilities. Lagakos and Waugh (2013) assume dependent Fréchet-distributed sector-specific latent abilities, while Pulido and Świecki (2019) adopt a joint-normal distribution of unobserved heterogeneity. These assumptions allow analytical tractability and closed-form aggregation but inherently shape the magnitude of selection: the Fréchet distribution's heavy tails mechanically amplify selection effects, producing large apparent heterogeneity even when underlying ability dispersion is moderate. Moreover, as Heckman and Honore (1990) emphasize, there is little empirical evidence that unobserved abilities are normally distributed, raising concerns about the sensitivity of estimates to distributional assumptions.

Acknowledging the limitations of distributional assumptions, Alvarez (2020) uses Brazilian administrative and survey panels (1993–2016) to compare wage changes for sector switchers and multi-sector workers, finding modest within-individual wage gains—9 log points for manufacturing and 4 for services—relative to an average inter-sector gap of 48 log points. Herrendorf and Schoellman (2018) analyze 13 countries' censuses and complement them with panel evidence from the PSID (United States) and other studies for Brazil and Indonesia, concluding that observed mobility frictions are modest relative to ability-based sorting. Although these studies avoid imposing functional forms on latent abilities, they rely on fixed-effects estimates for sector switchers to inform their models. This reliance draws inference from a sub-population of workers with marginal gains relative to non-switchers, limiting the representativeness of estimated selection effects.

Reduced-Form Approaches

Beyond using fixed effects as a microfoundation for structural calibration, some studies apply similar methods directly to estimate selection, exploiting switchers identified in longitudinal microdata. Aiming at directly evaluating the impact of individual sorting on APG, Hamory et al. (2021) employ the Two-Way Fixed Effect (TWFE) estimator in panel data

without a macro structural model and interpret the productivity gap across sectors between with and without controlling for the fixed effect as the impact of individuals' sorting on the APG. They use long panels for Kenya and Indonesia and estimate an average earnings gap of roughly 70 log points between agriculture and non-agriculture. Controlling for individual fixed effects reduces this to about 22–24 log points, implying that 67–92 percent of the observed APG reflects individual sorting. This framework assumes that fixed effects fully capture unobserved heterogeneity relevant to sectoral choice. In practice, however, they absorb both general ability and sector-specific traits, yielding a local estimate for marginal movers rather than a population-level measure of comparative advantage.

Related studies extend this framework to migration contexts to recover the selection effect on APG through rural-urban movers. Lagakos et al. (2020), using panel data from six developing countries, find that once worker fixed effects are included, estimated migration costs decline substantially and average returns to migration are around 23% in five of the six countries—suggesting limited misallocation. By contrast, Gai et al. (2021) exploit a natural policy intervention in rural China as an instrument and frame their analysis as an event study. They estimate the local treatment effect (LATE) for the agriculture-to-nonagriculture switchers affected by the policy intervention and find that those sector switchers face migration costs equivalent to approximately 55% of their earnings in the nonagricultural sector. Using a control function approach, they estimate an average treatment effect (ATE) on sectoral earnings in rural China during the study period of 2003-2012 and recover an underlying APG of 46 log points once the selection on unobserved sector-specific abilities is accounted for. Given that the observed APG in their data is 68 log points, they conclude that approximately 32% of the sectoral productivity gaps are explained by individuals' selection into sectors based on their latent skills.

Together, these results contribute to the ongoing debates on the relative roles of selection and misallocation.

Synthesis and limitations

Overall, the literature demonstrates that worker sorting contributes meaningfully to the APG, yet its precise magnitude remains uncertain. In terms of how unobserved comparative

advantage is identified, the two prevailing approaches are (a) panel fixed-effects estimation relying on switchers, and (b) structural modeling based on assumed distributions of latent abilities.

Despite valuable insights, these approaches face consequential limitations in addressing two core challenges when estimating the selection effect on APG: (1) measuring sector-specific heterogeneous abilities; and (2) disentangling endogeneity induced by unobserved comparative advantages. The panel approach that controls for individual fixed effects does not allow an appropriate formulation of comparative advantage, because it treats unobserved heterogeneity as an individual-specific component rather than distinguishing sector-specific abilities for each person. In addition, this approach ignores non-switchers and risks attributing all unobserved abilities to sectoral choice—including traits unrelated to sector selection, such as general ability or work ethic—thereby overstating the selection effect. By contrast, the structural approach based on the generalized Roy model estimates selection by assuming a specific distribution for unobserved sector-specific abilities (e.g., joint normal or Fréchet). However, these functional assumptions lack empirical support and can materially influence the estimated effects.

These limitations leave the magnitude of the selection effect highly sensitive to methodological choices. In practice, existing studies report wide-ranging estimates—significant but inconsistent in size—precisely because they conflate individual-level and lean on restrictive assumptions to handle heterogeneity and endogeneity. To address these challenges, this paper adopts an alternative framework that models sector-specific heterogeneity as individual deviation from sector-average productivity, while avoiding parametric distributional assumptions about unobserved heterogeneity. Specifically, I adapt the correlated random-coefficients (CRC) framework of Suri (2011), which avoids strong parametric restrictions and better isolates the sector-relevant component of individual abilities.

These limitations leave the magnitude of the selection effect highly sensitive to methodological choices. In practice, existing studies report wide-ranging estimates—significant but inconsistent in size—because empirical strategies often conflate sector-specific comparative advantage with general latent traits that do not govern sector choice, and they lean on restrictive assumptions to handle heterogeneity and endogeneity. To address these challenges,

this paper adopts an alternative framework that models sector-specific heterogeneity as individual deviation from sector-average productivity while avoiding parametric distributional assumptions about unobserved heterogeneity. Specifically, I adapt the correlated random-coefficients (CRC) framework of Suri (2011), in the distribution-free spirit of Lemieux (1998), which avoids strong parametric restrictions and better isolates the sector-relevant component of individual abilities. The next section details the implementation and identification of each component.

Appendix A provides a fuller description of Suri's original CRC framework—and its application to hybrid seed adoption in Kenya—for readers seeking additional background before turning to this paper's adaptation to the APG setting.

3 Empirical Model

This section outlines the empirical model of this paper by extending the approach in Suri (2011) to the APG literature. I begin by modelling individual earnings by using the Mincerian representation for human capital. Building on this foundation, I follow the method in Lemieux (1998) and Suri (2011) to incorporate heterogeneous latent skills across sectors into this model, where absolute and comparative advantages are defined. Next, I put together the main empirical model that enables the measurement of how self-selection affects individual earnings. Finally, I illustrate how to draw the impact of individual selection on the sectoral productivity gap through aggregation.

To anchor the selection effect of my model, I further map it to the types of selection proposed by Borjas (1987) in the classic Roy's model framework and discuss the selection effect in the adopted model in Appendix B.

3.1 Model Setup

As individual sorting is based on comparing the potential earnings in each sector, the starting point is to model how the choice of sector affects an individual's potential earnings. In an economy, an individual i at time t can choose to work in one of two sectors: $j \in \{n, a\}$, where n refers to the non-agricultural sector and a to the agricultural sector. An individual's

potential earnings in each sector at each period are determined by the sector productivity, A_t^j, and her own human capital, h_{it}^j, as expressed in equations (1) and (2). In these two equations, P_t^j represents the price level at each sector j.

$$W_{it}^n = P_t^n A_t^n h_{it}^n \tag{1}$$

$$W_{it}^a = P_t^a A_t^a h_{it}^a \tag{2}$$

Following Mincer, the human capital h_{it}^j can be expressed as an exponential function of observed and unobserved characteristics, see equations (3) and (4), where X_{it} is a vector of observed characteristics endowed by individual i at time t, and U_{it}^j is a vector of individual i's unobservable in sector j at time t.

$$h_{it}^n = exp(X_{it}\gamma^n + U_{it}^n) \tag{3}$$

$$h_{it}^a = exp(X_{it}\gamma^a + U_{it}^a) \tag{4}$$

Substitute equations (3) and (4) into (1) and (2), the individual potential earnings in each sector j and time t can be represented as equations (5) and (6).

$$W_{it}^n = P_t^n A_t^n exp(X_{it}\gamma^n + U_{it}^n) \tag{5}$$

$$W_{it}^a = P_t^a A_t^a exp(X_{it}\gamma^a + U_{it}^a) \tag{6}$$

Taking logs, I can obtain equations (7) and (8), where the lower cases represent the logarithmic terms.

$$w_{it}^n = \underbrace{p_t^n + a_t^n}_{=\delta_t^n} + X_{it}\gamma^n + U_{it}^n \tag{7}$$

$$w_{it}^a = \underbrace{p_t^a + a_t^a}_{=\delta_t^a} + X_{it}\gamma^a + U_{it}^a \tag{8}$$

In equations (7) and (8), p_t^j represents the price for each sector j at time t, and a_t^j stands for sectoral average productivity at time t. Together, they represent the returns from the sector-wide productivity at time t for sector j, denoted as δ_t^j. Hence, rewrite potential earnings for each sector as equations (9) and (10).

$$w_{it}^n = \delta_t^n + X_{it}\gamma^n + U_{it}^n \tag{9}$$

$$w_{it}^a = \delta_t^a + X_{it}\gamma^a + U_{it}^a \tag{10}$$

3.2 Absolute and Comparative Advantages

Comparative advantage is the difference in an individual's abilities between sectors. It is the differential returns based on this comparative advantage that affect sector decisions for economic agents. This comparative advantage comprises two components: observed characteristics (e.g. education), collected in the vector of X's, and unobserved traits–such as spatial intuition for arranging crops versus verbal dexterity in persuading clients–that differentially raise payoffs across sectors. Individuals know these latent abilities and internalize them when choosing sectors, while the econometrician does not as they are rarely recorded in data. As a result, ignoring this unobserved comparative advantage leads to selection bias in estimates of sectoral returns.

Therefore, I will introduce more structure to the unobserved error terms, U_{it}^j, in equations (9) and (10), reflecting the source of selection within the unobserved ability terms U_{it}^j. Following Lemieux (1998) and Suri (2011), decompose as in equation (11) and (12).

$$U_{it}^n = \theta_i^n + \xi_{it}^n \tag{11}$$

$$U_{it}^a = \theta_i^a + \xi_{it}^a \tag{12}$$

where θ_i^j is a time-invariant, sector-specific ability (permanent) and ξ_{it}^j is an idiosyncratic shock, which is unknown at choice, but follows a zero conditional mean (transitory). Hence, sorting depends on the permanent vector (θ_i^n, θ_i^a), not on ξ_{it}^j, shown in equation (13):

$$E(U_{it}^n - U_{it}^a) = \theta_i^n - \theta_i^a \tag{13}$$

Simply put, this transitory component of the error terms does not affect individuals' sector choices. Thus, when choosing a sector, individuals compare the difference of their expected potential earnings in each sectors (w_{it}^n vs. w_{it}^a), which essentially compare the expected difference in the sector-wide productivity ($\delta_t^n - \delta_t^a$), rewards to the observed characteristics ($X's$), and the individuals' sector-specific latent abilities ($\theta_i^n - \theta_i^a$). To separate abilities that influence sector choice from those that do not, linearly project absolute advantage in each sector (θ_i^j) onto the "difference of latent abilities across sector" ($\theta_i^n - \theta_i^a$), shown as equations

(14) and (15):

$$\theta_i^n = b_n(\theta_i^n - \theta_i^a) + \tau_i \tag{14}$$

$$\theta_i^a = b_a(\theta_i^n - \theta_i^a) + \tau_i \tag{15}$$

where τ_i is a common component (orthogonal to $\theta_i^n - \theta_i^a$) that moves productivity in both sectors equally (e.g., work ethic) and thus does *not* affect sector choice. The coefficients (b_n, b_a) are projection coefficients determined by the variance–covariance matrix of (θ_i^n, θ_i^a).[1]

Define the individual's *comparative advantage* as

$$\theta_i \equiv b_a(\theta_i^n - \theta_i^a), \tag{16}$$

and the *selection effect* as [2]

$$\beta \equiv \frac{b_n}{b_a} - 1. \tag{17}$$

Then individual sector-specific absolute advantages can be expressed as a function of comparative advantage and selection effect:

$$\theta_i^n = (1+\beta)\theta_i + \tau_i \tag{18}$$

$$\theta_i^a = \theta_i + \tau_i \tag{19}$$

Equations (18) and (19) are the key decompositions: (θ_i, β) describe the part of unobserved ability that *drives sorting* (comparative advantage and its sectoral loading), while τ_i is the sector-irrelevant component. This is precisely where TWFE regressions confound selection: they absorb *both* θ_i and τ_i into a single fixed effect, attributing common, sector-irrelevant traits to selection, such as hard work.

Hence, conditional on observables, sector choice is governed by the differential return to θ_i (the comparative advantage component), with loading $(1+\beta)$ in non-agriculture and 1 in agriculture. The common component τ_i is irrelevant for sorting.

[1] Closed-form expressions: $b_n = \frac{\sigma_n^2 - \sigma_{na}}{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}$ and $b_a = \frac{\sigma_{na} - \sigma_a^2}{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}$, where $\sigma_n = Var(\theta_i^n)$, $\sigma_a = Var(\theta_i^a)$, and $\sigma_{na} = COV(\theta_i^n, \theta_i^a)$.

[2] Later in this subsection, I'll further discuss on why β is selection effect.

In this formulation, the structural parameter β summarizes how strongly unobserved comparative advantage is rewarded *differentially* across sectors after netting out the correlation between sectoral abilities. Positive β indicates that the nonagricultural sector loads more on the relevant ability dispersion than agriculture; negative β indicates the opposite.

Intuitively, the parameter β measures how strongly differences in unobserved abilities are rewarded across sectors. When $\beta > 0$, variation in unobserved comparative advantage generates larger payoffs in non-agriculture than in agriculture. In this case, differences in individual talent translate more strongly into earnings outside agriculture, so workers with higher relative ability in non-agriculture have stronger incentives to sort into that sector, amplifying the observed productivity gap. When β is close to zero, the two sectors value differences in ability similarly, and sorting contributes little to the gap.

In the Appendix B, I formally map the selection effect β to types of selection in the classic Roy model. When $\beta > 0$, individuals are positively selected—drawn from the upper tail in agriculture and landing in the upper tail of non-agriculture. When $-1 < \beta < 0$, the sorting is negative, as workers come from the lower tail in both sectors. When $\beta < -1$, a 'refugee' case arises: workers below average in agriculture sort into non-agriculture but earn above its mean. Full algebra, the formal conditions, and the side-by-side comparison with Borjas's are in Appendix C.

Essentially, the selection effect and unobserved comparative advantages formulated by Lemieux (1998) capture the equivalent selection effect from unobserved heterogeneity that affects individuals' choices, as modelled in the classic Roy choice framework. The difference is that this formulation allows for the flexibility to estimate underlying unobserved comparative advantages avoiding distributional assumption.

3.3 Main Empirical Model

After the discussion on β, let's go back to the empirical model used in this paper. Building on equations (18)-(19), which decompose latent ability into comparative advantage and selection effect, I substitute into the potential earnings framework (equations (9) and (10)). I can rewrite the individual's log potential earnings at time t for each sector j in equations (20)

and (21).

$$w_{it}^{n} = \delta_{t}^{n} + (1+\beta)\theta_{i} + \tau_{i} + X_{it}\gamma^{n} + \xi_{it}^{n} \tag{20}$$

$$w_{it}^{a} = \delta_{t}^{a} + \theta_{i} + \tau_{i} + X_{it}\gamma^{a} + \xi_{it}^{a} \tag{21}$$

Let D_{it} be a dummy variable, taking the value one if an individual i chooses the primary job in the nonagricultural sector at time t and zero otherwise. Then, I can write the individual's observed log earnings as equation (22).

$$w_{it} = D_{it}w_{it}^{n} + (1 - D_{it})w_{it}^{a} \tag{22}$$

where

$$D_{it} = \begin{cases} 1 & \text{non-agricultural sector} \\ 0 & \text{agricultural sector} \end{cases}$$

An individual chooses non-agriculture if she compares expected potential earnings in both sectors and finds that she would earn more in the non-agricultural sector, i.e., $w_{it}^{n} > w_{it}^{a}$; otherwise, she selects an agricultural job.

Then, substitute equations (20) and (21) in (22) to obtain equation (23).

$$\begin{aligned} w_{it} = {} & \delta_{t}^{a} + (\delta_{t}^{n} - \delta_{t}^{a})D_{it} \\ & + \theta_{i} + \beta\theta_{i}D_{it} + X_{it}\gamma^{a} + X_{it}(\gamma^{n} - \gamma^{a})D_{it} + \tau_{i} + \epsilon_{it} \end{aligned} \tag{23}$$

where

$$\epsilon_{it} = D_{it}\xi_{it}^{n} + (1 - D_{it})\xi_{it}^{a} \tag{24}$$

Equation (23) is the main empirical model in this paper. I am interested in estimating the parameter β, the selection effect of comparative advantage, and recovering the distribution of comparative advantages θ_i. The fourth term in equation (23), $\beta\theta_i D_{it}$, contains the unobserved random variable θ_i, which is correlated with sectoral choice D_{it}; hence, this is a correlated random-coefficients (CRC) model. In this formulation, individual comparative advantage (θ_i) is explicitly defined as the individual's deviation from average sector productivity. If an individual works in the agricultural sector, sector-wide productivity is δ_t^a, and θ_i expresses how much more productive each individual is compared to the sector mean. If the

non-agricultural sector is chosen, sector-wide productivity is represented by $\delta_t^a + (\delta_t^n - \delta_t^a)$, and the individual deviation from the sector mean is $(1 + \beta)\theta_i$.

This empirical framework models comparative advantage as the difference between absolute advantages across sectors, scaled by the covariance-adjusted spread, while distinguishing which unobserved abilities are relevant to sector choices. It directly addresses the two core challenges identified in the Literature Review—heterogeneity and endogeneity. Moreover, this formulation, consistent with the Roy model of sectoral choice, is better equipped to estimate individual sorting than a TWFE estimator or models that impose fixed effects, as reviewed previously, and allows for the estimation of selection effects without functional assumptions on latent abilities.

In a fixed-effects estimator using panel data, individual heterogeneity is absorbed at the person level rather than at the person-sector level, effectively imposing $\theta_i^n = \theta_i^a$ and thus $\beta = 0$. In other words, although the TWFE estimator corrects for time-invariant unobserved heterogeneity across individuals, it assumes that each individual's latent ability is identical across sectors and therefore cannot identify differential returns to unobserved comparative advantage. Moreover, by uniformly absorbing both the general ability component (τ_i)—which is irrelevant to sector choice—and the comparative advantage component (θ_i)—which is central to sorting—the fixed-effects estimator conflates the two and removes the very variation needed to identify β.

Structural approaches, by contrast, identify β by imposing a joint distribution on sector-specific abilities, e.g., (θ_i^a, θ_i^n) following a Normal or Fréchet distribution. In this setup, the selection effect arises from the assumed shape and correlation structure of that joint distribution: the dependence between sectors, and the relative spread of abilities, determines how much unobserved heterogeneity translates into differential payoffs across sectors. Appendix B illustrates this relationship formally by mapping the CRC selection term to the classical Roy model. Consequently, the estimate of β in this type of structural settings inherits these functional-form restrictions, rather than completely emerging from the observed variation in sectoral choices.

By contrast, following Lemieux (1998), the CRC model adapted in this paper relaxes those distributional assumptions by linearly projecting absolute advantage onto the latent

abilities that determine sector choice (equations (18) and (19)). It retains the identity of the selection term β consistent with the classical Roy framework,[3] but achieves it through linear projection, thereby removing any functional-form requirement on θ_i. This relaxation makes it possible to use information contained in observed choice trajectories to infer unobserved comparative advantage θ_i while simultaneously identifying its impact, β.

In Section 5, I show how these choice trajectories reveal θ_i empirically and demonstrate how this framework recovers the distribution of comparative advantage without parametric assumptions. Before that, the next subsection discusses how the selection effect β contributes to the Agricultural Productivity Gap (APG).

3.4 Individual Sorting and Agricultural Productivity Gap (APG)

The objective of the main empirical model, as described in equation (23), is to estimate the structural parameter, the selection effect β, without imposing any distributional assumptions on latent skills. In the estimation section, I will describe how to obtain β without distributional assumptions. Now, suppose β is estimated from the data. What β measures is the extent to which individual sorting based on comparative advantages affects their earnings. Therefore, an aggregation is necessary to provide an answer to the research question posed in this paper: how much individual sorting affects sectoral productivity gaps?

In this two-sector economy (agriculture and non-agriculture), assume that both of which are perfectly competitive. The output in each sector is given by equations (25) and (26).

$$Y_n = A_n H_n \tag{25}$$

$$Y_a = A_a H_a \tag{26}$$

where Y_j represents the sector aggregate output, A_j is the sector-specific efficiency, H_j is the efficient labour in each sector. Furthermore, the efficient labour H_j is the product of sector-specific human capital, h_j, and the total number of workers L_j in each sector j, represented

[3] Appendix B provides the full mapping of β to the coefficient of the selection-correction term (the Inverse Mills Ratio) in a Roy model with a joint-normal distribution.

by equations (27) and (28).

$$H_n = h_n L_n \tag{27}$$

$$H_a = h_a L_a \tag{28}$$

Following the agricultural productivity gap (APG) defined by Gollin et al. (2014), under the assumption of a perfectly competitive labour and goods market in both sectors, the wage in each sector equals the marginal value product of labour, and it also equals the average value of output per labour in each sector at equilibrium. Let W_j be the wage in the sector j, and P_j be the price of good j. In equilibrium, APG–the ratio of value-added (VA) per worker between sectors j—can be expressed as the wage ratio at the sector level. $\frac{W_n}{W_a}$. See equation (29).

$$\underbrace{\frac{P_n Y_n / L_n}{P_a Y_a / L_a}}_{= \frac{VA_n/L_n}{VA_a/L_a} \equiv APG} = \frac{W_n}{W_a} \tag{29}$$

In the data, the difference in log earnings between sectors is APG, expressed in equation (30). This APG can be directly calculated by differencing the mean log earnings between two sectors, using the data.

$$log\left(\frac{W_n}{W_a}\right) = log(W_n) - log(W_a) = APG \tag{30}$$

Mathematically, take an expectation of observed log wage w_{it} for each group, using the main empirical model expressed in equation (23), and then plug them in APG equation (30). I can further rewrite APG as equations (31) and (32). [4]

$$APG = \underbrace{\overline{X}(\hat{\gamma^n} - \hat{\gamma^a})}_{APG_{observed}} + \underbrace{\delta^n - \delta^a}_{APG_\delta} + S_\theta \tag{31}$$

$$S_\theta = \underbrace{\beta E[\theta_i | D = 1]}_{\text{extra returns in nonag}} + \underbrace{(E[\theta_i | D = 1] - E[\theta_i | D = 0])}_{\text{mean diff in comparative advantages}} \tag{32}$$

As shown in equation (31), APG comprises three components at the sector level: (1) the gap from observed characteristics, $\overline{X}(\hat{\gamma^n} - \hat{\gamma^a})$; (2) the disparity from sector-wide productivity

[4] Detailed step-by-step derivations for equations (31) and (32) are in Appendix E.

gap, $\delta^n - \delta^a$; (3) the difference from individual sorting based on the unobserved comparative advantages, denoted as S_θ.

Because θ_i is unobserved, its mean cannot be calculated directly. Equation (33) offers an alternative route to recover the selection component of the APG. As a result, the impact of individual sorting based on comparative advantages is the share of APG from the unobserved component, $\frac{S_\theta}{APG}$.

$$S_\theta = APG - APG_{observed} - APG_\delta \tag{33}$$

To recap, the main empirical model in this section explicitly models individual unobserved comparative advantage θ_i as a deviation from the average productivity for each sector, addressing unobserved heterogeneity. Moreover, it formulates the absolute and comparative advantages to allow for the estimation of the selection effect β without imposing a functional form on latent abilities. In addition, the types of selection in this main model can be mapped to the selection types in the classic Roy model framework.

I aim first to recover the structural parameters β and θ_i, and then aggregate to assess the importance of individual selection based on the observed APG. Before getting into estimation strategies, I will first discuss the identification of this CRC model.

4 Identification

The identification of this CRC model is hinged on two key elements: First, the composite error terms $\tau_i + \epsilon_{it}$ is strictly exogenous. Second, the choice trajectory reveals unobserved comparative advantage θ_i.

4.1 A Strict Exogeneity Assumption

A strict exogeneity assumption of the composite error term, expressed in equation (34), delivers the identification for the main estimation equation (23).

$$E(\tau_i + \epsilon_{it} | \theta_i, D_{i1} ... D_{iT}, X_{i1} ... X_{iT}) = 0 \tag{34}$$

As τ_i represents individual i's unobserved abilities, regardless of sector choices, this strict exogeneity assumption is not overly restrictive for τ_i. Mathematically, equations (18)

and (19) indicate that τ_i is removed from the comparative advantages, θ_i; hence, τ_i does not affect the sectoral choices and other observed regressor in (23).

The main concern arises from the transitory error term ϵ_{it}. While τ_i is time-invariant and unrelated to sectoral choice, ϵ_{it} could be correlated with D_{it} if short-term shocks (e.g., crop failure, weather events) influence both earnings and the decision to switch sectors. However, such shocks are typically transitory and do not justify abandoning a productivity-maximizing sector. Moreover, switching sectors often requires retraining or relocation, making it implausible as a short-run adjustment. Thus, individuals primarily adjust hours worked rather than change sectors in response to transitory shocks. As long as these temporary adjustments are controlled for, the strict exogeneity assumption remains credible.

However, some transitory shocks could impact individuals' realized earnings, such as crop failure due to extreme weather conditions, which could raise concerns about the identification of the model. This concern arises because individuals are likely to take measures to maintain their normal income level during adverse shocks, such as switching sectors or increasing the number of hours worked. Notably, individuals selected for their primary sector based on comparative advantage are unlikely to respond to short-lived, sector-specific shocks by changing sectors. This tendency to remain in the original sector is plausible for two reasons. First, the shocks are transitory and do not justify abandoning a productivity-maximizing allocation. Second, without retraining or skill upgrading, individuals are unlikely to achieve higher earnings in an alternative sector that lacks a comparative advantage.

If individuals choose to work more hours in response to temporary shocks, that could raise concerns about identification. This is because temporary shocks contributes both adjustment on working hours and individual's earnings. To make the strict exogeneity assumption hold, such temporary shocks needs to be controlled for. The next section will further examine if this is possible in the data.

Hence, in this setting, the exogenous shocks that cause individuals to switch sectors must be cost-related, such as making the pursuit of the alternative sector cheaper. For instance, as more factories are established in villages, the cost of securing a stable wage in the manufacturing sector is substantially reduced, and some people may be persuaded to switch sectors.

4.2 Revealed Comparative Advantage

If the strict exogeneity assumption on the composite error term holds, I can follow the projection approach by Chamberlain (1982, 1984) and Suri (2011) to disentangle the correlation between unobserved comparative advantage θ_i and choice variable D_{it} in the fourth term $\beta\theta_i D_{it}$ in the main estimation equation (23). This step allows a reduced-form regression where the error terms follows conditional mean zero and the coefficients are the functions of the structural parameters. Then, using a minimium distance estimator (MDE), I can identify selection effect β and recover the distribution of θ_i.

In the subsection 3.2, I have discussed that the comparative advantage formulation by Lemieux (1998) only relies on linear projection and relaxes the distributional assumption on θ_i. Here, I further discuss how to obtain unobserved θ_i, which is free of a specific distributional form. In this setting, individuals make decisions by comparing expected potential earnings between two sectors, and the unobserved comparative advantage θ_i contributes to the differential rewards in their earnings. Hence, θ_i is correlated to individual's choices D_{it}. Intuitively, individuals tend to stay in the sector in which their comparative advantages are stronger.

To separate θ_i and D_{it}, Chamberlain (1982, 1984) linearly project time-invariant θ_i on choice variable D_{it}, taking advantage that choice is made in each period. Further, Suri (2011) proposes adding the interactions to purge this correlation entirely. To operationalize the model, I follow Chamberlain (1982, 1984) and project the unobserved θ_i onto observed choice histories D_{it}, exploiting within-individual variation across period. Equation (35) shows this step in a three-period panel.

$$\begin{aligned}\theta_i = \lambda_0 &+ \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1} D_{i2} \\ &+ \lambda_5 D_{i1} D_{i3} + \lambda_6 D_{i2} D_{i3} + \lambda_7 D_{i1} D_{i2} D_{i3} + \nu_i \end{aligned} \tag{35}$$

As choice variable D_{it} is a dummy, this linear projection is a saturated model. Both Chamberlain and Suri call this step as a purely technical procedure without any economic interpretation. This is because once I substitute Equation (35) in (23), the rearranged main estimation equation does not have θ_i, instead it is replaced by the function of choice variable D_{it} and their interaction terms. This substitution step allows two things: First, it allows to

perform a reduced-form regression as each regressor is observed in the data. Second, it links the coefficients of the reduced-form to the underlying structural parameters.[5]

Further on Equation (35), I propose an economic interpretation. Each period, individuals internalize their unobserved comparative advantages in order to decide on which sector is better rewarded. Hence, the choice histories reveal individuals' latent abilities. In the data, sector choices are observed, but comparative advantages θ_i are not. In my view, Equation (35) can be also a way to estimate unobserved comparative advantages by exploiting individuals' sectoral choice trajectories. So, I call estimated unobserved comparative advantage, $\hat{\theta}_i$, as revealed comparative advantage. Appendix G further discusses this reveal comparative advantage concept and shows that choice trajectories contain such information in the data.

Together, strict exogeneity and the revealed comparative advantage substitution identify the adapted CRC model. This approach allows estimation of the selection effect without assuming any specific distribution for latent abilities—unlike the Roy-type model, which hinges on joint normality. This revealed comparative advantage interpretation is what connects the structural framework to observable data, adding an economic meaning to a purely technical procedure.

Relative to the existing literature, this framework contributes by (i) redefining absolute advantage in line with the APG literature, (ii) linking the selection effect β to Borjas's (1987) Roy-model selection types, (iii) introducing the notion of revealed comparative advantage to add economic interpretation to the projection method, and (iv) decomposing the APG into observed, productivity, and sorting components.

The next section applies this empirical model to data from the first three waves of the IFLS.

5 Estimate Selection on APG

The previous section described the empirical framework based on Suri (2011), extended to quantify the role of individual sorting in explaining sectoral productivity gaps. This section

[5] Appendix F illustrates this mapping in a simple two-period case, showing explicitly how reduced-form coefficients recover θ_i, mathematically.

implements the model using the first three waves of the Indonesia Family Life Survey (IFLS), covering 1993–2000.[6]

I first describe the data and present descriptive statistics for the first three IFLS waves. Next, I estimate the reduced-form equations using Seemingly Unrelated Regressions (SUR) in the first stage, followed by recovering the structural parameters of interest—the selection effect β and the distribution of comparative advantage θ_i—in the second stage. Both stages are implemented using **randcoef**. Finally, I assess the contribution of individual sorting to the agricultural productivity gap (APG).

5.1 Data

This study draws on two data sources: the Indonesian Family Life Survey (IFLS) and provincial consumer price indices (CPI) from Statistics Indonesia (Badan Pusat Statistik, BPS). The IFLS is a rich longitudinal household survey spanning five waves (1993, 1997, 2000, 2007, and 2014). The baseline covers 13 of the then–26 provinces, selected to represent roughly 83% of Indonesia's population and to capture broad socioeconomic and geographic diversity (Strauss et al., 2016). Over this period, Indonesia transitioned from a low-income to a lower-middle-income country. According to World Bank estimates (2025), real GDP per capita (constant 2015 USD) rose from $1,693 in 1993 to $3,171 in 2014, passing through the Asian Financial Crisis and entering a phase of sustained post-crisis recovery and growth. The IFLS provides detailed, repeated observations on individuals' employment, sectoral choices, earnings, and household characteristics, making it well suited to study how individual sorting contributes to sectoral productivity gaps.

During the IFLS period (1993–2014), Indonesia underwent major institutional and political changes. The authoritarian New Order regime collapsed with the ouster of President Suharto in May 1998, initiating the *Reformasi* transition to democratic governance (Freedom House, 1998). Subsequently, Parliament passed Laws 22/1999 and 25/1999, mandating a sweeping transfer of authority and revenues to regional governments, effective 2001—a shift

[6]The estimation uses the **randcoef** package in Stata, developed by Cabanillas, Michler, Michuda, and Tjernström (2018). The **tuple** package must be installed beforehand, both available via the Stata Community-Contributed programs repository.

known as the "Big Bang" decentralization (Hofman and Kaiser, 2002). Given the magnitude of this institutional break and its economic implications, this project naturally divides the IFLS data into two periods: 1993–2000 and 2000–2014. The current paper focuses on the first period, while a companion study will examine the later waves.

To account for spatial and temporal price variation, I incorporate province–year CPI data from Statistics Indonesia (BPS) as a control in all earnings regressions. Figure 1 plots the log CPI across the 16 provinces represented in the sample. The original IFLS baseline covered 13 provinces, but a small number of respondents moved to new provinces in later waves, producing limited observations for those areas (e.g., Riau, East Kalimantan, and Central Kalimantan). Overall, provincial price levels rose steadily between 1993 and 2000, with particularly sharp increases in Jakarta, West Java, and West Sumatra. Because both earnings and CPI enter the regression in logarithmic form, the specification effectively adjusts for inflation and spatial price differences without explicitly deflating nominal earnings.

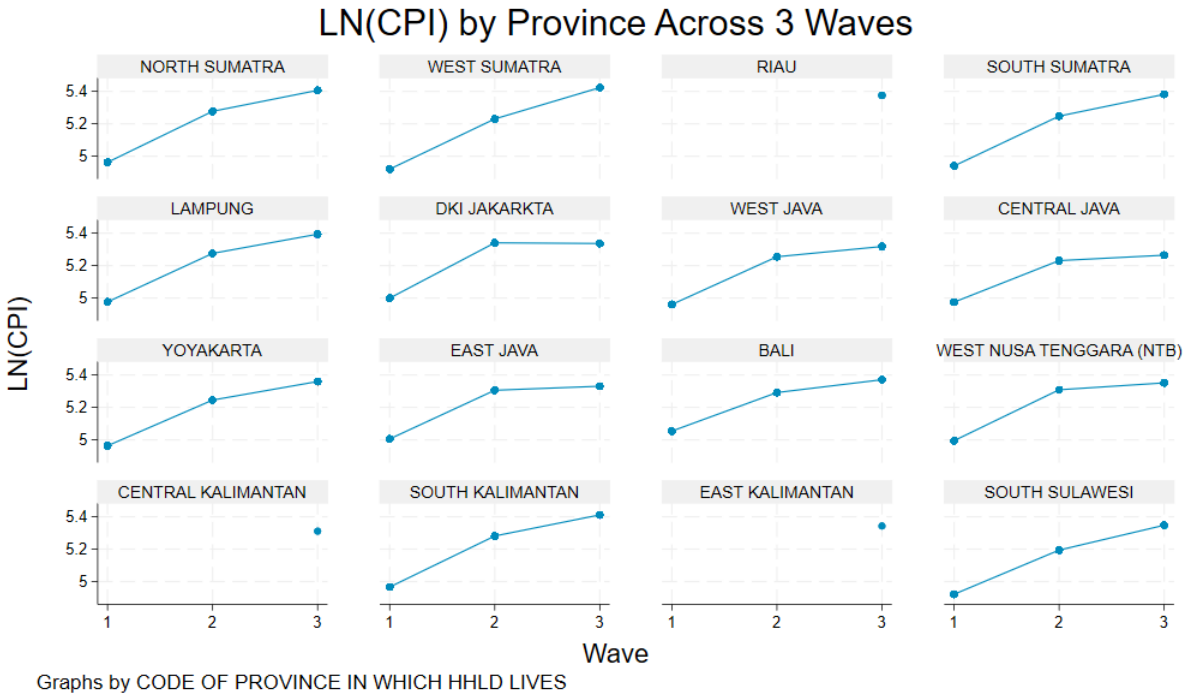

Figure 1: Log CPI

5.2 Descriptive Statistics

Because the empirical model requires observing both earnings and sector choice in every wave, I keep only individuals with complete information in all periods. The resulting panel contains 4,615 individuals followed over three waves, totalling 13,845 person-wave observations. These individuals come from 3,963 households in wave 1, 3,991 in wave 2, and 4,012 in wave 3; the small increase reflects household splits over time. About 12% of sampled individuals share a household with another sampled adult (i.e., multiple earners from the same household), and roughly 78% of the individuals are household heads.

Table 3 summarizes key characteristics (means/proportions) for the restricted and unrestricted samples. The restricted sample is systematically different, with more males (73% vs. 54%), slightly older respondents (mean age 44 vs. 40), and a higher marriage rate (90.8% vs. 75.9%).

This gender imbalance does not accurately reflect the gender composition of the full IFLS sample, but instead arises from sample restrictions: many women in the broader population work without pay in family enterprises or are out of the labour force as homemakers, and thus are excluded from the earnings-based analysis.

These restrictions arise mechanically from the need to observe both sectoral employment and earnings, rather than from survey attrition. In contrast, the sample remains broadly representative along key economic dimensions—education, average log earnings, average log hours worked, rural–urban composition, and the agricultural employment share are nearly identical to those in the unrestricted sample. While the restricted sample is not fully representative of the broader IFLS population—particularly along demographic dimensions—it is well suited for examining within-individual sectoral sorting, since it ensures consistent earnings and sector observations across all periods.

Tables 4 and 5 report summary statistics for the balanced analytical sample at the individual level. On average, individuals in the balanced panel were 40.4 years old in 1993, with mean age increasing steadily across waves as expected. Roughly 44% of the people resided in urban areas, a proportion that remained stable across survey rounds. This stability likely reflects the nature of the balanced sample, which includes only individuals with complete

earnings and sector information in all three waves.

The share of individuals in non-agricultural work is relatively stable, declining slightly from 65.5% in 1993 to 63.0% in 2000. Approximately 45–40% report waged employment across waves,[7] implying a gradual rise in informality—from about 54% in 1993 to 60% in 2000. In this paper, sector is defined by the individual's primary job in each wave; both wage and self-employment (informal) are included within agriculture and non-agriculture. Individuals reporting activities in both sectors are classified by the sector of their main job. Notably, non-agriculture has a much higher share of waged employment (54%) than agriculture (24%), and both sectors exhibit a decline in wage employment over time. Average monthly nominal income rose substantially—from IDR 126,195 in 1993 to IDR 432,583 in 2000—reflecting both real income growth and high inflation around the Asian financial crisis. This is corroborated by the rise in the Consumer Price Index (CPI), which climbed from 145.2 to 209.4 over the same period. Meanwhile, the average number of hours worked per month remained stable, hovering around 174 hours, with a slight decrease in working hours over time.

Table 6 reports household-level characteristics for the same balanced sample. Average household size increased modestly over time, from 4.78 members in 1993 to 5.80 in 2000. Family business ownership rose sharply—from 68% in the first wave to 80% by the third period. The share of households operating farm businesses remained relatively stable (rising from about 43% to 48%), while the share engaged in non-farm businesses increased significantly, from 37% to 52% over the three periods. Nominal asset values in both categories rose over time: farm business assets grew from IDR 7.1 million in 1993 to IDR 26.6 million in 2000, while non-farm business assets increased from IDR 6.4 million to IDR 9.5 million. Household assets not tied to any business—mainly real estate—were substantially higher, increasing from IDR 11.4 million to IDR 32.1 million. This comparison with property values suggests that most family businesses are relatively small in scale compared with overall household wealth holdings.

The IFLS also provides detailed information on economic shocks experienced by households in the five years preceding each survey. In 1993, 31.2% of households reported at least

[7] In this paper, I use 'wage' ≈ formal and 'non-wage' ≈ informal

one shock, with an average of 0.39 shocks per household. In 1997, during the onset of the Asian financial crisis, 40% of households reported at least one shock, with an average of 0.57 shocks. By 2000, both the incidence (35%) and intensity (0.44 shocks on average) of reported shocks had declined.

The descriptive figures further illuminate income dynamics and sectoral choices in the dataset. Throughout, sector is defined by the individual's primary job in each wave; secondary activities are not used because those data are sparse and inconsistent. Figure 2 displays the distribution of monthly income (log) by wave and sector. Across all three rounds, individuals in non-agriculture earn more on average than those in agriculture. In agriculture, outliers cluster on the low side each wave (including a handful near zero), while high-side outliers are rare. In non-agriculture, low-side outliers are relatively few, but a recurrent upper tail of high earners protrudes above the upper whisker. This pattern aligns with Tables 4–5, where within-sector earnings dispersion is large in each wave.

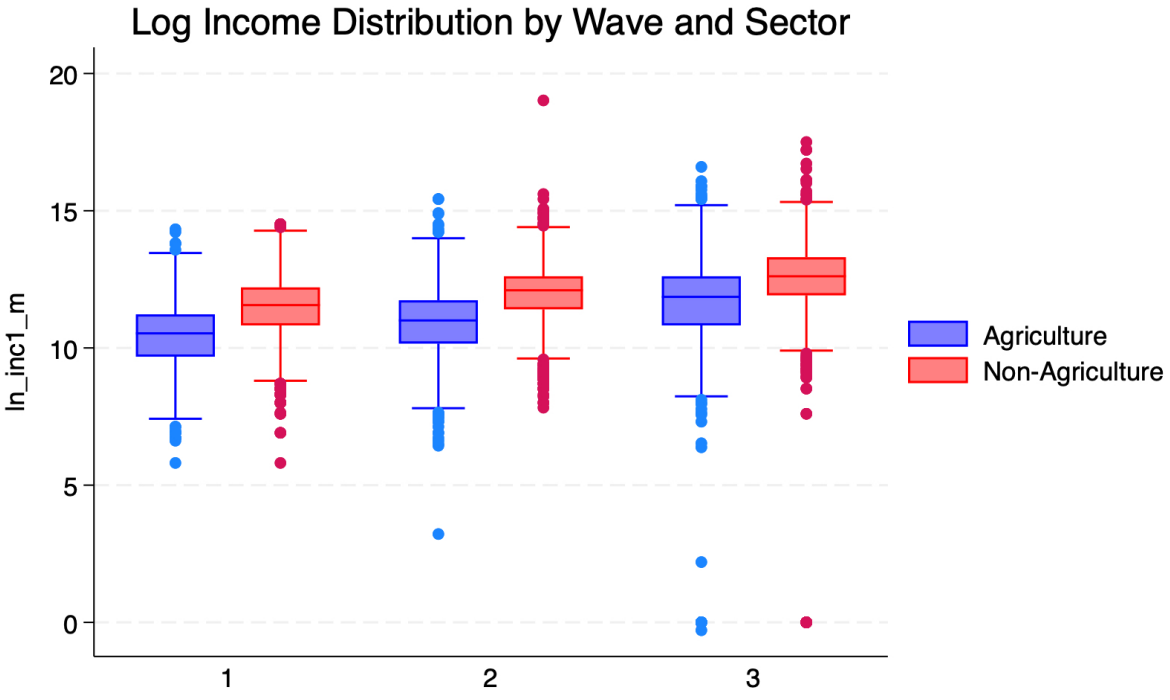

Figure 2: Log Income Non-agriculture vs. Agriculture over Three Waves

In Figure 3, urban areas show a sizable but declining APG: the non-ag/Ag earnings ratio falls from about 1.8 (1993) to 1.3 (2000). Rural areas remain higher on average—around 2.0

to 1.7—so the gap is generally larger outside cities. The only exception is 1997, when the urban gap briefly exceeds the rural one, consistent with the crisis hitting sectors differentially. I include urban–rural as a control in the reduced-form regressions.

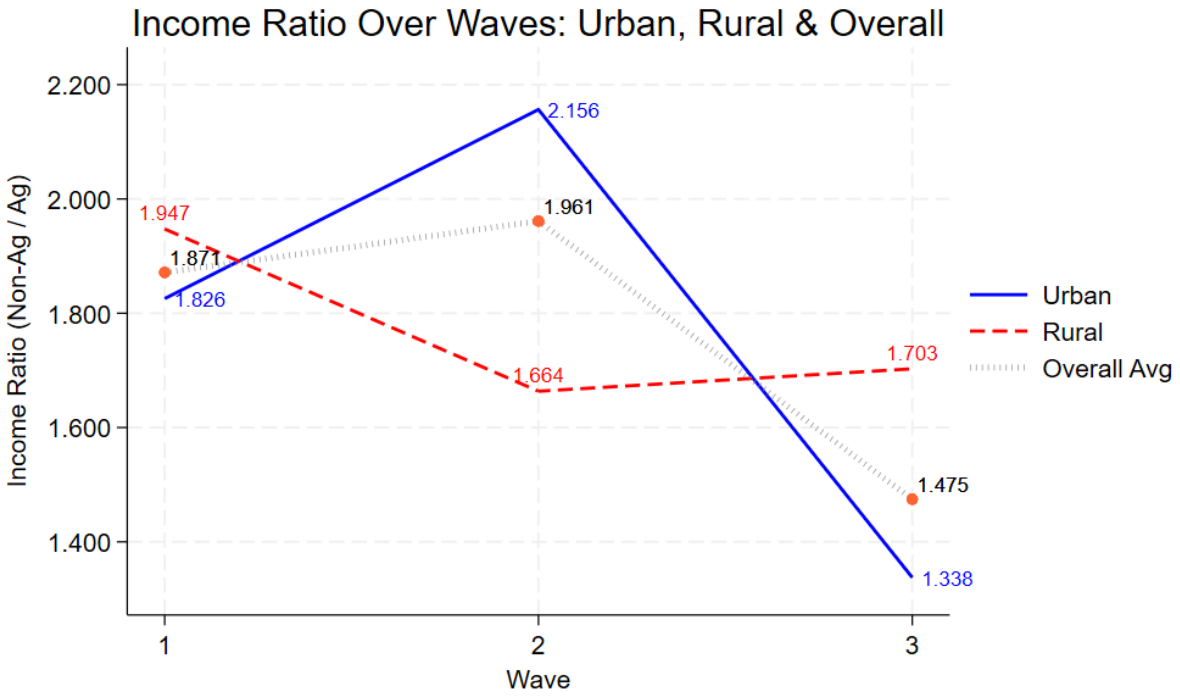

Figure 3: Sectoral Earnings Gaps Urban vs. Rural over Three Waves

Figures 2 and 3 show a persistent—but narrowing—agricultural productivity gap (APG) over 1993–2000. As outlined in the model section, such a decline can reflect (i) smaller sector-wide productivity differences and/or (ii) changes in individual sorting via observed characteristics or latent comparative advantage. While the analysis is not designed to decompose rural–urban trends per se, it will help shed light on the relative roles of these mechanisms and deepen our understanding of why the APG falls over this period.

Figure 4 displays the distribution of sectoral transition patterns across the three survey rounds. A majority of individuals exhibit strong sectoral attachment: 55% remain in non-agriculture across all waves (Nonag–Nonag–Nonag), and 25% stay continuously in agriculture. In contrast, sector switchers constitute about 20% of the sample, with no single transition path exceeding 5%. This persistence suggests that individuals make sectoral choices early in life—likely based on comparative advantage—and seldom alter them after-

ward. It lends empirical support to the assumption that unobserved sector-specific abilities, which drive sorting, are time-invariant over the study period. The stability of sectoral attachment may also reflect high costs of switching sectors, further discouraging mobility. At the same time, the roughly one-fifth of individuals who switch at least once provide meaningful variation essential for identifying the selection effect in the CRC model. The three-wave structure of the panel strengthens this identification by revealing not only whether individuals switch but also the direction and persistence of transitions (e.g., Ag–Nonag–Ag vs. Ag–Ag–Nonag), offering additional information to recover θ_i more precisely. Finally, when combined with the information on economic shocks across waves, these transition patterns reinforce the identification arguments developed in the previous section.

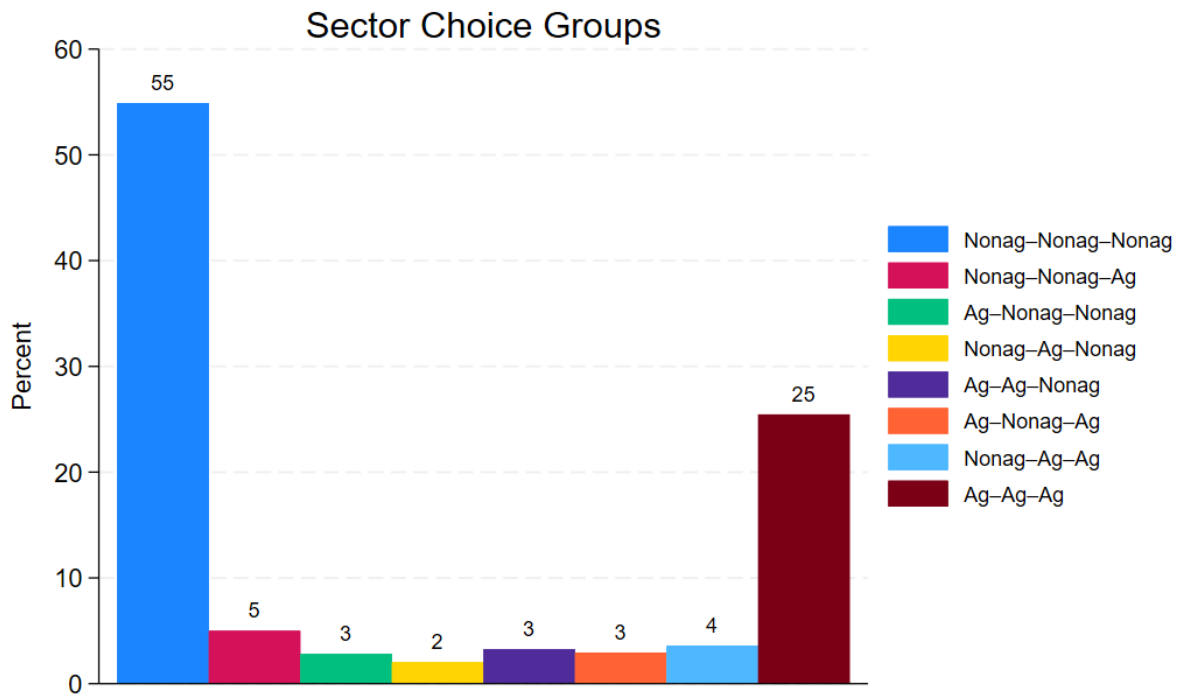

Figure 4: Choice Trajectories over Three Waves

Complementing the evidence on sectoral persistence, Figure 5 presents the share of individuals engaged in non-agricultural employment by urban and rural location. Participation in non-agriculture consistently exceeds 85% in urban areas and remains around 45–50% in rural areas, with these spatial differences remarkably stable across survey rounds. Sectoral classification refers to individuals' main jobs only; while some workers hold secondary activ-

ities in the other sector, those are not included because secondary-job data are sparse and incomplete. Urban and rural designations follow the administrative classifications recorded in the IFLS (based on BPS definitions). The sizable share of non-agricultural employment in rural areas reflects the prevalence of non-farm opportunities outside cities and indicates that sectoral transitions need not always coincide with geographic migration. This challenges the common strategy of using rural–urban migration as a proxy for sectoral transitions in the Indonesian context.

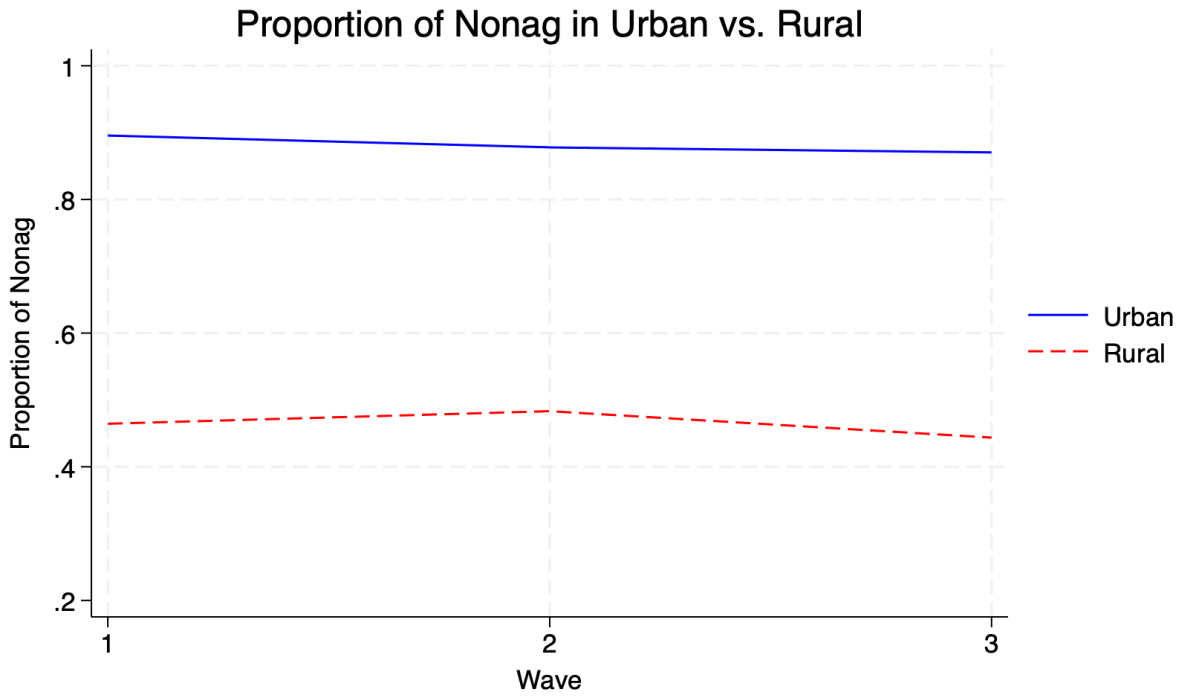

Figure 5: Non-agriculture Share Urban vs. Rural over Three Waves

Figure 6 plots the share of wage-employed workers by sector. Waged employment accounts for just one-quarter of agricultural workers, compared with around half of non-agricultural workers. Both sectors show a mild decline in formal (waged) employment between 1993 and 2000, implying a corresponding rise in informality. This figure shows especially pronounced informality in agriculture, where roughly three-quarters of workers are self-employed.

Together, these descriptive statistics highlight five key empirical patterns in the balanced panel from the first three IFLS waves: (1) income gaps between sectors are large but decline

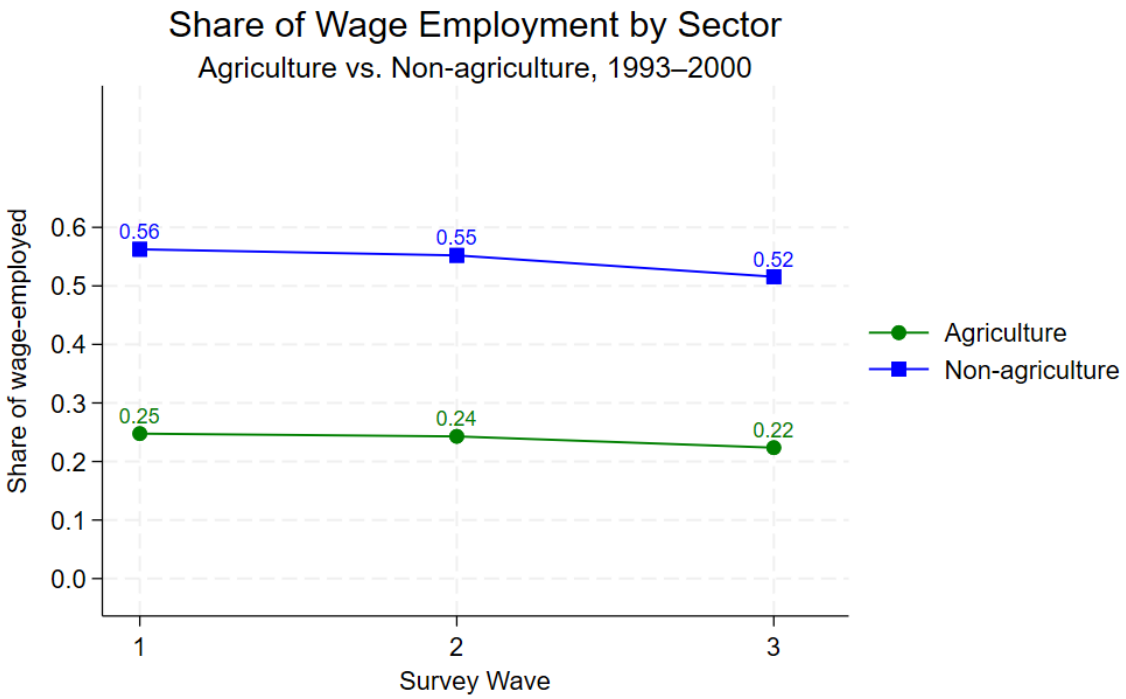

Figure 6: Formal vs. Informal Workers (Wage employment ≈ formal; self-employed ≈ informal)

over time, with distinct patterns across urban and rural areas; (2) earnings dispersion is substantial in both sectors; the left tail in agriculture likely reflects both true heterogeneity and some measurement error. (3) roughly 80% of workers remain in their initial sector—55% in non-agriculture and 25% in agriculture—while sector switchers account for 20%; (4) sectoral composition differs sharply between urban and rural areas, a distinction explicitly controlled for in the regression analysis; and (5) informality rises in both sectors, and remains much higher in agriculture.

Together, these descriptive patterns also provide indirect evidence for the identification assumptions introduced earlier. In particular, the strong persistence of sectoral attachment and the stability of sector composition suggest that short-term shocks are unlikely to drive sectoral mobility. The next subsection examines this point directly by using IFLS data on households' exposure to and coping strategies for economic shocks.

5.3 Empirical Evidence to the Identification Assumption

Empirical evidence from the Indonesia Family Life Survey (IFLS) supports the plausibility of the strict exogeneity assumption discussed in Section 4. In the first wave, respondents were asked whether they had experienced major economic shocks in the previous five years—such as crop failure, business loss, illness of a household member, or income loss (Figure 7)—and how they coped with those events (Figure 8). The most frequently reported coping mechanisms were taking on additional jobs, borrowing or receiving transfers from relatives, using savings, and reducing expenditures. Notably, switching sectors does not appear among the recorded responses, indicating that individuals tend to remain in their primary sector and absorb short-term shocks through adjustments within the same occupation rather than by changing sectors. This pattern aligns closely with the behavioural mechanism underlying the identification strategy: sectoral choices are stable, and temporary shocks do not alter the long-run comparative advantage that drives selection.

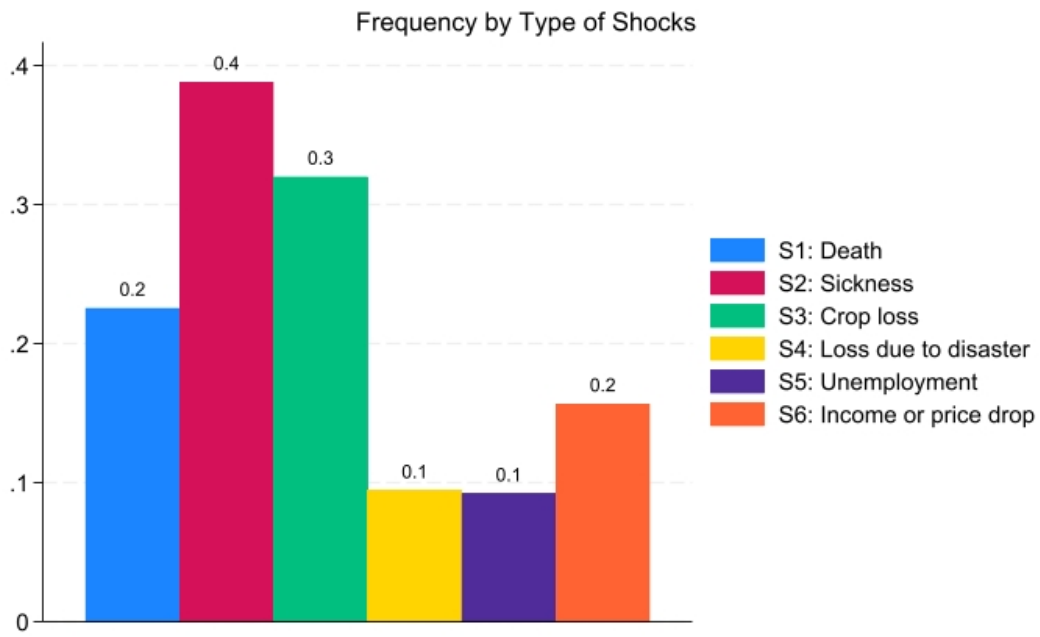

Figure 7: Types of Economic Shocks Reported by Households (IFLS Wave 1)

Temporary shocks may still influence short-run earnings through changes in inputs such as hours worked or secondary activities. Since the IFLS provides information on major

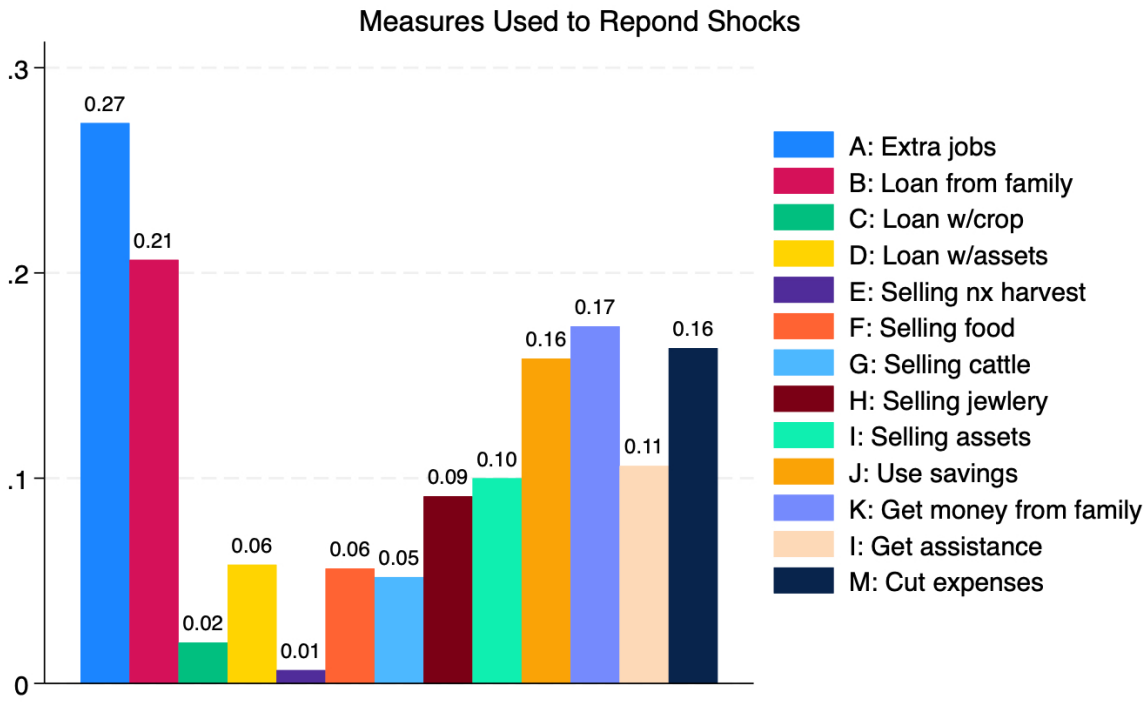

Figure 8: Coping Mechanisms Reported by Households (IFLS Wave 1)

household-level shocks, these factors can be explicitly controlled for in the estimation. Including such variables mitigates potential correlation between shocks, labor supply adjustments, and earnings, thereby strengthening the credibility of the strict exogeneity assumption for the transitory error term ϵ_{it}. In practice, this ensures that observed income fluctuations associated with temporary shocks do not bias the estimation of the structural parameters of interest.

Together with the descriptive evidence presented earlier—showing limited sectoral mobility and high persistence in primary occupation—these findings lend empirical support to the identification assumptions of the model. The next subsection turns to estimating the reduced-form coefficients as the first step toward recovering the structural parameters.

5.4 Estimates Reduced-Form Coefficients

Section 3.3 introduced the main empirical model (23). Equation (35) projects the unobserved comparative advantage θ_i onto the choice histories. Substituting (35) into (23) yields

reduced-form log earnings equation at each period.

$$
\begin{aligned}
w_{it} = & \underbrace{\delta_t^a}_{\equiv\eta} + \underbrace{(\delta_t^n - \delta_t^a)}_{\equiv\alpha} D_{it} \\
& + \underbrace{\lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1}D_{i2} + \lambda_5 D_{i1}D_{i3} + \lambda_6 D_{i2}D_{i3} + \lambda_7 D_{i1}D_{i2}D_{i3} + \nu_i}_{=\theta_i} \\
& + \beta(\underbrace{\lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1}D_{i2} + \lambda_5 D_{i1}D_{i3} + \lambda_6 D_{i2}D_{i3} + \lambda_7 D_{i1}D_{i2}D_{i3}}_{=\theta_i \text{(continued below)}} \\
& \underbrace{+\nu_i}) D_{it} + X_{it}\gamma^a + X_{it}(\gamma^n - \gamma^a) D_{it} + \tau_i + \epsilon_{it}, \quad t = \{1,2,3\} \qquad (36)
\end{aligned}
$$

To simplify calculation, let $\delta_t^a \equiv \eta$ and $\alpha \equiv \delta_t^n - \delta_t^a \ \forall t$. Equation (36) makes clear that the coefficients on the choice indicators and their interactions are functions of the underlying structural parameters. The observed characteristics X_{it} are included as covariates to improve precision. Rearranging terms yields the first-stage reduced-form regressions:[8]

$$
\begin{aligned}
w_{i1} = & \ \eta_1 + \phi_1 D_{i1} + \phi_2 D_{i2} + \phi_3 D_{i3} + \phi_4 D_{i1}D_{i2} + \phi_5 D_{i1}D_{i3} + \phi_6 D_{i2}D_{i3} + \phi_7 D_{i1}D_{i2}D_{i3} \\
& + X_{i1}'\gamma_{11} + X_{i2}'\gamma_{12} + X_{i3}'\gamma_{13} + e_{i1}, \qquad (37) \\
w_{i2} = & \ \eta_2 + \phi_8 D_{i1} + \phi_9 D_{i2} + \phi_{10} D_{i3} + \phi_{11} D_{i1}D_{i2} + \phi_{12} D_{i1}D_{i3} + \phi_{13} D_{i2}D_{i3} + \phi_{14} D_{i1}D_{i2}D_{i3} \\
& + X_{i1}'\gamma_{21} + X_{i2}'\gamma_{22} + X_{i3}'\gamma_{23} + e_{i2}, \qquad (38) \\
w_{i3} = & \ \eta_3 + \phi_{15} D_{i1} + \phi_{16} D_{i2} + \phi_{17} D_{i3} + \phi_{18} D_{i1}D_{i2} + \phi_{19} D_{i1}D_{i3} + \phi_{20} D_{i2}D_{i3} + \phi_{21} D_{i1}D_{i2}D_{i3} \\
& + X_{i1}'\gamma_{31} + X_{i2}'\gamma_{32} + X_{i3}'\gamma_{33} + e_{i3}. \qquad (39)
\end{aligned}
$$

Different from the ones in the Appendix F (which shows the two-period, no-covariate case), the reduced-form regression equations (37)-(39) add a third period, the corresponding interaction terms, and the X_{it} covariates as controls.[9]

The goal of the first-stage reduced-form regression is to estimate coefficients on the choice trajectories that map to the underlying structural parameters. To ensure that these estimates

[8] Appendix F presents the full algebraic derivation of the reduced-form regressions in a two-period case without covariates. Here, I simply state the reduced-form regressions in the first stage.

[9] The full set of interaction terms is written explicitly because the reported reduced-form tables (Tables 7–9) present the associated coefficients.

are not confounded, I include a comprehensive set of controls guided by the earlier descriptive analysis and identification assumptions. The covariates include indicators for urban/rural location, province, waged work (formal vs. informal), and household-level economic shocks, alongside standard observed characteristics such as age, education, marital status, hours worked, and gender. In addition, the logarithm of the provincial consumer price index (CPI) is included to account for spatial and temporal variation in price levels.

Tables 7, 8, and 9 report the reduced-form estimates for the 1993, 1997, and 2000 waves, respectively. Each table presents four specifications: Column (1) includes sectoral choices and their interactions across waves, without any controls. Column (2) adds location controls (urban/rural and province). Column (3) incorporates individual and job characteristics, including hours worked, waged work status, age, gender, marital status, religion, and education. Column (4) further adds province-level CPI (in logs) and a binary indicator for whether the household experienced any economic shock in the past five years. Because CPI is defined at the province level, province dummies are omitted in this full specification. Column (4) serves as the preferred model, while Column (1) provides a benchmark showing how coefficients shift when no controls are included.

Coefficients are in the first row for each regressor, and the corresponding standard errors are beneath them. A double-asterisk, **, denotes statistical significance at the 5% level, and a single-asterisk, *, indicates the 10% level. The end of each column reports the number of observations (N) and the R-squared value (R^2) for each regression. Five key patterns emerge from the reduced-form regressions:

1. Initial sectoral choice as a persistent predictor: The sector of employment in the first wave significantly influences earnings across all periods; by 2000, the interaction terms with choice in period 3 become significant, showing early choices matter for later earnings (persistence of sorting). In contrast, sectoral choices in later waves exhibit little effect. This persistence supports the assumption of time-invariant unobserved comparative advantage, consistent with the observed pattern of a relatively high tendency to stay in the initially chosen sector. Additionally, the contemporaneous sector-choice indicator exhibits significant explanatory power for earnings in the respective period.

2. The impact of urban-rural residence is more important than provincial locations in terms of explaining earnings. Column (2) adds two spatial controls: urban and province. Urban is a dummy variable equal to 1 if an individual resides in an urban area. The province is a categorical variable with 16 provinces: 13 initial provinces from the survey and 3 additional provinces due to households' relocation. The share of families that have moved across provinces is less than 0.1% in the balance sample. Once control for urban-rural locations, the provinces do not have significant explanatory power for earnings.

3. Urban-rural location in the initial period has significant explanatory power for earnings, but this explanatory power diminishes in the third period. This substantial impact of initial location is consistent with the low observed urban-rural mobility in the balanced sample, where the proportions are: 1,877 always-urban, 2,453 always-rural, and 285 switchers.

4. Hours worked exhibit a strong positive correlation with the earnings in the same period. Moreover, the dummy variable "wagedwork" captures the type of work for each individual, with a value of 1 for wage-paid jobs and 0 for self-employment. The type of employment has a significant impact on earnings, with formal employment paying higher, which aligns with the descriptive graph shown in the previous section.

5. CPI fluctuations align with economic shocks: The inclusion of log province-level CPI captures geographic price variation and some time-varying transitory shocks. After controlling for CPI, the economic shocks are not significant in the first period; however, they exhibit greater explanatory power for earnings in subsequent waves, suggesting that CPI may also absorb some of the transitory variation.

Finally, the covariates that commonly explain earnings behave as expected: age and education are consistently strong predictors of earnings. These results establish a credible reduced-form foundation for identifying structural parameters in the second-stage estimation. In addition, the point estimates are stable across the four specifications as the controls gradually increase.

Across specifications, R^2 rises as controls are added; standard earnings covariates (education, age, gender) enter with expected signs and magnitudes; and identification-motivated controls (urban/rural, CPI, shocks, hours, wage status) explain meaningful variation. Most importantly, the pattern that history matters—significant effects of initial non-ag status and the 2000 interaction terms—shows that the reduced form is capturing stable sectoral sorting rather than contemporaneous noise. Taken together, these results give confidence that the coefficients on the choice indicators and their interactions contain the right information to recover the underlying structural objects (β, θ_i). I therefore map the reduced-form estimates onto the structural parameters and quantify the extent to which the APG is due to the selection effect β rather than sector-wide productivity α.

5.5 Recovering Structural Parameters

The second-stage estimation recovers the structural parameters that underlie the reduced-form coefficients—specifically, the selection effect (β) and the distribution of unobserved comparative advantage (θ_i).[10] [11]

As shown in the reduced-form equations (37)–(39), coefficients of choice at each period and their interaction terms $(\phi' s)$ are functions of the underlying structural parameters: sector-wide efficiency gap α, selection effect β, and λ_1–λ_7, where these λ's can be used to recover the distribution of unobserved comparative advantage θ_i. Recall the main estimation

[10] The estimation is implemented using the **randcoef** package in Stata (Cabanillas et al., 2018). The estimation in the second stage uses the minimum-distance estimator (MDE).

[11] The complete mathematical formulation of the second-stage calculation refers to the paper by Cabanillas, Michler, Michuda, and Tjernström (2018), who developed this **randcoef** STATA package.

equation with the substitution of the linear projection of θ_i:

$$
\begin{aligned}
w_{it} = & \underbrace{\delta_t^a}_{\equiv\eta} + \underbrace{(\delta_t^n - \delta_t^a)}_{\equiv\alpha} D_{it} \\
& + \underbrace{\lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1}D_{i2} + \lambda_5 D_{i1}D_{i3} + \lambda_6 D_{i2}D_{i3} + \lambda_7 D_{i1}D_{i2}D_{i3} + \nu_i}_{=\theta_i} \\
& + \beta(\underbrace{\lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1}D_{i2} + \lambda_5 D_{i1}D_{i3} + \lambda_6 D_{i2}D_{i3} + \lambda_7 D_{i1}D_{i2}D_{i3}}_{=\theta_i \text{(continued below)}} \\
& \underbrace{+\nu_i}) D_{it} + X_{it}\gamma^a + X_{it}(\gamma^n - \gamma^a) D_{it} + \tau_i + \epsilon_{it}, \quad t = \{1, 2, 3\}
\end{aligned} \tag{36}
$$

Equation (36) displays all the structural parameters of interest: α, β, and λ_1–λ_7. Appendix F derives each coefficient's mathematical representation as the underlying structural parameters for a two-period case; extending to three periods simply adds one more period and additional interaction terms. Here, I only state the key structural parameters for clarity.

Table 1 presents the estimation results for these underlying structural parameters. Each column in Table 1 corresponds to one of the four model specifications used in the first-stage estimation (Tables 7–9). Column (1) excludes all covariates; Column (2) adds urban-rural residence and province; Column (3) further includes log hours worked, waged work, age, gender, education, religion and marital status; and Column (4) adds province-level log CPI and household-level economic shocks. The structural parameter estimates exhibit three key insights:

1. The selection effect (β) is positive but *not* statistically significant. The point estimate of β captures the extent to which sectoral income differentials stem from the individual sorting based on unobserved comparative advantages. Its values increase from 0.154 log points in the baseline specification to 0.39 under the full controls. In all the cases, $\beta > 0$ suggests positive selection: individuals with a stronger comparative advantage in non-agriculture are more likely to enter that sector. Moreover, the people who choose the non-agricultural jobs are those who are better farmers, i.e. earning higher than the average farmers. However, the standard errors are substantially large, ranging from 0.368 to 0.613, thereby rendering the estimates statistically insignificant.

   This insignificant result aligns with the substantial variation in income observed in the

data. As shown in Table 4, the pooled average monthly income from the primary job is IDR 268,329, while the standard deviation is IDR 1,711,989, more than six times larger than the mean. Moreover, this pattern of widespread monthly earnings is persistent in each wave. This high dispersion in earnings likely implies high dispersion in latent abilities. Such dispersion means that even if selection is strong at the individual level for some, it may not translate into a substantial sectoral wage gap after considering the distribution of comparative advantages. Here is where the distribution assumption matters to the estimation results.

This finding highlights the critical role of the underlying distribution of comparative advantages. A more concentrated distribution could yield a more significant aggregate selection effect, whereas a widely dispersed distribution, as observed here, dilutes its statistical significance. Hence, distributional assumptions on unobserved comparative advantages raise concerns for the empirical estimation of the selection effect, even though it can offer profound theoretical insights.

2. The estimated distribution of comparative advantages (θ_i) deviates substantially from normality. Recall Equation (35) expresses the unobserved comparative advantages θ_i in a three-period model. After obtaining estimates of λ_1–λ_7, normalize θ_i such that $\sum \theta_i = 0$. Then, I can obtain λ_0 by calculating the intercept of θ_i, using Equation (40).

$$\begin{aligned}\theta_i &= \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i1} D_{i2} \\ &\quad + \lambda_5 D_{i1} D_{i3} + \lambda_6 D_{i2} D_{i3} + \lambda_7 D_{i1} D_{i2} D_{i3} + \nu_i \qquad (35)\\ \lambda_0 &= -\lambda_1 \overline{D_{i1}} - \lambda_2 \overline{D_{i2}} - \lambda_3 \overline{D_{i3}} \\ &\quad - \lambda_4 \overline{D_{i1} D_{i2}} - \lambda_5 \overline{D_{i1} D_{i3}} - \lambda_6 \overline{D_{i2} D_{i3}} - \lambda_7 \overline{D_{i1} D_{i2} D_{i3}} \qquad (40)\end{aligned}$$

Once all the λ's are available, linear prediction recovers the empirical distribution of the revealed comparative advantages, $\hat{\theta}_i$. Figure 9 plots the recovered empirical distribution of $\hat{\theta}_i$ under specification (4) in Table 1. This figure has the normalized value of $\hat{\theta}_i$ labelled on the horizontal axis and the group share in the sample on the vertical axis. The distribution of comparative advantage is far from normal.

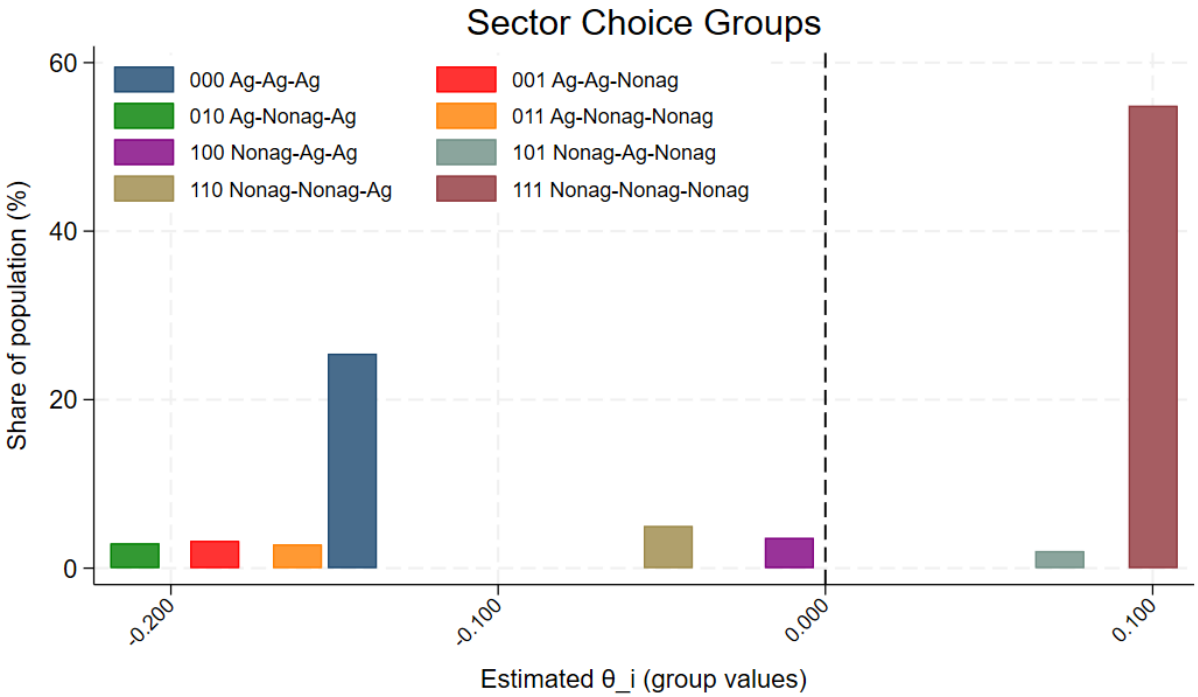

Figure 9: Recovered Distribution of Comparative Advantage for Specification (4)

This finding on the empirical distribution of comparative advantages aligns with extensive studies in the labour economics literature that question the conventional assumption of log-normality in the Roy model selection framework (Heckman and Honore, 1990). The two peaks located on both tails and the broad dispersion of θ_i call into question the empirical estimation approach that relies on strong parametric assumptions about unobserved heterogeneity in the Indonesian context in this study period.

Importantly, for individual selection to meaningfully influence sectoral productivity gaps, a sufficiently large share of the population must sort into sectors based on comparative advantages in the same direction, i.e the same sign in β. Without such alignment, even substantial individual-level sorting may fail to generate aggregate effects large enough to shift the observed APG.

3. Sectoral productivity differences (α) explain a substantial and statistically significant share of the observed earnings gap. The estimated sectoral premium in the baseline specification without controls is 0.509 log points, which declines slightly to 0.42 in

the complete specification with all the controls. Importantly, as the average values of α decrease, their associated standard errors also contract, enhancing the precision and statistical significance of the estimates. As shown in Table 1, the estimated α values across columns (1) to (4) are 0.509, 0.491, 0.413, and 0.420, with corresponding standard errors of 0.090, 0.082, 0.066, and 0.061. This consistent precision suggests that sector-wide productivity differences remain a robust and one of the primary factors in explaining the APG, even after controlling for a rich set of covariates and selection.

Table 1: Structural Parameters

Structural Parameters	**(1)**	**(2)**	**(3)**	**(4)**
λ_1	0.222 **	0.135 **	0.127 **	0.134 **
	(0.063)	(0.063)	(0.058)	(0.056)
λ_2	-0.037	-0.063	-0.097	-0.066
	(0.072)	(0.071)	(0.067)	(0.064)
λ_3	-0.055	-0.073	-0.054	-0.042
	(0.063)	(0.062)	(0.054)	(0.053)
λ_4	-0.031	-0.009	0.028	0.029
	(0.093)	(0.091)	(0.081)	(0.077)
λ_5	0.308 **	0.295 **	0.155	0.125
	(0.124)	(0.124)	(0.099)	(0.093)
λ_6	0.193 *	0.185 *	0.118	0.092
	(0.106)	(0.106)	(0.091)	(0.085)
λ_7	0.015	-0.063	-0.024	-0.026
	(0.153)	(0.148)	(0.127)	(0.112)
α	0.509 **	0.491 **	0.413 **	0.420 **
	(0.090)	(0.082)	(0.066)	(0.061)
β	0.154	0.119	0.180	0.390
	(0.368)	(0.444)	(0.508)	(0.613)

Notes: Numbers in parentheses are standard errors. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

In contrast, the selection effect (β) increases in magnitude across specifications but remains statistically insignificant throughout. The estimated values of β rise from 0.154 in column (1) to 0.39 in column (4), yet the standard errors increase even more sharply, from 0.368 to 0.613, resulting in wide confidence intervals and imprecise inference. This pattern suggests that while selection may play a role in shaping the sectoral choices at the individual level, its aggregate contribution to explaining the APG is weak and highly uncertain in this specific context.

One might argue that the wide dispersion of the estimated β reflects imprecise inference arising from large variance or potential misspecification of the first-stage reduced form. However, the reduced-form regressions based on the Mincer framework exhibit

reasonable explanatory power, with standard covariates behaving as expected and estimates remaining stable as additional controls are introduced. Moreover, as shown in the following Discussion section, when a joint-normal distribution of latent abilities is imposed—using the same dataset—the selection effect becomes statistically significant. This suggests that the insignificance of β in the present framework is not driven by weak identification or noise, but by the model's more flexible treatment of unobserved heterogeneity.

Overall, the recovered structural parameters indicate that sector-wide productivity differences (α) are the dominant source of the observed APG. As the model incorporates richer controls, α becomes more precisely estimated, while the contribution of individual sorting (β) remains weak and empirically fragile. This fragility stems from the wide dispersion and non-normality of the recovered distribution of θ_i, which dilutes aggregate selection effects even when they may be present at the individual level.

5.6 Individual Selection on APG

As discussed in the previous subsection, the selection effect β is not statistically significant in explaining the observed sector-level earnings. However, this extra reward for individuals in the non-agricultural sector captures only a partial selection effect from unobserved comparative advantages at the aggregate level. As described by the Equation (32) in Section 3.4, the impact of individual sorting based on latent abilities at the sector level comprises two components: one is the additional returns from the average workers by choosing the non-agricultural vs. agricultural sector, and the other is the average differences of the unobserved comparative advantages between the two sectors. β only reflects the selection effect from the former. Recall Equation (32):

$$S_\theta = \underbrace{\beta E[\theta_i|D=1]}_{\text{extra returns in nonag}} + \underbrace{(E[\theta_i|D=1] - E[\theta_i|D=0])}_{\text{mean diff in comparative advantages}} \tag{32}$$

To infer the second component of the sorting at the sector level, I need to further compute the difference of the conditional means of θ_i for the workers between the non-agricultural and

agricultural sectors. Since the estimated $\hat{\lambda}_0$ is not at level, I cannot calculate $E[\theta_i|D=1]$ and $E[\theta_i|D=0]$ separately. However, the difference in these two conditional means cancels the intercept λ_0 shown in the Equation (35). Therefore, I can sum the estimated $\hat{\theta}_i$ values in each sector at each wave, and then take the mean difference of these two estimated $\hat{\theta}_i$'s in the respective sectors. At each period, the mean difference between the estimated $\hat{\theta}_i$'s in two sectors yields the conditional mean difference in the unobserved comparative advantage, which is the second component in the Equation (32) and refers to how different the latent abilities are between the farmers and non-farmers on average at a given wave.

Table 2 shows the share of this conditional mean difference in APG at each period and the average over the waves. The observed log earnings differences at the sector level, which is APG, are in the second column. The difference in the means of latent abilities between farmers and non-farmers is in the third column, and the share of this conditional mean difference in the APG is in the far right column. The APG from the data are 1.0655, 1.0208, and 1.1224 log points in waves one to three, respectively, and 1.0696 log points on average. The latent skills between non-farmers and farmers are slightly negative, -0.0063, -0.0361, and -0.0285 log points in the sequential waves, and -0.0236 log points on average. Overall, the impact of this latent ability's conditional mean difference between non-agricultural and agricultural workers accounts for only 2.22% of the APG, with a direction opposite to that of the observed earnings gap.

Table 2: Conditional Mean Difference in APG

Wave	APG	$E[\theta_i\|D=1]-E[\theta_i\|D=0]$	$\frac{\Delta E[\theta_i]}{APG}$
1	1.0655	-0.0063	-0.0059
2	1.0208	-0.0361	-0.0354
3	1.1224	-0.0285	-0.0254
Average	1.0696	-0.0236	-0.0222

Notes: This table reports the estimated conditional mean difference in unobserved comparative advantage between non-agriculture ($D=1$) and agriculture ($D=0$) groups, corresponding to the second component in equation 32.

These results suggest that, on average, the latent abilities of non-agricultural and agricultural workers are not markedly different once individual heterogeneity is averaged out within each sector. This does not imply that farmers and non-farmers are identical in their unobserved comparative advantages—individual heterogeneity remains substantial—but rather that the sectoral means of these abilities are similar. The weak difference in latent abilities across sectors is consistent with the earlier finding that the estimated selection effect (β) is positive but dispersed widely: although some individuals self-select strongly based on comparative advantage, the aggregate sorting effect largely cancels out when averaged across the population.

In sum, aggregate sorting on unobserved comparative advantage plays a minor role in explaining the APG. The structural selection parameter (β) is positive but statistically insignificant, and the cross-sector mean difference accounts for only about 2.2% of the observed gap (with the opposite sign), implying that individual heterogeneity does not align strongly enough to shift sector averages. Substantively, this points to sector-wide productivity differences and observable characteristics—rather than who sorts where—as the dominant drivers of the APG.

This analysis of the first three waves of the IFLS yields three main insights. First, the sector-wide productivity gap remains the primary source of the APG, underscoring the importance of sectoral structure and technology in explaining aggregate productivity disparities. This highlights the central role of sector-level allocation efficiency and technological improvement in closing the APG. Second, while the selection effect is positive in magnitude, its statistical insignificance reflects the wide dispersion of latent abilities—suggesting that individual comparative advantages are highly heterogeneous but largely offset across sectors. A high average selection effect at the individual level may therefore translate into only modest aggregate consequences, depending on the underlying distribution of unobserved heterogeneity. Third, the empirical distribution of revealed comparative advantages features two distinct peaks and departs from normality, implying that imposing restrictive parametric assumptions could misrepresent the true structure of heterogeneity and selection.

6 Discussion

Contrary to the prevailing consensus in the APG literature, which attributes a substantial share of sectoral productivity gaps to self-selection, this paper finds that individual sorting contributes minimally to the average earnings disparities between the agricultural and non-agricultural sectors. Moreover, this study identifies a persistent and substantial sector-wide productivity gap as a crucial driver of the APG. Hence, the combination of the sector-wide technology difference and individual observed heterogeneity explains the majority of the sectoral productivity gaps when using a balanced panel in the first three waves of the IFLS data.

To reconcile my findings with the consensus in the current literature, I conduct a series of comparison exercises using the same balanced panel data in this study. First, I estimate the selection effect using a two-way fixed effects (TWFE) model and calculate the difference between the sectoral productivity gaps with and without controlling for individual fixed effects to infer the selection effect. This method yields a significant selection effect on APG. Next, I apply the canonical Heckman two-step estimator, which assumes that the distribution of the unobserved components is jointly normal. The results from both pooled and panel data exhibit significant selection on the sectoral productivity gaps. Hence, both the TWFE and distributional assumption methods produce estimates that align with the relevant findings in the APG literature.

Finally, I implement the selection bias correction procedure for panel data developed by Wooldridge (1995), which relaxes the joint normality assumption on individual latent abilities and applies a control function to correct for selection bias in panel settings. On the same dataset, once I relax the distributional assumptions on unobserved abilities, the estimated selection effect becomes statistically insignificant, which reconciles with my findings. The key takeaway from these comparison exercises shows that the estimation of the selection effect depends critically on the choice of the method. Therefore, it is essential to recognize the limitations of various approaches and choose the most suitable one to address the research question at hand.

6.1 Estimating Selection Using a TWFE Approach

Several studies in the APG literature rely on TWFE models to estimate or infer the impact of latent skills on the sectoral earnings gap, concluding that there is a large and significant selection effect on APG. This subsection shows that using the TWFE on the first three-wave IFLS dataset also yields a significant selection effect.

Table 10 reports pooled OLS for three IFLS waves with log primary-job earnings as the outcome. Adding covariates and year fixed effects reduces the raw APG, with the fully specified model yielding an observed gap of 0.651 log points. Table 11 estimates the same specifications in panel form; even-numbered columns include individual and year fixed effects (TWFE), odd-numbered columns omit individual fixed effects. Standard errors are clustered at the person level. Introducing individual fixed effects substantially lowers the agriculture–nonagriculture gap across specifications. Table 12 quantifies these reductions: pooled OLS vs. TWFE differs by 0.556–0.213 log points (Panels b–d), and adding individual fixed effects within the panel reduces the gap by 0.451–0.191 log points (Panels c–d). All differences are statistically significant at the 1% level, consistent with the claim in the literature that selection effects are large.

While these reductions are statistically significant and align with previous works, TWFE's interpretation as "selection on comparative advantage" is problematic for two reasons. First, identification is local to switchers. Identification in TWFE comes solely from people who change sectors: for the 80% who never switch, their choices (D_{it}) are time-invariant and differenced out by the individual fixed effect. As a result, the estimated coefficient of (D_{it}) reflects the behaviour of the 20% switchers—who may be systematically different from the full population relevant for APG. Second, TWFE absorbs *all* time-invariant heterogeneity. In the main model (eqs. (18)–(19)), permanent unobserved ability decomposes into a sector-relevant component, θ_i (comparative advantage), and a sector-irrelevant component, τ_i. The earnings equation (eq. (23)) shows that TWFE differences out both θ_i and τ_i. Because TWFE cannot separate sector-specific ability from common ability, it risks attributing earnings changes due to τ_i—e.g., general work ethic or family networks that help in either sector—to sector choices. In short, TWFE could conflate general individual heterogeneity unrelated

to sectoral choice with the actual sorting based on the latent skills, thereby overstating the contribution of the selection effect to the observed APG.

Unlike the fixed-effect method, the CRC approach models sector-specific latent abilities and exploits choice histories, isolating the comparative-advantage component θ_i rather than incorporating all time-invariant traits into a single undifferentiated fixed effect for each individual.

6.2 Empirical Consequences of Distributional Assumptions

After evaluating the limitations of the TWFE method in modelling selection based on latent skills, I turn to a second dominant strategy in the APG literature: parametric selection corrections rooted in Roy's (1951) framework. This approach assumes individuals choose sectors by comparing potential earnings, but only the chosen earnings are observed. While this setup mirrors the structure developed in Section 3, the literature typically departs in one critical respect: it imposes strong distributional assumptions on unobserved heterogeneity, most commonly joint normality of sector-specific abilities.

Why does this matter? Because the parametric assumption, rather than the data alone, often drives the size and significance of the estimated selection effect. For example, several studies—including Pulido and Świecki (2019)—extend the Roy model with selection and mobility frictions and estimate their frameworks on the IFLS. Using indirect inference (Gouriéroux et al., 1993), Pulido and Świecki match wage regressions and mobility patterns under the joint normality of latent abilities and idiosyncratic shocks. Within this structure, they find that selection accounts for 45–70% of the APG, depending on substitution elasticities. These results illustrate how significant selection effects can emerge under parametric assumptions, rather than directly from the data-generating process.

To assess how these assumptions shape empirical results in my context, I re-estimate the selection effect using the first three waves of IFLS under three standard parametric corrections: the canonical Heckman two-step (Heckman, 1979), the panel MLE estimator **xtheckman**, and Wooldridge's (1995) control function approach. For clarity, the full model setup for each method, as well as the discussion of exclusion restrictions, are provided in Appendix H; here I focus on comparative interpretation. These exercises demonstrate how

imposing or relaxing distributional assumptions on latent abilities changes the magnitude, and even the sign, of the estimated selection effect.

Table 15 reports the Heckman two-step results for the pooled data. The IMR coefficients are statistically significant in both sectors, implying systematic sectoral selection. Specifically, the estimates indicate negative selection into non-agriculture (IMR coefficient of around -0.14) and positive selection into agriculture (IMR coefficient ranging from 0.20 to 0.25). To bolster credibility, I impose exclusion restrictions: variables such as age and non-farm business are used for non-agriculture, while rural-born, marital status, and farm business are used for agriculture. As Tables 13–14 show, these instruments strongly predict sector choice but are not significant determinants of sectoral wages, supporting their validity.

While these exclusion restrictions are imperfect, they are motivated by Indonesia's structural context. Rural origin and household enterprise ownership strongly influence sectoral attachment and access to sector-specific opportunities, yet conditional on sector and standard covariates, they are plausibly orthogonal to individual productivity within sectors. Age and marital status, while potentially correlated with earnings, mainly capture lifecycle and household-constraint effects that shape sectoral participation rather than productivity conditional on employment.

Table 16 summarizes the implied contribution of selection to the observed APG. Depending on specification, selection into non-agriculture explains 21–22% of the earnings gap, while selection into agriculture accounts for 31–38%. These magnitudes align with the findings of Pulido and Świecki (2019), underscoring the strong influence of the joint normality assumption on estimates of selection.

To account for the panel nature of the data, I next apply `xtheckman`, a maximum likelihood estimator that extends Heckman's framework to panel settings. Rather than computing an IMR for each observation, this method directly estimates the correlation between unobserved sectoral fixed effects and time-varying error terms. Convergence proved challenging, with none of the specifications reaching full convergence despite extensive iterations. Nonetheless, where results are available (Table 17), the correlations between unobservables across sectors are large and statistically significant. This suggests that, under the joint normality assumption, selection bias remains substantial even in the panel framework. These

results broadly align with Pulido and Świecki's (2019) indirect inference estimates, highlighting that, regardless of estimation technique, assuming bivariate normality produces significant selection effects. Appendix H.2 provides further details.

Finally, I turn to Wooldridge's (1995) control function approach, which relaxes the joint normality assumption. Unlike Heckman or `xtheckman`, it does not hinge on any parametric assumption about the joint distribution of unobserved sectoral abilities. This method employs panel differencing to eliminate time-invariant heterogeneity and incorporates generalized residuals from a first-stage probit regression as control functions in the wage equation. Applied to the same dataset, the estimated coefficients on the control functions are statistically insignificant across specifications and both sectors (Table 20). Appendix H.3 provides the full model exposition and implementation details.

Taken together, these results demonstrate that distributional assumptions, rather than the data alone, drive the size and even the sign of the estimated selection effect. Under joint normality (as in Heckman or `xtheckman`), selection appears large and significant, explaining a non-trivial share of the APG. Under weaker assumptions (as in Wooldridge), the selection effect disappears. This sensitivity highlights the methodological risk of interpreting parametric Roy-model estimates as structural facts about the labour market: what appears to be strong evidence of self-selection may instead be an artifact of functional form assumptions imposed on unobserved heterogeneity.

While the assumption of joint normality in sector-specific latent abilities (θ_i^a, θ_i^n) can generate earnings distributions that resemble log-normality in the data, this consistency arises mechanically from the imposed functional form rather than from identification through variation in the data. In contrast, the CRC framework used here does not require a parametric distribution for unobserved comparative advantage. Instead, it infers the shape of θ_i directly from individuals' sectoral choice trajectories. Although this approach relies on a discrete support of estimated coefficients (λ_1–λ_7) rather than a continuous density, it allows the data to reveal the heterogeneity pattern empirically instead of forcing it to conform to a prespecified distribution. Thus, the resulting distribution of unobserved abilities is an empirical outcome, not an assumption, providing a conceptually distinct source of discipline relative to the parametric Roy framework.

In sum, the fixed-effect approach is not well-suited for studying self-selection based on latent skills because unobserved comparative advantages necessitate distinguishing individual fixed effects at the sector level, which the fixed-effect method cannot achieve. On the other hand, imposing distributional assumptions on latent skills across sectors often has a consequential impact on the estimation results of the selection effect. In this section, I demonstrate that the selection on APG would be significant if I were to implement either a fixed-effects or distributional assumption. This illustration reconciles the findings in my paper with the relevant studies in the APG literature. However, using the different empirical method by Suri (2011), I find that individual sorting is insignificant in terms of sectoral productivity gaps in Indonesia.

This paper estimates sector-specific comparative advantage without imposing any distributional assumptions. This method treats sectoral choices over time as informative signals of latent comparative advantage. It exploits the panel structure of the data to recover structural parameters non-parametrically. However, this method also faces limitations. As Tjernström et al. (2023) emphasize, identification requires variation in choice trajectories and earnings over time. When incomes are very similar across different choice groups or choice transitions are incomplete, the system of equations may become weakly identified or even break down. Moreover, the method requires a balanced panel and assumes that unobserved comparative advantages are time-invariant.

Despite these caveats, the Suri-inspired approach offers a valuable alternative to existing methods by avoiding functional form assumptions and directly modelling unobserved comparative advantages. Given the stark contrast in results across the three parametric approaches examined in this section, and the fragility of distributional assumptions in this context, the Suri-based estimator provides a theoretically and empirically grounded alternative to revisit long-standing claims about the role of selection in explaining agricultural productivity gaps.

In the IFLS setting, the evidence points to sector-wide productivity differences—rather than selection on latent skills—as the primary driver of the APG. This shifts the policy margin toward sectoral fundamentals (technology adoption, input and output market access, infrastructure, and management/extension) and away from policies premised on re-sorting

workers across sectors. I return to these implications in the conclusion, where I outline concrete levers consistent with the Indonesian context and the patterns documented above.

7 Conclusion

This paper revisits a core question: how much of the agricultural productivity gap (APG) can be explained by individual sorting on unobserved comparative advantage? Using the Indonesia Family Life Survey (IFLS), I focus on the pre-decentralization 1990s—a relatively stable decade before the 2000 "Big Bang" reforms. In contrast to much of the APG literature that pools all five IFLS waves and finds large selection effects, the evidence here shows that sorting contributed only a small share to the gap in this earlier period. Instead, sector-wide productivity differences and observable heterogeneity account for most of the earnings disparity between agriculture and non-agriculture.

This difference in findings is the result of adopting an alternative empirical strategy that (i) separates fixed effects unrelated to sectoral choice from sector-relevant latent abilities, (ii) uses information from both stayers and switchers, and (iii) avoids strong distributional assumptions on comparative advantage. As shown in the Discussion section, applying prevailing methods to the same data (TWFE and parametric selection corrections) yields significant selection effects—illustrating how assumptions and design choices, rather than the data alone, can drive results.

The paper contributes to the literature in two dimensions. Methodologically, it extends the correlated random coefficient (CRC) framework beyond its original context to analyze sectoral productivity gaps. Three modifications are central. First, it redefines absolute advantage as sector-specific—θ_i^n and θ_i^a—rather than as a single, sector-neutral component as in Suri (2011). This restores the theoretical link between absolute and comparative advantage emphasized in the APG literature and matches how individuals choose sectors—by comparing sector-specific returns. Second, it adds an aggregation step that translates individual selection into sector-level contributions to the observed APG, decomposing sorting into (i) the selection effect β and (ii) the mean difference in latent abilities across sectors. Third, it provides an economic interpretation of what was a purely technical projection step in the

original CRC approach: separating latent abilities (θ_i) from choices (D_{it}) is reinterpreted as revealed comparative advantage, connecting estimation mechanics to the decision process behind sectoral sorting. Together, these extensions tie individual comparative advantage to aggregate productivity outcomes in a way that is theoretically grounded and empirically tractable.

Conceptually, this paper bridges the CRC and Roy-model traditions through two mappings. First, it links the CRC selection effect β to the coefficient on the inverse Mills ratio in Heckman-type estimators—clarifying when these objects coincide and how they differ when distributional assumptions are relaxed. Second, it maps the two components of aggregate sorting—the extra returns to unobserved advantage and the cross-sector mean difference in latent abilities—onto conditional means in the classic Roy model. This synthesis clarifies how micro-level sorting aggregates into sectoral productivity differences.

Sorting affects the APG through two channels, not one. Moreover, the empirical distribution of latent abilities in IFLS deviates from normality, with mass in both tails. Under joint normality, the sign of the selection effect β and the sign of the cross-sector mean difference in latent abilities typically align (differing mainly in magnitude). Without such parametric assumptions, they *need not* align—indeed, in the IFLS they diverge. Policies guided solely by β could therefore miss (or even counteract) the implications of cross-sector mean differences when the underlying distribution is skewed or fat-tailed.

Two illustrative cases highlight the stakes. In Sub-Saharan Africa (2000s–2010s), many skills and entrepreneurship programs delivered modest gains. A plausible contributing factor is that they implicitly assumed general training would raise non-farm earnings, where β was weak or negative and mean ability gaps were large, moving workers with poor sectoral fit may have limited payoffs. Program design, demand constraints, capital scarcity, and local frictions also likely mattered. By contrast, during Indonesia's Green Revolution, rice-intensification programs in the 1970s–early 1990s directly raised agricultural productivity. The final phase (Repelita V, 1989–1994) overlaps the beginning of the IFLS period studied here and is consistent with narrowing earnings gaps without large-scale reallocation. When sector-wide efficiency (α) dominates sorting (β), policies that boost within-sector productivity—rather than encouraging cross-sector migration—can yield more equitable and durable gains.

Taken together, these results caution against treating selection as a single-number object. Effective policy should consider both terms—returns to comparative advantage and average ability gaps across sectors—and recognize that distributional assumptions about latent abilities could tilt policy toward reallocation or other directions.

In summary, adopting a distribution-free CRC framework and focusing on Indonesia's pre-decentralization years, the paper shows that individual sorting accounted for only a small portion of the APG in the 1990s. The gap was driven mainly by sector-wide differences in technology and observables. More broadly, revealed comparative advantage provides a practical lens for diagnosing whether policy should prioritize reallocation or within-sector productivity—an empirical question with answers that can vary across time, place, and institutions.

From a broader perspective, these findings speak directly to structural transformation in the Lewis dual-economy tradition. In the Lewis model, growth arises from reallocating surplus labour from a low-productivity agricultural sector to a high-productivity modern sector. Yet in the 1990s Indonesia, transformation appears to have relied less on large-scale reallocation and more on within-sector productivity improvements—especially in agriculture. The weak selection effect (β) alongside a dominant sectoral efficiency gap (α) suggests an economy moving beyond the classic surplus-labour phase: convergence owed more to productivity deepening than to worker migration. At the same time, the later rise of low-end services raises a distinct concern: if future APG reductions come mainly from absorption into low-productivity services rather than true convergence, the growth implications differ sharply. Understanding how APG falls—through reallocation, absorption, or within-sector progress—remains central for assessing the quality and sustainability of structural transformation.

Table 3: Comparison Between Restricted and Unrestricted Samples (IFLS Waves 1–3)

Variable	**Restricted Sample**	**Unrestricted Sample**
Age (years)	44	40
Male (%)	72.7	53.7
Married (%)	90.8	75.9
Education level (%)		
Unscholled	15.1	15.7
Primary	51.4	45.8
Junior high	11.9	13.1
Senior high	16.5	17.2
College / university	5.1	6.2
Others	0	2
Urban (%)	43.9	46.2
Non-agriculture (%)	64.7	64
Waged work (%)	43.6	42.6
Log earnings	11.62	11.59
Log hours worked	5.04	4.95

Notes: The restricted sample only includes individuals with complete information on both earnings and sectoral choices in all three waves.

Table 4: Descriptive Statistics for Individuals – Part A

	Number of Survey Round			
Variable	**1**	**2**	**3**	**Total**
N	4,615	4,615	4,615	13,845
	(33.3%)	(33.3%)	(33.3%)	(100.0%)
hh_count	3,963.000	3,991.000	4,012.000	4,012.667
	(98.8%)	(99.5%)	(100.0%)	(100.0%)
gender	0.727	0.727	0.727	0.727
	(0.446)	(0.446)	(0.446)	(0.446)
age	40.407	44.335	47.308	44.016
	(11.250)	(11.173)	(11.277)	(11.504)
non-agriculture (primary job)	0.655	0.657	0.630	0.647
	(0.475)	(0.475)	(0.483)	(0.478)
wagedwork (primary job)	0.454	0.446	0.408	0.436
	(0.498)	(0.497)	(0.491)	(0.496)
income	126,194.800	246,299.956	432,582.566	268,359.107
	(167,450.486)	(2,692,023.335)	(1,212,964.123)	(1,711,988.882)
ln_income	11.127	11.629	12.096	11.617
	(1.209)	(1.213)	(2.026)	(1.582)
hours worked	177.964	171.676	172.585	174.088
	(74.269)	(74.998)	(81.850)	(77.168)
ln_hours worked	5.076	5.027	5.014	5.039
	(0.505)	(0.546)	(0.596)	(0.551)
cpi	145.195	194.337	209.439	182.990
	(4.604)	(7.064)	(9.432)	(28.083)
ln_cpi	4.978	5.269	5.343	5.197
	(0.031)	(0.036)	(0.045)	(0.168)
ruralborn	0.746	0.749	0.792	0.763
	(0.435)	(0.434)	(0.406)	(0.425)
moved	0.533	0.465	0.462	0.487
	(0.499)	(0.499)	(0.499)	(0.500)
urban	0.442	0.439	0.436	0.439
	(0.497)	(0.496)	(0.496)	(0.496)

Notes: Top line reports means (or counts for *N*, *hh_count*); second line reports standard deviations for continuous/binary variables and percentages for *N* and *hh_count*. Monetary values in Indonesian Rupiah. Statistics use the balanced panel (waves 1–3).

Table 5: Descriptive Statistics for Individuals – Part B (Categorical Variables)

	Number of Survey Round			
Category	**1**	**2**	**3**	**Total**
N	4,615 (33.3%)	4,615 (33.3%)	4,615 (33.3%)	13,845 (100.0%)
MARITAL STATUS				
Not yet married	99 (2.1%)	54 (1.2%)	38 (0.8%)	191 (1.4%)
Married	4,231 (91.7%)	4,194 (90.9%)	4,146 (89.8%)	12,571 (90.8%)
Separated	19 (0.4%)	22 (0.5%)	24 (0.5%)	61 (0.4%)
Divorced	56 (1.2%)	56 (1.2%)	59 (1.5%)	181 (1.3%)
Widowed	210 (4.6%)	289 (6.3%)	342 (7.4%)	841 (6.1%)
EDUCATION				
Unschooled	698 (15.1%)	651 (14.1%)	619 (13.4%)	1,968 (14.2%)
Primary	2,372 (51.4%)	2,434 (52.8%)	2,362 (51.2%)	7,168 (51.8%)
Junior high	548 (11.9%)	509 (11.0%)	485 (10.5%)	1,542 (11.1%)
Senior high	760 (16.5%)	745 (16.1%)	667 (14.5%)	2,172 (15.7%)
College/University	233 (5.1%)	273 (5.9%)	341 (7.4%)	847 (6.1%)
Others	0 (0.0%)	2 (0.0%)	136 (3.0%)	138 (1.0%)
RELIGION				
Islam	4,007 (86.8%)	4,030 (87.3%)	4,020 (87.1%)	12,057 (87.1%)
Protestant	187 (4.1%)	185 (4.0%)	192 (4.2%)	564 (4.1%)
Catholic	90 (2.0%)	92 (2.0%)	94 (2.0%)	276 (2.0%)
Hindu	286 (6.2%)	281 (6.1%)	284 (6.2%)	851 (6.1%)
Buddhist	23 (0.5%)	22 (0.5%)	19 (0.4%)	64 (0.5%)
Others	22 (0.5%)	7 (0.2%)	4 (0.1%)	33 (0.2%)

Notes: Each cell shows count with share in parentheses. Shares sum to 100% within each block.

Table 6: Descriptive Statistics for Households

	Number of Survey Round			
Variable	**1**	**2**	**3**	**Total**
N	3,963	3,991	4,012	11,966
	(33.1%)	(33.4%)	(33.5%)	(100.0%)
panel	.	0.980	0.978	0.979
	(.)	(0.141)	(0.146)	(0.144)
household size	4.777	5.361	5.801	5.315
	(1.981)	(2.180)	(2.408)	(2.237)
urban	0.430	0.426	0.427	0.428
	(0.495)	(0.495)	(0.495)	(0.495)
own farm business	0.434	0.397	0.476	0.436
	(0.496)	(0.489)	(0.500)	(0.496)
farm business assets	7,144,077.782	10,330,989.860	26,572,186.430	15,729,198.326
	(33,921,100.065)	(21,343,301.269)	(83,485,328.313)	(57,595,891.874)
ln_farm business assets	14.064	14.798	15.550	14.864
	(2.216)	(1.998)	(2.127)	(2.208)
owns non-farm business	0.374	0.398	0.523	0.432
	(0.484)	(0.490)	(0.500)	(0.495)
non-farm business assets	6,374,035.384	5,439,189.255	9,510,752.091	7,351,954.719
	(59,736,751.708)	(24,837,647.647)	(46,064,887.772)	(45,183,582.166)
ln_non-farm business assets	12.257	12.976	13.350	12.931
	(2.356)	(2.339)	(2.434)	(2.423)
own family business	0.675	0.672	0.803	0.717
	(0.469)	(0.469)	(0.398)	(0.451)
own assets	0.976	0.999	0.999	0.991
	(0.153)	(0.035)	(0.035)	(0.093)
total assets not for business	14,005,326.895	20,270,537.043	37,537,032.473	24,060,081.890
	(78,347,639.350)	(52,369,917.086)	(85,931,361.293)	(74,282,490.131)
ln_total assets not for business	14.709	15.654	16.392	15.595
	(1.785)	(1.621)	(1.586)	(1.801)
real estate not for business	11,356,720.903	18,508,233.980	32,133,196.522	20,998,405.810
	(56,657,512.411)	(49,100,027.450)	(67,800,563.543)	(59,108,696.917)
ln_real estate not for business	14.826	15.623	16.309	15.611
	(1.573)	(1.519)	(1.522)	(1.651)
shock	0.312	0.402	0.345	0.353
	(0.463)	(0.490)	(0.475)	(0.478)
numbers of shock	0.392	0.567	0.443	0.467
	(0.653)	(0.821)	(0.700)	(0.732)

Notes: Top line reports means (or counts for N); second line reports standard deviations (or column shares for N). Monetary values are in Indonesian Rupiah.

Table 7: SUR: Outcome Variable, log Earnings in 1993

ln_earnings_1	**(1)**	**(2)**	**(3)**	**(4)**
nonag_1	0.693 **	0.592 **	0.53 **	0.539 **
	(0.09)	(0.089)	(0.08)	(0.079)
nonag_2	-0.036	-0.057	-0.097	-0.092
	(0.099)	(0.097)	(0.088)	(0.087)
nonag_3	0.012	-0.009	0.037	0.059
	(0.094)	(0.093)	(0.082)	(0.082)
nonag_1 * nonag_2	0.028	0.043	0.075	0.100
	(0.148)	(0.145)	(0.13)	(0.129)
nonag_1 * nonag_3	0.246	0.220	0.076	0.052
	(0.169)	(0.166)	(0.148)	(0.147)
nonag_2 * nonag_3	0.207	0.188	0.129	0.114
	(0.163)	(0.160)	(0.144)	(0.143)
nonag_1 * nonag_2 * nonag_3	0.002	-0.097	-0.056	-0.058
	(0.228)	(0.224)	(0.201)	(0.199)
urban_1		0.452 **	0.185 **	0.185 **
		(0.097)	(0.087)	(0.086)
urban_2		-0.192	0.151	0.109
		(0.114)	(0.103)	(0.101)
urban_3		0.067	-0.067	-0.047
		(0.084)	(0.075)	(0.074)
ln_hrsworked_1			0.362 **	0.354 **
			(0.031)	(0.030)
ln_hrsworked_2			0.053 *	0.051 **
			(0.028)	(0.028)
ln_hrsworked_3			0.026	0.026
			(0.026)	(0.025)
wagedwork_1			0.106 **	0.105 **
			(0.042)	(0.042)
wagedwork_2			0.086 **	0.088 **
			(0.045)	(0.044)
wagedwork_3			-0.065	-0.057
			(0.042)	(0.042)
ln_cpi_1				-3.051 **
				(0.941)
ln_cpi_2				3.609 **
				(0.726)
ln_cpi_3				1.082 **
				(0.408)
shock_1				-0.021
				(0.031)
shock_2				-0.008
				(0.030)
shock_3				0.038
				(0.030)
province	N	Y	Y	N
age	N	N	Y **	Y **
gender	N	N	Y	Y
marital_status	N	N	Y	Y
religion	N	N	Y	Y
education	N	N	Y **	Y **
constant	10.412 **	10.391 **	6.919 **	-2.680
	(0.032)	(0.050)	(0.225)	(3.741)
N	4,615	4,614	4,513	4,510
R^2	0.189	0.221	0.392	0.402

Notes: Numbers in parentheses are **standard errors**. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

Table 8: SUR: Outcome Variable, log Earnings in 1997

ln_earnings_2	**(1)**	**(2)**	**(3)**	**(4)**
nonag_1	0.277 **	0.186 *	0.165 **	0.177 **
	(0.091)	(0.090)	(0.079)	(0.078)
nonag_2	0.462 **	0.437 **	0.340 **	0.352 **
	(0.099)	(0.098)	(0.087)	(0.087)
nonag_3	-0.045	-0.064	-0.070	-0.055
	(0.095)	(0.093)	(0.082)	(0.081)
nonag_1 * nonag_2	-0.071	-0.054	-0.036	-0.021
	(0.149)	(0.147)	(0.130)	(0.128)
nonag_1 * nonag_3	0.295	0.276 *	0.141	0.114
	(0.170)	(0.168)	(0.147)	(0.146)
nonag_2 * nonag_3	0.114	0.105	0.049	0.030
	(0.164)	(0.162)	(0.143)	(0.142)
nonag_1 * nonag_2 * nonag_3	0.137	0.358	0.087	0.101
	(0.230)	(0.226)	(0.199)	(0.198)
urban_1		0.438 **	0.180 **	0.199 **
		(0.097)	(0.087)	(0.086)
urban_2		-0.127	0.051	0.041
		(0.115)	(0.102)	(0.101)
urban_3		0.131	0.0003	-0.041
		(0.084)	(0.075)	(0.074)
ln_hrsworked_1			0.099 **	0.087 **
			(0.030)	(0.030)
ln_hrsworked_2			0.27 **	0.269 **
			(0.028)	(0.028)
ln_hrsworked_3			0.057 **	0.058 **
			(0.025)	(0.025)
wagedwork_1			-0.018	-0.008 **
			(0.041)	(0.041)
wagedwork_2			0.185 **	0.181 **
			(0.044)	(0.044)
wagedwork_3			-0.165	-0.011
			(0.042)	(0.041)
ln_cpi_1				-5.449 **
				(0.933)
ln_cpi_2				4.849 **
				(0.720)
ln_cpi_3				0.444
				(0.405)
shock_1				0.063 **
				(0.031)
shock_2				-0.044
				(0.030)
shock_3				0.003
				(0.030)
province	N	Y	Y	N
age	N	N	Y **	Y **
gender	N	N	Y	Y
marital_status	N	N	Y	Y
religion	N	N	Y	Y
education	N	N	Y **	Y **
constant	10.906 **	10.992 **	7.902 **	7.007 **
	(0.032)	(0.050)	(0.224)	(3.713)
N	4,615	4,614	4,513	4,510
R^2	0.185	0.212	0.406	0.417

Notes: Numbers in parentheses are **standard errors**. Asterisks indicate significance: ***** $p < 0.10$, ****** $p < 0.05$.

Table 9: SUR: Outcome Variable, log Earnings in 2000

ln_earnings_3	**(1)**	**(2)**	**(3)**	**(4)**
nonag_1	0.111	-0.048	0.014	0.038
	(0.161)	(0.161)	(0.156)	(0.156)
nonag_2	-0.221	-0.237 **	-0.408 **	-0.403 **
	(0.176)	(0.175)	(0.173)	(0.172)
nonag_3	0.173	0.163	0.071	0.044
	(0.168)	(0.168)	(0.162)	(0.162)
nonag_1 * nonag_2	-0.059	-0.036	0.159	0.164
	(0.264)	(0.264)	(0.256)	(0.255)
nonag_1 * nonag_3	0.729 **	0.723 **	0.592 **	0.586 **
	(0.302)	(0.264)	(0.291)	(0.290)
nonag_2 * nonag_3	0.516 *	0.511 *	0.454	0.467 *
	(0.291)	(0.291)	(0.283)	(0.282)
nonag_1 * nonag_2 * nonag_3	-0.065	-0.097	-0.216	-0.202
	(0.407)	(0.402)	(0.394)	(0.392)
urban_1		0.294	0.001	0.020
		(0.177)	(0.171)	(0.170)
urban_2		-0.192	-0.002	-0.055
		(0.208)	(0.202)	(0.200)
urban_3		0.161	0.077	0.071
		(0.152)	(0.148)	(0.146)
ln_hrsworked_1			0.035	0.029
			(0.060)	(0.060)
ln_hrsworked_2			0.098 *	0.106 **
			(0.055)	(0.055)
ln_hrsworked_3			0.323 **	0.313 **
			(0.050)	(0.050)
wagedwork_1			-0.149 *	-0.139 *
			(0.082)	(0.082)
wagedwork_2			0.069	0.075
			(0.088)	(0.087)
wagedwork_3			0.308 **	0.303 **
			(0.082)	(0.082)
ln_cpi_1				-3.748 **
				(1.854)
ln_cpi_2				3.671 **
				(1.431)
ln_cpi_3				0.937
				(0.805)
shock_1				0.201 **
				(0.061)
shock_2				0.006
				(0.059)
shock_3				-0.242 **
				(0.060)
province	N	Y	Y	N
age	N	N	Y **	Y
gender	N	N	Y	Y
marital_status	N	N	Y	Y
religion	N	N	Y	Y
education	N	N	Y **	Y **
constant	11.418 **	11.441 **	8.704 **	3.066
	(0.057)	(0.091)	(0.443)	(7.377)
N	4,615	4,614	4,513	4,510
R^2	0.083	0.087	0.168	0.175

Notes: Numbers in parentheses are **standard errors**. Asterisks indicate significance: * $p < 0.10$, ** $p < 0.05$.

ln_inc1_m	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
nonag_main	1.053 **	1.071 **	0.666 **	0.654 **	0.653 **	0.656 **	0.646 **	0.651 **
	0.027	0.033	0.03	0.034	0.029	0.034	0.029	0.033
ln_hrsworked1_m			0.337 **	0.349 **	0.353 **	0.35 **	0.351 **	0.349 **
			0.022	0.028	0.021	0.028	0.021	0.028
wagedwork_main			0.188 **	0.202 **	0.195 **	0.203 **	0.186 **	0.196 **
			0.027	0.026	0.026	0.026	0.026	0.026
age			0.004 **	-0.006 **	-0.005 **	-0.006 **	-0.005 **	-0.006 **
			0.001	0.001	0.001	0.001	0.001	0.001
gender			0.392 **	0.406 **	0.401 **	0.405 **	0.403 **	0.407 **
			0.03	0.03	0.029	0.03	0.029	0.03
education			0.368 **	0.323 **	0.331 **	0.323 **	0.333 **	0.325 **
			0.012	0.013	0.011	0.013	0.011	0.013
marital_status			-0.038 **	-0.038 *	-0.039 **	-0.039 **	-0.036 **	-0.037 **
			0.018	0.019	0.017	0.019	0.017	0.02
urban			0.238 **	0.257 **	0.252 **	0.253 **	0.249 **	0.251 **
			0.027	0.029	0.026	0.029	0.026	0.029
ln_cpi					2.457 **	1.926 **	2.469 **	1.933 **
					0.073	0.328	0.073	0.328
shock							-0.086 **	-0.07 **
							0.024	0.026
Year								
1997		0.5 **		0.537 **		-0.024		-0.021
		0.016		0.017		0.097		0.097
2000		0.996 **		1.013 **		0.307 **		0.306 **
		0.029		0.029		0.124		0.124
cons	10.936 **	10.425 **	8.01 **	7.956 **	-4.381 **	-1.649	-4.395 **	-1.65
	0.0215	0.027	0.126	0.017	0.388	1.658	0.388	1.658
clustered	N	Y	N	Y	N	Y	N	Y
N	13,845	13,845	13,742	13,742	13,742	13,742	13,739	13,739
R^2	0.101	0.167	0.225	0.2886	0.285	0.291	0.286	0.291

Table 10: Pooled OLS Regressions, With and Without Year Fixed Effects

ln_inc1_m	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
nonag_main	0.966 **	0.515 **	0.634 **	0.446 **	0.636 **	0.446 **	0.629 **	0.438 **
	0.035	0.072	0.035	0.069	0.035	0.069	0.035	0.069
ln_hrsworked1_m			0.326 **	0.268 **	0.328 **	0.269 **	0.326 **	0.264 **
			0.028	0.037	0.028	0.037	0.028	0.037
wagedwork_main			0.217 **	0.227 **	0.217 **	0.226 **	0.208 **	0.221 **
			0.026	0.055	0.026	0.055	0.026	0.054
age			-0.007 **	0.0001	-0.006 **	0.0001	-0.006 **	-0.001
			0.001	0.007	0.001	0.007	0.001	0.007
gender			0.406 **	0.874	0.405 **	0.863	0.408 **	0.946 *
			0.031	0.531	0.03	0.539	0.305	0.495
education			0.3 **	0.3	0.302 **	0.031	0.303 **	0.032
			0.013	0.013	0.013	0.031	0.013	0.031
marital_status			-0.041 **	-0.027	-0.041 **	-0.027	-0.038 *	-0.014
			0.019	0.036	0.019	0.036	0.019	0.037
urban			0.271 **	0.084	0.267 **	0.084	0.264 **	0.082
			0.029	0.081	0.029	0.08	0.029	0.081
ln_cpi					1.532 **	0.274	1.529 **	0.219
					0.335	0.449	0.335	0.449
shock							-0.09 **	-0.15 **
							0.026	0.029
Year								
1997	0.5 **	0.501 **	0.54 **	0.518 **	0.093	0.439 **	0.102	0.469 **
	0.016	0.016	0.017	0.328	0.099	0.135	0.099	0.135
2000	0.994 **	0.982 **	1.02 **	1.005 **	0.458 **	0.905 **	0.462 **	0.932 **
	0.287	0.029	0.029	0.057	0.126	0.173	0.126	0.173
cons	10.494 **	10.936 **	8.158 **	8.676 **	0.512	7.316 **	0.57	7.306 **
	0.029	0.0215	0.163	0.517	1.688	2.303	1.689	2.299
individual fixed-effect	N	Y	N	Y	N	Y	N	Y
clustered	Y	Y	Y	Y	Y	Y	Y	Y
sigma_u	0.774	1.084	0.557	1.032	0.549	1.031	0.55	1.039
sigma_e	1.212	1.215	1.206	1.206	1.206	1.206	1.205	1.205
rho	0.289	0.445	0.176	0.423	0.172	0.422	0.172	0.426
N	13,845	13,845	13,742	13,742	13,742	13,742	13,739	13,739
group	4,615	4,615	4,615	4,615	4,615	4,615	4,615	4,615
F-Test	2,912.75	579.99	5,589.42	196.35	5,652.31	178.58	5,693.10	176.01

Table 11: Panel Regressions with Year Fixed Effects vs. TWFE

ln_inc1_m	(1)	(2)	(3)	(4)
(a) nonag_main	1.053 **	0.666 **	0.653 **	0.646 **
pooled & no year fixed effect	0.027	0.03	0.029	0.029
(b) nonag_main	1.071 **	0.654 **	0.656 **	0.651 **
pooled, year fixed effect & clustered	0.033	0.034	0.034	0.033
(c) nonag_main	0.966 **	0.634 **	0.636 **	0.629 **
panel, year fixed effect & clustered	0.035	0.035	0.035	0.035
(d) nonag_main	0.515 **	0.446 **	0.446 **	0.438 **
panel, TWFE & clustered	0.072	0.069	0.069	0.069
differences before & after controlling for individual fixed effect:				
(b) - (d)	**0.556 ** **	**0.208 ** **	**0.21 ** **	**0.213 ** **
	0.079	**0.077**	**0.077**	**0.076**
(c) - (d)	**0.451 ** **	**0.188 ** **	**0.19 ** **	**0.191 ** **
	0.080	0.077	0.077	0.077

Table 12: Estimated Selection Effects on the APG using TWFE

ln_inc1_m *non-agriculture*	(1)	(2)	(3)	ln_inc1_m *agriculture*	(1)	(2)	(3)
age	-0.001	-0.001	-0.001	age	-0.012 **	-0.011 **	-0.012 **
	0.002	0.002	0.002		0.002	0.002	0.002
gender	0.401 **	0.402 **	0.403 **	gender	0.513 **	0.504 **	0.502 **
	0.032	0.032	0.032		0.082	0.082	0.083
educlevel2	0.351 **	0.352 **	0.352 **	educlevel2	0.237 **	0.236 **	0.238 **
	0.013	0.013	0.013		0.035	0.035	0.034
marital_status	-0.08 **	-0.078 **	-0.077 **	marital_status	0.021	0.015	0.017
	0.214	0.021	0.021		0.038	0.038	0.038
urban	0.177 **	0.176 **	0.176 **	urban	0.262 **	0.27 **	0.269 **
	0.03	0.03	0.03		0.087	0.086	0.086
wagedwork_main	0.144 **	0.149 **	0.147 **	wagedwork_main	0.239 **	0.253 **	0.245 **
	0.034	0.034	0.034		0.069	0.069	0.069
ln_hrsworked_m	0.307 **	0.309 **	0.309 **	ln_hrsworked_m	0.435 **	0.435 **	0.432 **
	0.028	0.028	0.028		0.064	0.064	0.064
nfarmbiz	0.017	0.024	0.025	nfarmbiz	0.129 **	0.135 **	0.135 **
	0.026	0.026	0.026		0.064	0.064	0.065
farmbiz	-0.091 **	-0.083 **	-0.078 **	farmbiz	-0.134 *	-0.112	-0.094
	0.034	0.034	0.034		0.072	0.073	0.074
ruralborn	-0.105 **	-0.102 **	-0.103 **	ruralborn	-0.138	-0.122	-0.117
	0.029	0.029	0.029		0.108	0.107	0.107
ln_cpi	**	1.447 **	1.451 **	ln_cpi		2.599 **	2.629 **
		0.331	0.331			0.713	0.715
shock			-0.03	shock			-0.104 *
			0.026				0.056
wave				wave			
2	0.513 **	0.089	0.091	2	0.601 **	-0.154	-0.151
	0.018	0.098	0.099		0.038	0.209	0.209
3	0.99 **	0.463 **	0.462 **	3	1.044 **	0.075	0.069
	0.028	0.125	0.125		0.064	0.276	0.276
cons	8.796 **	1.567	1.553	cons	7.926 **	-5.048	-5.153
	0.166	1.666	1.665		0.38	3.629	0.209
clustered	Y	Y	Y	clustered	Y	Y	Y
N	8,867	8,867	8.867	N	4,875	4,875	4,875
R^2	0.333	0.335	0.335	R^2	0.113	0.116	0.117

Table 13: Exclusion Restriction Evaluation in Outcome Equations

nonag_main	(1)		ag_main	(2)	
age	-0.003	**	age	0.003	**
	0.0004			0.0004	
nfarmbiz	0.256	**	nfarmbiz	-0.256	**
	0.008			0.008	
gender	-0.175	**	gender	0.175	**
	0.01			0.01	
educlevel2	0.06	**	educlevel2	-0.06	**
	0.004			0.004	
marital_status	-0.022	**	marital_status	0.022	**
	0.006			0.006	
urban	0.159	**	urban	-0.159	**
	0.01			0.01	
wagedwork_main	0.153	**	wagedwork_main	-0.153	**
	0.01			0.01	
ln_hrsworked_m	0.05	**	ln_hrsworked_m	-0.05	**
	0.006			0.006	
farmbiz	-0.289	**	farmbiz	0.289	**
	0.01			0.01	
ruralborn	-0.016	*	ruralborn	0.016	*
	0.008			0.008	
wave			wave		
2	-0.0009		2	0.0009	
	0.0057			0.0057	
3	-0.028	**	3	0.028	**
	0.007			0.007	
cons	0.421	**	cons	0.579	**
	0.044			0.044	
clustered	Y		clustered	Y	
N	13,742		N	13,742	
R^2	0.463		R^2	0.463	

Table 14: Exclusion Restriction Evaluation in Selection Equations

ln_inc1_m non-agriculture	(1)	(2)	(3)	ln_inc1_m agriculture	(1)	(2)	(3)
Step 2 Outcome				*Step 2 Outcome*			
gender	0.444 **	0.445 **	0.446 **	gender	0.601 **	0.592 **	0.578 **
	0.029	0.029	0.029		0.093	0.093	0.093
educlevel2	0.339 **	0.339 **	0.339 **	educlevel2	0.195 **	0.197 **	0.203 **
	0.011	0.011	0.011		0.035	0.035	0.036
marital_status	-0.074 **	-0.072 **	-0.071 **	age	-0.011 **	-0.01 **	-0.01 **
	0.016	0.016	0.016		0.002	0.002	0.002
urban	0.151 **	0.146 **	0.144 **	urban	0.173 **	0.188 **	0.197 **
	0.03	0.03	0.03		0.094	0.094	0.094
wagedwork_main	0.127 **	0.127 **	0.125 **	wagedwork_main	0.182 **	0.193 **	0.187 **
	0.025	0.025	0.025		0.081	0.084	0.081
ln_hrsworked1_m	0.296 **	0.297 **	0.296 **	ln_hrsworked1_m	0.402 **	0.404 **	0.403 **
	0.021	0.027	0.021		0.051	0.05	0.05
ln_cpi		1.513 **	1.514 **	ln_cpi		2.601 **	2.627 **
		0.289	0.289			0.657	0.657
shock			-0.03	shock			-0.096 *
			0.024				0.052
wave				wave			
2	0.517 **	0.078	0.079	2	0.595 **	-0.159	-0.155
	0.027	0.089	0.089		0.063	0.201	0.201
3	0.984 **	0.438 **	0.437 **	3	1.062 **	0.093	0.089
	0.027	0.109	0.109		0.063	0.253	0.253
cons	8.888 **	1.345	1.349	cons	7.743 **	-5.216	-5.276
	0.134	1.449	1.449		0.283	3.287	3.286
Step 1: Selection				*Step 1: Selection*			
age	-0.014 **	-0.013 **	-0.014 **	age	0.01 **	0.01 **	0.01 **
	0.001	0.001	0.001		0.001	0.001	0.001
nfarmbiz	1.12 **	1.129 **	1.128 **	fambiz	1.109 **	1.11 **	1.103 **
	0.03	0.031	0.031		0.03	0.03	0.03
gender	-0.853 **	-0.853 **	-0.846 **	gender	0.904 **	0.904 **	0.903 **
	0.037	0.037	0.037		0.036	0.036	0.036
educlevel2	0.257 **	0.261 **	0.265 **	educlevel2	-0.309 **	-0.31 **	-0.311 **
	0.129	0.014	0.014		0.014	0.014	0.014
marital_status	-0.074 **	-0.074 **	-0.067 **	marital_status	0.134 **	0.134 **	0.132 **
	0.019	0.019	0.019		0.019	0.019	0.019
urban	0.933 **	0.934 **	0.926 **	urban	-0.689 **	-0.689 **	-0.689 **
	0.031	0.031	0.031		0.032	0.032	0.032
wagedwork_main	0.868 **	0.872 **	0.853 **	wagedwork_main	-0.087 **	-0.087 **	-0.084 **
	0.032	0.032	0.032		0.03	0.03	0.03
ln_hrsworked1_m	0.244 **	0.242 **	0.237 **	ln_hrsworked1_m	-0.195 **	-0.195 **	-0.194 **
	0.024	0.024	0.024		0.024	0.024	0.024
ruralborn	-0.261 **	-0.255 **	-0.252 **	ruralborn	0.201 **	0.201 **	0.2 **
	0.042	0.042	0.042		0.043	0.043	0.043
ln_cpi		-0.226 **	-0.201 **	ln_cpi		0.011 **	0.005
		0.086	0.087			0.085	0.085
shock			-0.182 **	shock			0.047 *
			0.028				0.029
cons	-0.979 **	0.149	0.099	cons	-0.523 **	-0.579	-0.561
	0.156	0.459	0.461		0.158	0.458	0.459
Inverse Mills Ratio				*Inverse Mills Ratio*			
λ	-0.138 **	-0.142 **	-0.143 **	λ	0.245 **	0.226 **	0.202 *
	0.052	0.052	0.052		0.105	0.105	0.106
rho	-0.133	-0.137	-0.138	rho	0.139	0.129	0.116
sigma	1.038	1.036	1.036	sigma	1.756	1.752	1.749
N	13,742	13,742	13,739	N	13,742	13,742	13,739
Selected	8,867	8,867	8,864	Selected	4,875	4,875	4,875
Wald chi2	3,777.43 **	3,868.35 **	3,867.37 **	Wald chi2	523.65 **	540.57 **	544.21 **

Exclusion Restriction (Nonag)
age
nfarmbiz

Exclusion Restrictions (Ag)
marital_status
farmbiz
ruralborn

Table 15: Heckman Two-Step Estimation (Pooled IFLS 1-3 Waves)

ln_inc1_m *pooled, year fixed effect, Heckman twostep*	(1)	(2)	(3)
nonag			
selection effect (λ^n)	-0.138 **	-0.142 **	-0.143 **
	0.052	0.052	0.052
ag			
selection effect (λ^a)	0.245 **	0.226 **	0.202 *
	0.105	0.105	0.106
observed APG	0.654 **	0.656 **	0.651 **
pooled, year fixed-effect & clustered	0.034	0.034	0.033
Selection effect explains observed APG			
nonag	***21.1%***	***21.6%***	***22.0%***
ag	***37.5%***	***34.5%***	***31.0%***

Table 16: Heckman Selection Effect on APG (Pooled)

xthechman	(1)	(2)
Outcome Equation		
ln_inc1_m		
gender	0.451 **	0.651 **
	0.041	0.033
educlevel2	0.375 **	0.267 **
	0.014	0.012
marital_status	-0.02	-0.067 **
	0.019	0.019
urban	0.204 **	-0.034
	0.042	0.03
wagedwork_main	0.056 *	-0.033
	0.029	0.028
lm_hrworked1_m	0.258 **	0.206 **
	0.023	0.021
ln_cpi		2.486 **
		0.059
cons	9.241 **	-2.723 **
	0.156	0.329
Selection Equation		
nonag_main		
age	-0.029 **	-0.024 **
	0.003	0.002
nfarmbiz	1.303 **	1.21 **
	0.054	0.046
gender	-1.319 **	-1.466 **
	0.094	0.075
educlevel2	0.413 **	0.357 **
	0.029	0.025
marital_status	-0.112 **	-0.117 **
	0.041	0.035
urban	1.787 **	1.553 **
	0.075	0.059
wagedwork_main	1.128 **	0.944 **
	0.058	0.051
ln_hrsworked1_m	0.347 **	0.312 **
	0.041	0.035
cons	-0.909 **	-0.197
	0.283	0.564
var(e.ln_inc1_m)	0.972	0.845
	0.019	0.015
var(ln_inc1_m[i])	0.266	0.412
	0.019	0.021
var(nonag_main[i])	2.45	2.314
	0.17	0.148
corr(e.nonag_main, e.ln_inc1_m)	-0.235 **	-0.973
	0.093	.
corr(nonag_main[i], ln_inc1_m[i])	0.166 **	-0.241 **
	0.08	0.148
N	13,742	13,742
Selected	8,867	8,867
# of groups	4,615	4,615
Wald chi2	1,602.49 **	3,124.52 **

Exclusion Restriction (Nonag)
age
nfarmbiz

***Note: Convergence not achieved in (1) and (2)**

Table 17: Panel Heckman Estimation using **xtheckman** (IFLS Waves 1–3)

A Appendix: Background on Suri's Empirical Approach

This appendix provides background on the empirical approach of Suri (2011), which this paper adapts to study agricultural productivity gaps. While the main text explains how the framework is modified for the APG context, this appendix reviews Suri's original application to technology adoption in Kenya and outlines the underlying logic of the correlated random coefficient (CRC) model. The purpose is to provide readers less familiar with this method with a clear understanding of its origins, intuition, and technical foundations. Readers already acquainted with Suri's work may skip directly to Section 3, where the adapted framework is presented.

To address the two challenges identified in the previous subsection, I adopt the Correlated Random Coefficient (CRC) model, as employed by Tavneet Suri (2011), when explaining the low adoption rates of hybrid seeds in Kenya, despite their high yields. This empirical approach allows me to estimate individual sorting based on the sector-specific unobserved abilities without imposing parametric distributional assumptions.

In Suri's empirical strategy, expected potential returns determine each farmer's adoption decision on hybrid seeds, which follows Heckman and Vytlacil (1998) under the generalized Roy's model framework (Roy, 1951); what's more, this method does not assume any functional form for unobserved heterogeity by exploring the fact that farmer's adoption choice history contains the information on farmer's net benefits from using hybrid seeds, which is in the spirit of Chamberlain's estimation of fixed-effect in panel data (Chamberlain, 1982, 1984). This framework consists of two key appealing features: First, it considers each farmer's net benefit as a deviation from the average net benefit of hybrid seed adoption, which explicitly models heterogeneity. Second, it exploits the revealed comparative advantages that can be projected by hybrid seeds adoption trajectories, which allows for estimating the selection effect without distributional assumptions. As a result, Suri's (2011) empirical approach is preferable for tackling the two challenges that the research question of this paper must overcome.

B Selection Effect and Map to the Classic Roy Model

This appendix derives the link between the CRC model selection parameter β and the selection terms in the classic Roy framework represented in Borjas (1987). To further illustrate what exactly the selection effect, β, is measured by this model, I first map it to different types of selection in the classic Roy model framework. Then, I discuss how the selection is determined by the variance of the latent skills in each sector and their correlations.

In this model, individual sorting based on the unobserved comparative advantage is summarized by the structural parameter β. Let $\sigma_n = Var(\theta_i^n)$, $\sigma_a = Var(\theta_i^a)$, and $\sigma_{na} = COV(\theta_i^n, \theta_i^a)$. Then, b_n and b_a as coefficients for equations (18) and (19) take form, as shown in equation (41). Therefore, in equation (42), the numerator $\sigma_n^2 - \sigma_{na}$ is the covariance-adjusted dispersion of latent absolute advantage in non-agriculture; the denominator $\sigma_{na} - \sigma_a^2$ is the analogous term for agriculture. Intuitively, the selection effect, β, measures the relative, covariance-adjusted dispersion of unobserved absolute advantages in non-agriculture relative to agriculture — i.e., how much more the non-agriculture sector loads on the latent skill variation once the skill correlation between the two sectors is netted out.

$$\begin{aligned}
\beta &\equiv \frac{b_n}{b_a} - 1 \\
&= \frac{(\sigma_n^2 - \sigma_{na})/(\sigma_n^2 + \sigma_a^2 - 2\sigma_{na})}{(\sigma_{na} - \sigma_a^2)/(\sigma_n^2 + \sigma_a^2 - 2\sigma_{na})} - 1 \qquad (41) \\
&= \frac{\sigma_n^2 - \sigma_{na}}{\sigma_{na} - \sigma_a^2} - 1 \qquad (42)
\end{aligned}$$

Under the assumption of a joint normal distribution for the latent skills, the inverse Mills ratio (IMR) can be derived as a selection-bias factor that summarizes the selectivity of the sample and adjusts the results by serving as a proxy for latent abilities. Therefore, the coefficient of the IMR in a regression represents the selection effect that arises when a joint-normal distribution is imposed on the latent skills.

Borjas (1987) studies the earnings and immigration choices in the United States and assumes a joint normal distribution for latent skills between two countries. Under the classic Roy's model framework, Borjas captures the selection effect and bias correction by Q_1 and Q_0, which are defined as the differential earnings between the average immigrants and the average in the country of destination (referred to as Country 1) and in the country of origin

(referred to as Country 0), respectively. As shown in the equations (43) and (44), $\frac{\phi(z)}{1-\Phi(z)}$ is IMR and $1/\sigma_\nu$ is the scaling factor for IMR as it is transformed to the standard normal when applying the closed-form solution. Hence, the selection effect in Borjas' paper is captured by the coefficients, $\sigma_0\sigma_1(\rho_{0,1} - \frac{\sigma_0}{\sigma_1})$ and $\sigma_0\sigma_1(\frac{\sigma_1}{\sigma_0} - \rho_{0,1})$ for the country of origin and destination, respectively.

In Borjas (1987), the Roy framework is applied to immigration: country 0 represents the origin country and country 1 the destination (United States). In my APG setting, these roles naturally map to agriculture (a) as the origin sector and non-agriculture (n) as the destination. Borjas defines two selection terms, Q_0 and Q_1, which measure differential earnings between immigrants and sectoral averages in the origin and destination, respectively:

$$Q_1 = \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho_{0,1}\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right) \tag{43}$$

$$Q_0 = \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\rho_{0,1} - \frac{\sigma_0}{\sigma_1}\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right) \tag{44}$$

These two coefficients can be further expressed as equations (45) and (46) by multiplying $\sigma_0\sigma_1$ to the terms within the first bracket in each equation (43) and (44), respectively. The country 0 in Borjas' paper corresponds to the agricultural sector, and country 1 represents the nonagricultural sector in my setting. Notably, β in my model includes the selection effect formulation in the classic Roy model. Instead of measuring selection in each sector separately, β refers to the differences in the selection effects between two sectors, with the selection effect in the agricultural sector serving as a benchmark.

$$\begin{aligned}\sigma_0\sigma_1\left(\frac{\sigma_1}{\sigma_0} - \rho_{0,1}\right) &= \sigma_1^2 - \rho_{0,1}\sigma_0\sigma_1 \\ &= \sigma_1^2 - \sigma_{0,1}\end{aligned} \tag{45}$$

$$\begin{aligned}\sigma_0\sigma_1\left(\rho_{0,1} - \frac{\sigma_0}{\sigma_1}\right) &= \rho_{0,1}\sigma_0\sigma_1 - \sigma_0^2 \\ &= \sigma_{0,1} - \sigma_0^2\end{aligned} \tag{46}$$

Analogously, I define two distribution-free counterparts in my model: Δ_a for agriculture and Δ_n for non-agriculture. These capture the differential earnings between sectoral switchers and the corresponding sectoral averages, just as Q_0 and Q_1 do in Borjas. The critical difference is that Δ_a and Δ_n do not rely on a joint-normality assumption on latent skills.

Instead, they are constructed directly from the Roy-style choice problem. Formal definitions and derivations are provided in Appendix C, but for interpretation it suffices to note that (Δ_a, Δ_n) serve as the natural analogues to (Q_0, Q_1) in a distribution-free setting.

Under a weak assumption on monotone sorting, the selection effects in this paper can be mapped into the positive selection, negative selection, and refugee cases illustrated in Borjas (1987). The derivation of the conditions for the different types of selection effect β in Appendix C. The summary of the conditions compared to those in Borjas (1987) is in Table 18

Table 18: Selection Types in Roy vs. CRC Model

Selection Type	**Roy Model (Borjas, 1987)**	**CRC (Lemieux, 1998)**
Positive Selection	$\rho_{0,1} > \frac{\sigma_0}{\sigma_1},\ \sigma_1 > \sigma_0$ $Q_0 > 0,\ Q_1 > 0$ In both upper tails	$\beta > 0 \iff \rho_{na} > \frac{\sigma_a}{\sigma_n},\ \sigma_n > \sigma_a$ $\Delta_a > 0,\ \Delta_n > 0$ In both upper tails
Negative Selection	$\rho_{0,1} < \frac{\sigma_0}{\sigma_1},\ \sigma_0 > \sigma_1$ $Q_0 < 0,\ Q_1 < 0$ In both lower tails	$-1 < \beta < 0 \implies \rho_{na} < \frac{\sigma_a}{\sigma_n}$ $\Delta_a < 0,\ \Delta_n < 0$ In both lower tails
Refugee Selection	$\rho_{0,1} < \min\left\{\frac{\sigma_0}{\sigma_1}, \frac{\sigma_1}{\sigma_0}\right\}$ $Q_0 < 0,\ Q_1 > 0$	$\beta < -1 \implies \rho_{na} < \frac{\sigma_a}{\sigma_n}$ $\Delta_a < 0,\ \Delta_n > 0$
Null on One-Side	No such case	$\beta = -1 \implies \rho_{na} < \frac{\sigma_a}{\sigma_n}$ $\Delta_a < 0,\ \Delta_n = 0$

Notes: In Borjas' paper, 1 refers to the country of destination, 0 to the country of origin. In this paper, n is non-agriculture and a is agriculture. Δ_a and Δ_n are the differential earnings between the average of those who choose to switch to non-agriculture and the average in agriculture and non-agriculture, respectively.

When $\beta > 0$, indicating positive selection—individuals are drawn from the upper tail of agriculture and fall in the upper tail of non-agriculture. However, when $\beta < 0$, there are three cases: (1) If $-1 < \beta < 0$, workers are drawn from the lower tail in agriculture and contribute to the lower tail in non-agriculture, hence negative selection. (2) If $\beta < -1$, it is drawn from the lower tail in agriculture, but those workers earn higher wages than the average workers in non-agriculture. This case would occur when the negative selection is

sufficiently large. (3) If $\beta = -1$, the switchers from agriculture earn the same as the average workers in non-agriculture, which is not in Borjas (1987).

Essentially, the selection effect and unobserved comparative advantages formulated by Lemieux (1998) capture the equivalent selection effect from unobserved heterogeneity that affects individuals' choices, as modelled in the classic Roy choice framework. The difference is that this selection effect is not the coefficient of IMR, thereby allowing for the flexibility to estimate underlying unobserved comparative advantages, as opposed to assuming a specific functional form.

C Appendix: Sector Choice and Differential Returns

This appendix lays out the sector choice individual faces in the Roy's model framework and then defines the differential returns sector switchers relative to the average earnings in each sector, Δ_n and Δ_a.

I start from the potential earnings representation in Section 3.3 and make the individual choice rule explicit under the Roy model choice framework. We then define the differential earnings objects Δ_n and Δ_a and show how they correspond to Borjas's Q_1 and Q_0.

Step 1. Potential earnings in each sector. Recall the log potential earnings for individual i in sector $j \in \{n, a\}$ at time t (see Eqs. (20)–(21)):

$$w_{it}^n = \delta_t^n \; + \; (1+\beta)\,\theta_i \; + \; \tau_i \; + \; X_{it}\gamma^n \; + \; \xi_{it}^n, \tag{26}$$

$$w_{it}^a = \delta_t^a \; + \; \theta_i \; + \; \tau_i \; + \; X_{it}\gamma^a \; + \; \xi_{it}^a. \tag{27}$$

Here, θ_i is the unobserved comparative advantage (sector-relevant, time-invariant), τ_i is the sector-invariant component (irrelevant for choice), β captures the relative loading of latent skills across sectors, X_{it} are observables, and ξ_{it}^j are transitory shocks with zero conditional mean.

Step 2. Sectoral choice rule. Individual i chooses non-agriculture ($D_{it} = 1$) iff $w_{it}^n \geq w_{it}^a$. Using (26)–(27):

$$D_{it} = 1 \quad \Longleftrightarrow \quad (\delta_t^n - \delta_t^a) \; + \; \beta\, \theta_i \; + \; X_{it}(\gamma^n - \gamma^a) \; + \; (\xi_{it}^n - \xi_{it}^a) \; \geq \; 0. \tag{47}$$

Conditional on observables, sorting is governed by the comparative advantage term $\beta\, \theta_i$; the common component τ_i cancels in the difference.

Step 3. Differential earnings objects. Define selection premia as differences between conditional and unconditional sector means:

$$\Delta_n \equiv E[w^n \mid D = 1, X] - E[w^n \mid X]\,, \tag{48}$$

$$\Delta_a \equiv E[w^a \mid D = 0, X] - E[w^a \mid X]\,. \tag{49}$$

Intuitively, Δ_n measures how the earnings of those who select into non-agriculture differ from the non-agriculture sector mean; Δ_a is the analogous object for agriculture.[12]

Using (26)–(27) and the choice rule (47), we can write

$$\Delta_n = (1 + \beta)\,\Big(E[\theta_i \mid D = 1, X] - E[\theta_i \mid X]\Big), \tag{50}$$

$$\Delta_a = \qquad\quad \Big(E[\theta_i \mid D = 0, X] - E[\theta_i \mid X]\Big). \tag{51}$$

Thus, both selection premia are linear in the *comparative advantage* component; β scales the non-agriculture premium relative to agriculture.

Step 4. Connection to Borjas (1987). In the Borjas–Roy migration setting, country 1 (destination) and 0 (origin) abilities are jointly normal with variances σ_1^2, σ_0^2 and covariance σ_{01}. The standard selection corrections are

$$Q_1 \;=\; \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho_{01}\right)\lambda(z), \qquad Q_0 \;=\; \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\rho_{01} - \frac{\sigma_0}{\sigma_1}\right)\lambda(z), \tag{52}$$

where $\lambda(z) = \phi(z)/[1 - \Phi(z)]$ is the inverse Mills ratio and $\sigma_\nu > 0$ is a scale from the latent index. The *coefficients on the IMR* are

$$\underbrace{\sigma_1^2 - \sigma_{01}}_{\text{destination (1)}}\,, \qquad \underbrace{\sigma_{01} - \sigma_0^2}_{\text{origin (0)}}. \tag{53}$$

[12] We suppress t and condition on X to lighten notation. The transitory shocks ξ_{it}^j integrate out by zero conditional mean.

Identify agriculture with origin $(0 \leftrightarrow a)$ and non-agriculture with destination $(1 \leftrightarrow n)$, so that

$$\sigma_1^2 - \sigma_{01} \;\longleftrightarrow\; \sigma_n^2 - \sigma_{na}, \qquad \sigma_{01} - \sigma_0^2 \;\longleftrightarrow\; \sigma_{na} - \sigma_a^2. \tag{54}$$

From Section 3.2, the CRC selection parameter satisfies

$$\beta + 1 \;=\; \frac{\sigma_n^2 - \sigma_{na}}{\sigma_{na} - \sigma_a^2}. \tag{55}$$

Comparing (53)–(54) with (55) yields the algebraic identity

$$\boxed{\beta + 1 \;=\; \frac{\text{IMR coefficient at destination (non-agriculture)}}{\text{IMR coefficient at origin (agriculture)}}} \tag{56}$$

up to a common positive scale that cancels in the ratio.[13]

Finally, observe that (50)–(51) are distribution-free analogues of Borjas's Q_1, Q_0:

$$\Delta_n \;\leftrightarrow\; Q_1, \qquad \Delta_a \;\leftrightarrow\; Q_0, \tag{57}$$

where the CRC framework replaces the IMR with conditional means of θ_i governed by the monotone selection rule (47).

D Appendix: Further Analysis of β

This appendix examines how the variance and correlation of latent abilities between sectors determine selection. I can rewrite β in equation (42) into the expression in equation (58). Define r as the ratio of the standard deviation of unobserved absolute advantages between two sectors, as in equation (59), which represents how widely spread the absolute advantages in nonagriculture are relative to those in agriculture. Then, β can be further expressed as in equation (60), where ρ is the correlation coefficient of the covariance of absolute advantages between agriculture and nonagriculture, taking values $-1 \leq \rho \leq 1$.

[13] Any multiplicative factors such as $1/\sigma_\nu$ and $\lambda(z)$ in (52) are common across Q_1, Q_0 conditional on the selection index and therefore cancel in the ratio.

$$\begin{aligned}\beta &= \frac{\sigma_n^2 - \sigma_{na}}{\sigma_{na} - \sigma_a^2} - 1 \\ &= \frac{(\sigma_n^2 - \sigma_{na}) - (\sigma_{na} - \sigma_a^2)}{\sigma_{na} - \sigma_a^2} \\ &= \frac{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}{\sigma_{na} - \sigma_a^2}\end{aligned} \tag{58}$$

$$r \equiv \frac{\sigma_n}{\sigma_a} \tag{59}$$

$$\beta = \frac{\sigma_n^2 + \sigma_a^2 - 2\sigma_{na}}{\sigma_{na} - \sigma_a^2} = \frac{r^2 + 1 - 2\rho r}{\rho r - 1} \tag{60}$$

$$\frac{\partial \beta}{\partial \rho} = \frac{r(1 - r^2)}{(\rho r - 1)^2} \tag{61}$$

Given a fixed r in the market, meaning how spread unobserved absolute advantages in nonagriculture relative to agriculture, the partial derivatives of β with respect to ρ, shown in equation (61), indicate four cases:

(i) If $r > 1$, then β decreases in ρ;

(ii) If $0 < r < 1$, then β increases in ρ;

(iii) If $r = 0$, then $\beta = -1$ for all feasible ρ;

(iv) If $r = 1$, then $\beta = -2$ for all feasible $\rho < 1$.

Since r is the ratio of standard deviations, its value will always be non-negative. In the case (i), this ratio of spread between latent absolute advantages is sufficiently large ($r > 1$). Holding r constant, as those latent skills become more similar across sectors (an increase in ρ), it dulls the comparative advantage in the respective sector that workers enjoy and weakens the selection effect; hence, β falls. In the case (ii), the dispersion of latent absolute advantages is alike in two sectors ($0 < r < 1$). For a fixed r, as the latent skills become more transferable across two sectors (a rising ρ), it raises the comparative advantage in the sector where workers don't possess and strengthens the selection effect; thus, β rises. The remaining

two cases are not very interesting. Both cases will make the selection effect degenerate into a constant number, where case (iii) is when absolute advantages in nonagriculture are without any dispersion, and case (iv) is when the spread of latent skills is precisely the same in two sectors.

Furthermore, the values of the selection effect β can provide some policy insights. When $\beta > 0$, the selection effect is bounded from below, $\beta \in [r-1, +\infty)$, which can only occur in the positive selection case. The minimum value of the selection effect is the differential spread of latent skills between sectors after removing the correlation coefficient at the upper boundary ($\rho = 1$). When $\beta < 0$, there are two cases: (i) If the spreads of latent skills across sectors are sufficiently large ($r > 1$), the selection flips to a negative value when passing the threshold, $\frac{1}{r}$. In this case, β is bounded from above, $\beta \in (-\infty, -(r+1)]$, where the least negative value is at two skills perfectly negatively correlated ($\rho = -1$). (ii) If the spreads of latent skills are similar ($0 < r < 1$), the selection is bounded, $\beta \in [-(r+1), r-1]$.

The threshold $\rho = \frac{1}{r}$ is where the positive and negative selection switches. In the environment where the spread in two sectors is sufficiently large, when the correlation of latent skills approaches the threshold, the selection effect becomes a large positive number on the right side and a large negative number on the left side. In the environment, the dispersions in latent skills are similar, and the selection effect is tightly bound by the correlation coefficient $\rho \in [-1, 1]$. Policies promoting transferable skills can influence the correlations, and interventions enhancing educational levels may bridge the disparities in latent skills between the two sectors. As the selection effect intensifies at the threshold, this insight can help design policies that align the selection effect with the main policy goals, thereby avoiding unintended consequences arising from the selection effect.

E APG Decomposition and the Roy Framing

This appendix derives the components of Agricultural Productivity Gap (APG) in the main text (expressed by equations (31) and (32)). This decomposition in the main text is not imposed ex ante but follows algebraically from aggregating the empirical model (Equation (23)) across sectors. In addition, it demonstrates algebraic equivalent object in the Roy

representation. The goal is two-fold: first, to provide the derivation of S_θ, the share of individual selection effect in APG, illustrated in the main text; second, to anchor the aggregation of selection effect to the classical Roy model framework.[14]

E.1 Notation and the main model

Recall the observed log earnings (suppressing t for exposition):

$$w_i = \delta^a + (\delta^n - \delta^a)D_i \;+\; \tau_i + \theta_i + \beta\,\theta_i D_i \;+\; X_i\gamma^a + X_i(\gamma^n - \gamma^a)D_i \;+\; \varepsilon_i, \tag{62}$$

where $D_i \in \{0,1\}$ indicates non-agriculture, τ_i is individual latent ability that is irrelevant to the sector choice, θ_i is unobserved comparative advantage that determines selection, and β is the *relative* loading of non-agriculture on θ_i.

APG defined in equation (30) can be express as

$$APG = \underbrace{E[w_i \mid D=1]}_{\text{avg. log earnings in non-ag}} - \underbrace{E[w_i \mid D=0]}_{\text{avg. log earnings in ag}} \tag{63}$$

E.2 Deriving the APG Expression

Take the expectation of (62) for each sector to obtain the sector average log earnings

$$\begin{aligned} E[w_i \mid D=1] = \delta^a + (\delta^n - \delta^a) \;+\; \underbrace{E[\tau_i \mid D=1]}_{=E[\tau_i]} + E[\theta_i \mid D=1] \;+\; \beta\, E[\theta_i \mid D=1] \\ +\; \gamma^a E[X_i \mid D=1] \;+\; (\gamma^n - \gamma^a)E[X_i \mid D=1] + \underbrace{E[\epsilon_i | D=1]}_{=0}, \end{aligned} \tag{64}$$

$$\begin{aligned} E[w_i \mid D=0] = \delta^a \;+\; \underbrace{E[\tau_i \mid D=0]}_{=E[\tau_i]} \\ +\; E[\theta_i \mid D=0] \;+\; \gamma^a E[X \mid D=0] + \underbrace{E[\epsilon_i \mid D=0]}_{=0}. \end{aligned} \tag{65}$$

[14] Appendix B links β to the Roy selection coefficients under joint normality; Appendix C defines distribution-free selection premia (Δ_n, Δ_a) and connects them to Borjas's (Q_1, Q_0).

Note that the conditional mean of τ_i equals its unconditional value due to its irrelevance to the sector choice. Moreover, the error terms follows conditional mean zero, shown in equations (64) and (65).

Then, subtracting equation (65) from (64) gives

$$\begin{aligned} APG = & \underbrace{\gamma^n E[X_i \mid D=1] - \gamma^a E[X_i \mid D=0]}_{\equiv \overline{X}(\hat{\gamma}^n - \hat{\gamma}^a)} + (\delta^n - \delta^a) \\ & + \beta\, E[\theta_i \mid D=1] + \big(E[\theta_i \mid D=1] - E[\theta_i \mid D=0]\big), \end{aligned} \tag{66}$$

Now, define the gap arising from observed characteristics between two sectors as $\overline{X}(\hat{\gamma^n} - \hat{\gamma^a})$, which is $APG_{observed}$. Hence, APG can be rewritten as:

$$\begin{aligned} APG = & \underbrace{\overline{X}(\hat{\gamma}^n - \hat{\gamma}^a)}_{APG_{\text{observed}}} + \underbrace{(\delta^n - \delta^a)}_{APG_\delta} \\ & + \underbrace{\beta\, E[\theta_i \mid D=1] + \big(E[\theta_i \mid D=1] - E[\theta_i \mid D=0]\big)}_{S_\theta}, \end{aligned} \tag{67}$$

$\theta^n - \theta^a$ is the sector-wide efficiency gap between two sectors, which is APG_δ. The impact of the unobserved comparative advantage is defined as S_θ, which consists two components: one is the extra returns to unobserved comparative advantages for those who choose non-agriculture, in relative to that in agriculture, that is $\beta E[\theta_i \mid D=1]$; the other is the how different on average in unobserved comparative advantage between workers in two sectors, that is $E[\theta_i \mid D=1] - E[\theta_i \mid D=0]$.

Hence, this gives the equations (31) and (32) in the main text:

$$APG = \underbrace{\overline{X}(\hat{\gamma}^n - \hat{\gamma}^a)}_{APG_{\text{observed}}} + \underbrace{\delta^n - \delta^a}_{APG_\delta} + S_\theta, \tag{31}$$

$$S_\theta \equiv \underbrace{\beta\, E[\theta_i \mid D=1]}_{\text{extra returns in non-ag}} + \underbrace{E[\theta_i \mid D=1] - E[\theta_i \mid D=0]}_{\text{mean diff. in } \theta}. \tag{32}$$

E.3 Mapping to the Roy parameterization

After showing the derivation of S_θ illustrated in the main text, I map this CRC model to the classic Roy model framework. One of the main differences between the CRC model and

the classical Roy model is that the selection term β in this paper is relative to agriculture. In contrast, the selection is measured for each sector.

Let the *raw* comparative advantage be

$$\tilde{\theta}_i \equiv \theta_i^n - \theta_i^a.$$

This is the comparative advantage before being scaled by b_a. In the CRC model, the unobserved comparative advantage is scaled by the agricultural loading:

$$\theta_i \equiv b_a\,\tilde{\theta}_i, \qquad b_n \equiv (1+\beta)\,b_a \iff 1+\beta = \frac{b_n}{b_a}.$$

In the classical Roy model, each sector has its own loading

$$(b_n, b_a)$$

on the same latent ability. The CRC normalization absorbs b_a into θ_i and estimates the relative loading $1+\beta$ directly. Substituting $\theta_i = b_a\tilde{\theta}_i$ into S_θ, equation (32), yields the Roy-form equivalence:

$$\begin{aligned}
S_\theta &= \beta E[b_a\tilde{\theta}_i \mid D=1] \;+\; E[b_a\tilde{\theta}_i \mid D=1] - E[b_a\tilde{\theta}_i \mid D=0] \\
&= \beta b_a E[\tilde{\theta}_i \mid D=1] \;+\; b_a\big(E[b_a\tilde{\theta}_i \mid D=1] - E[\tilde{\theta}_i \mid D=0]\big) \\
&= \underbrace{(1+\beta)b_a}_{b_n} E[\tilde{\theta}_i \mid D=1] \;-\; b_a E[\tilde{\theta}_i \mid D=0] \\
&= b_n\,E[\tilde{\theta}_i \mid D=1] - b_a\,E[\tilde{\theta}_i \mid D=0]. \qquad (68)
\end{aligned}$$

Equation (68) shows that the CRC decomposition is algebraically identical to the Roy representation once the normalization $\theta_i = b_a\tilde{\theta}_i$ and the relative loading $1+\beta = b_n/b_a$ are accounted for. In the CRC model, the agricultural coefficient is normalized into θ_i (equal to 1), and the non-ag coefficient is relative $(1+\beta)$.

If $b_a > 0$, a higher raw comparative advantage $\tilde{\theta}_i$ maps to a higher θ_i by scaling a b_a. It preserves each individual's ranking. If $b_a < 0$, the scaling of raw comparative advantage reversed the rank via a reflection about zero for each individual. In both cases, sectoral returns load θ_i as 1 in agriculture and $1+\beta$ in non-agriculture, so $\beta\theta_i$ is the selection premium. The mapping to the Roy model shown in Equation (68) remains valid. Now, I'll demonstrate this difference in normalization using both a positive and a negative b_a in a classical Roy model framework.

E.4 Visualization of the normalization

As Equation (68) shows, S_θ is equivalent to the difference in conditional group means, I can take advantage of the tractability of the Roy model to visually present this normalization by using a joint-normal distribution on the latent abilities. Appendix B shows that the CRC model can map to the three types of selection in the classical Roy model.

Figure 10 visualizes the normalization in a positive selection case in the Roy model. In this case, $b_a > 0$, $b_n > 0$ and $\beta > 0$. I draw two normal distributions with mean 0 and 0.5 for agriculture and non-agriculture sector, respectively. The standard variations are 0.8 and 1.2, with a wider dispersion in the non-agriculture. The absolute advantages in two sectors in correlated, $\rho = Corr(\theta_i^n, \theta_i^a) = 0.75$.

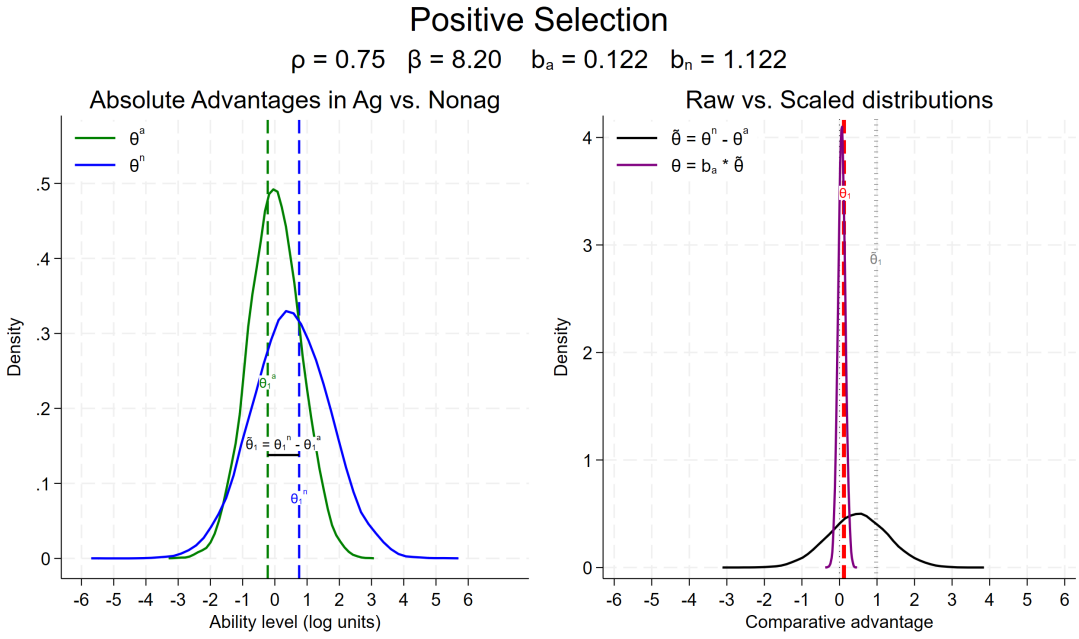

Figure 10: **Raw vs. Scaled Comparative Advantage** When $b_a > 0$ preserves the ranking of unobserved comparative advantage for each individual. The dispersions of absolute advantages and their correlation determine the location and compression of the scaled comparative advantage θ_i

In this figure, the left panel shows sector-specific latent abilities (θ^a, θ^n), with the blue line for non-agriculture and the green line for agriculture. Pick an individual, call this person no. 1, her absolute advantage in the non-agriculture is labelled by the blue dashed line, while the green dashed line represents her agricultural ability ranking. The difference is

the raw comparative advantage for person 1, $\tilde{\theta_1} = \theta_1^n - \theta_1^a$, which is the distance between two dash lines. Then, I can plot this difference for each individual on the right panel. The right panel shows the raw comparative advantage $\tilde{\theta_i} = \theta_i^n - \theta_i^a$ (black) and the scaled comparative advantages $\theta_i = b_a\tilde{\theta_i}$ (magenta). The dotted black line labels the rank of $\tilde{\theta_1}$, and dash magenta line shows where person 1 stands in the scaled distribution. The location and compression of the scaled distributions is jointly determined by the dispersion of the two latent skills (i.e. σ_n, σ_a and ρ). Note that when $b_a > 0$, it preserves the rank for each individual.

Now, what if $b_a < 0$? I accomplish this by simply change the $\rho < \frac{\sigma_a}{\sigma_n}$ (See the full details on why this switches sign of b_a in Appendix B). To generate $b_a < 0$, I set $\rho = 0.4 (< \frac{\sigma_a}{\sigma_n}$, holding means and variances fixed. This is refugee case in the classical Roy model, $b_n > 0, b_a < 0$ and $\beta < -1$.

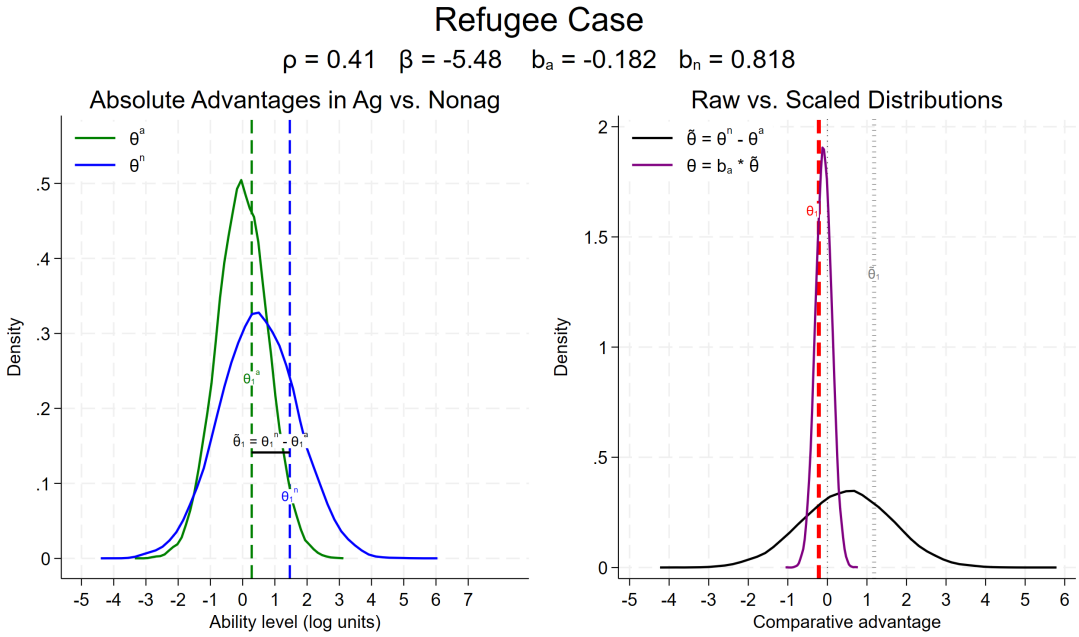

Figure 11: **Raw vs. Scaled Comparative Advantage** When $b_a < 0$ changes the ranking of unobserved comparative advantage for each individual via a reflection around zero. $\tilde{\theta_1}$ is on the right side of zero in the raw comparative advantage; whereas the scaled θ_1 is on the left side of the zero.

Figure 11 shows a refugee case, where $b_n > 0$ and $b_a < 0$. On the right panel, the raw comparative advantage $\tilde{\theta_i}$ represents in black line. The dotted black line represent the

location of $\tilde{\theta}_1$, which is at the right side of the mean in the raw comparative advantage. Since $b_a < 0$, the θ_i is symmetrically reflected around the zero in the scaled comparative advantage θ_i distribution.

In the negative selection, where $b_n < 0, b_a < 0$, the ranking of individual is also a reflection around zero in the scaled comparative advantage. See Figure 12.

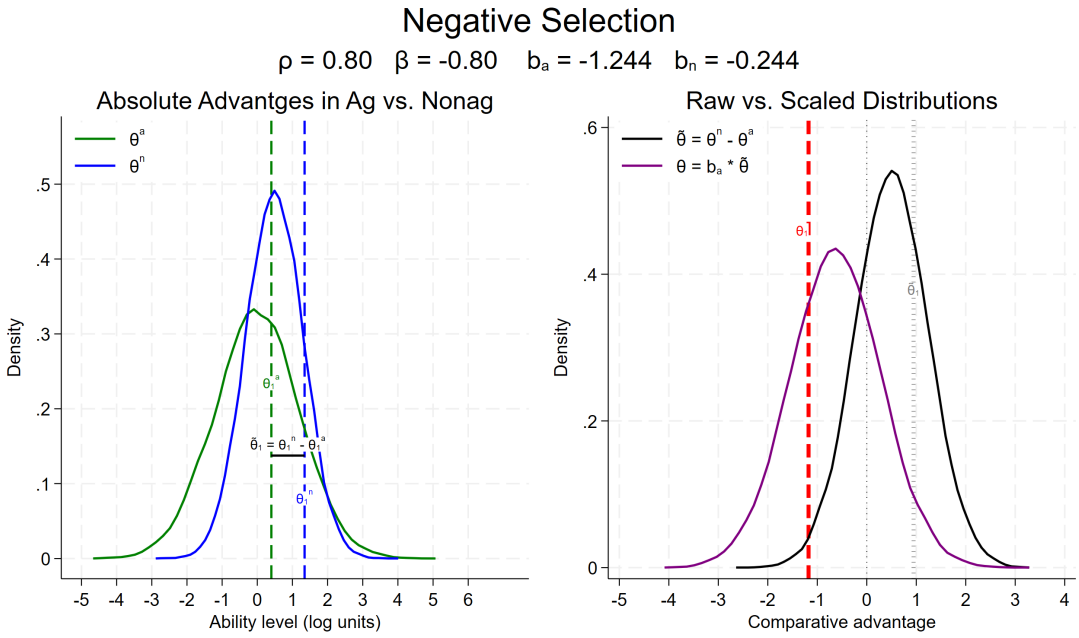

Figure 12: **Raw vs. Scaled Comparative Advantage** This is a negative selection case, where $b_n < 0, b_a < 0$. The ranking of individual in the scaled comparative advantage follows the same pattern as refugee case.

Regardless of $b_a > 0$ or $b_a < 0$, the scaled comparative advantage loads the return in agriculture sector as 1 (shown in Equation (19), and loade the return of unobserved comparative advantage in non-agriculture as $1 + \beta$ (shown in Equation (18). Hence, the selection term in CRC model β is the relative differential returns of latent abilities, which aligns with the mathematical form in Equation (17) and how it relates to the selection terms in the classical Roy model in Appendix B.

Take together, this section has shown the step-by-step derivation of the APG decomposition illustrated in subsection 3.4 in the Empirical Model, which are expressed by Equations (31) and (32). To anchor the aggregation expression, Equation (68) show the equivalent object in the classical Roy model framework. Moreover, the visualization of the three typical

selection cases illustrates how raw and scaled comparative advantages are related. I show the figures in a join-normal distribution on sector-specific latent abilities. Importantly, although the visualization assumes joint normality for clarity, the adapted CRC framework itself does not rely on this assumption. θ_i is estimated directly as a structural parameter.

F Appendix: Recovering Structural Parameters

This appendix is to demonstrate how structure parameter unobserved comparative advantage, θ_i, is recovered in a simplified two-period no covariant model. θ_i is unobserved in the data. I will, now, demonstrate how the structural parameters of interest can be recovered without imposing distributional assumptions. Following the procedure proposed by Suri (2011), I present the recovery of structural parameters in the simplest setting, without covariates, over two periods, as expressed in equation (69).

$$w_{it} = \eta + \alpha D_{it} + \theta_i + \beta \theta_i D_{it} + u_{it} \tag{69}$$

where $\delta_t^a = \eta \quad \forall t$, $\alpha \equiv \delta_t^n - \delta_t^a \quad \forall t$, and $u_{it} \equiv \tau_i + \epsilon_{it}$.

I can do this because structural parameters β and θ_i do not enter covariates in the main estimation equation (23). First, I disentangle the dependency between θ_i and D_{it} by linearly projecting θ_i onto the entire history of sector choices and their interaction terms. This method was first developed by Chamberlain (Chamberlain, 1982, 1984) to estimate individual unobserved fixed effects in panel data. Later, Suri's (2011) generalized Chamberlain's fixed-effect estimation by including interactions of choice histories to purge the dependency between unobserved abilities and individual choices fully. Chamberlain and Suri treat this as a purely technical step to separate θ_i into two parts: one is related to sectoral choice (D_{it}), and the remainder is orthogonal to the sectoral choices. Since individuals choose, D_{it}, is a dummy variable, the projection θ_i onto a complete history and interaction terms will be a saturated model to purge the correlation between θ_i and D_{it}, which is formally expressed in the equation (70). However, I interpret the equation (70) as an individual choice trajectories that reveal her unobserved comparative advantages, θ_i.

$$\theta_i = \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i1} D_{i2} + \nu_i \tag{70}$$

Next, I substitute equation (70) into the wage equation (69) to obtain the log earnings for each period as a function of choice histories and their interactions, see equations (71) and (72).

$$\begin{aligned} w_{i1} = {} & (\eta + \lambda_0) + (\lambda_1(1+\beta) + \alpha + \beta\lambda_0) D_{i1} \\ & + \lambda_2 D_{i2} + (\lambda_3(1+\beta) + \beta\lambda_2) D_{i1} D_{i2} + (\nu_i + \beta\nu_i D_{i1} + u_{i1}) \end{aligned} \tag{71}$$

$$\begin{aligned} w_{i2} = {} & (\eta + \lambda_0) + \lambda_1 D_{i1} \\ & + (\lambda_2(1+\beta) + \alpha + \beta\lambda_0) D_{i2} + (\lambda_3(1+\beta) + \beta\lambda_1) D_{i1} D_{i2} \\ & + (\nu_i + \beta\nu_i D_{i2} + u_{i2}) \end{aligned} \tag{72}$$

Since sector choices are observed in each period, I can run a reduced-form regression of earnings on the choice history and their interactions in this stacked system of equations (71) and (72). To simplify the coefficients in the reduced form regression in equations (71) and (72), I can rewrite them as equations (73) and (74).

$$w_{i1} = \eta_1 + \phi_1 D_{i1} + \phi_2 D_{i2} + \phi_3 D_{i1} D_{i2} + e_{i1} \tag{73}$$

$$w_{i2} = \eta_2 + \phi_4 D_{i1} + \phi_5 D_{i2} + \phi_6 D_{i1} D_{i2} + e_{i2} \tag{74}$$

The reduced form regression of each period earnings on the entire history of the sector choice, including interaction terms across periods, will obtain reduced form coefficients ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_5, and ϕ_6. Combining equations (71) to (74), the information that I have learned from the reduced form coefficients can yield the following equations:

$$\phi_1 = \lambda_1(1+\beta) + \alpha + \beta\lambda_0 \tag{75}$$

$$\phi_2 = \lambda_2 \tag{76}$$

$$\phi_3 = \lambda_3(1+\beta) + \beta\lambda_2 \tag{77}$$

$$\phi_4 = \lambda_1 \tag{78}$$

$$\phi_5 = \lambda_2(1+\beta) + \alpha + \beta\lambda_0 \tag{79}$$

$$\phi_6 = \lambda_3(1+\beta) + \beta\lambda_1 \tag{80}$$

Equations (75) to (80) explicitly express the reduced-form coefficients as functions of underlying structural parameters. Solving this system of equations, I can estimate five underlying parameters: λ_1, λ_2, λ_3, α, and β. Under the condition that $\lambda_1 \neq \lambda_2$, the parameter β is identified. The first objective is to obtain the structural parameter β, which captures the extra returns of comparative advantages in nonagriculture, i.e., the selection effect. Then, the second task is to estimate the distribution of comparative advantage θ_i by using equation (70) and normalizing $\sum \theta_i = 0$. Specifically, I can obtain λ_0 by using $\lambda_0 = -\lambda_1\overline{D_{i1}} - \lambda_2\overline{D_{i2}} - \lambda_3\overline{D_{i1}D_{i2}}$. It is noted that λ_1, λ_2, and λ_3 can be obtained by solving the system of equations, and sectoral choices are observed in the data. Once λ_0 is available, I can use (70), λ's, and D_{it}'s to estimate the distribution of the unobserved comparative advantage, θ_i.

G Appendix: Revealed Comparative Advantages

This appendix shows that the projection of θ_i can be interpreted as revealed comparative advantages, which contain rich information empirically in the context of Indonesia.

Although Chamberlain (1982, 1984) and Suri (2011) explicitly emphasize that equation (70) is primarily a technical device for eliminating correlation between θ_i and choice variable D_{it}, it can equivalently be interpreted as a regression of the latent comparative advantage θ_i on indicators of the choice trajectory (choices at each t and their interaction). In this formulation, the fitted values $\hat{\theta}_i$ represent the component of individual underlying comparative advantage explained by the trajectory, thereby providing an empirical measure of unobserved heterogeneity across groups.

Drawing loosely on the intuition of revealed preference theory (Samuelson, 1938, 1948), sectoral choices can be interpreted as revealing information about underlying comparative advantages. In this context, individuals implicitly conduct a cost–benefit analysis, where potential earnings in each sector represent are the benefits, and constraints such as schooling, time, and ability represent the costs. Over three waves, each agent's sequence of sectoral choices yields one of eight possible trajectories ($2^3 = 8$), reflecting how unobserved comparative advantages shape decisions. Although this analogy to revealed preference is only suggestive rather than a formal extension, the observed choice histories and their interactions provide an empirical basis for capturing latent abilities. I therefore refer to the fitted values from this procedure, $\hat{\theta}_i$, as revealed comparative advantages.

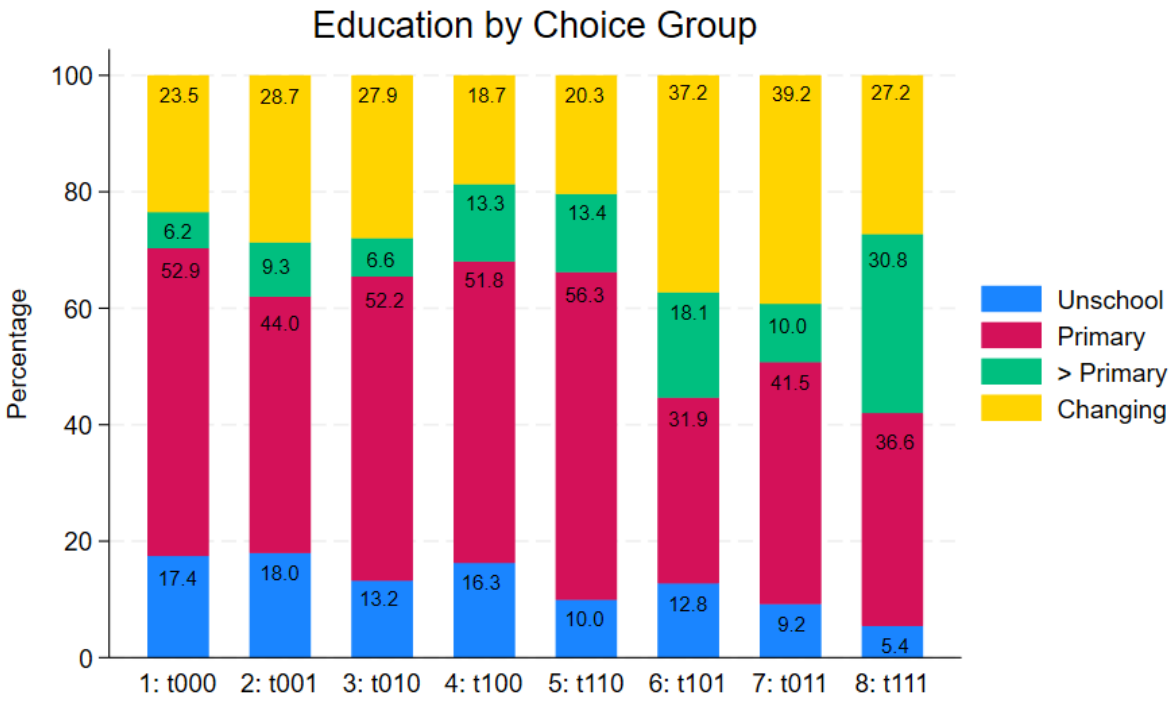

Figure 13

Figures 13 and 14 illustrate the composition of education levels and waged work across the eight trajectory groups, denoted by $t000, t001, t010, t100, t110, t101, t011, t111$, where 1 indicates non-agriculture and 0 indicates agriculture in a given period. For example, $t000$ corresponds to staying in agriculture in all three waves, while $t111$ corresponds to remaining in non-agriculture. Figure 13 shows clear contrasts: non-agricultural stayers ($t111$) are

disproportionately drawn from individuals with education beyond primary school, whereas agricultural stayers ($t000$) contain a higher share of unschooled individuals. Sector switchers, by contrast, display larger shifts in educational attainment across waves. A parallel pattern emerges in Figure 14: switchers exhibit greater transitions between self-employment and waged work, while stayers tend to remain more stable in their employment types.

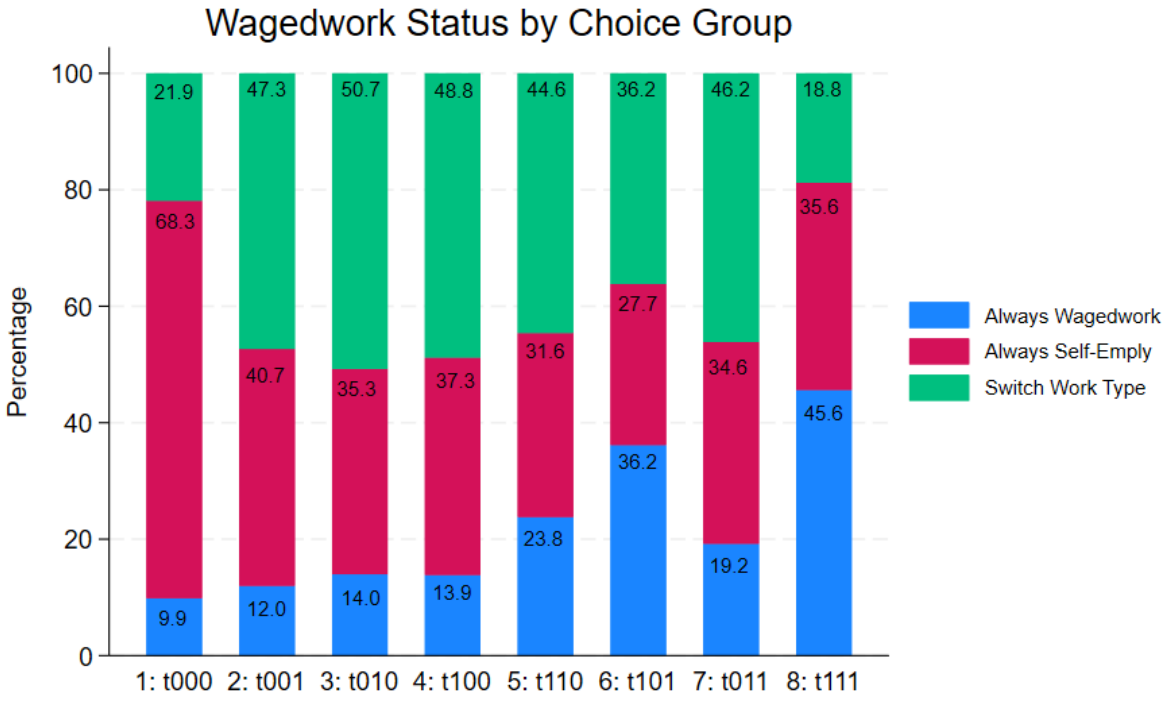

Figure 14

Figures 15 and 16 provide additional evidence that choice trajectories reveal information about underlying comparative advantages. A well-documented puzzle in the APG literature using IFLS data is that individuals who move from non-agriculture to agriculture appear to experience substantial earnings losses (Pulido and Świecki, 2019; Hamory et al., 2021). When earnings are instead examined by trajectory groups, as shown in Figure 15, the distributions of log earnings for all groups shift to the right over time, indicating earnings growth for each trajectory group. This perspective suggests that trajectory groups differ in their initial mean earnings, reflecting underlying abilities and costs across sectors. When outcomes are aggregated to sector-level averages, these initial differences are masked, creating the appearance of earnings losses that are not evident once trajectories are taken into account.

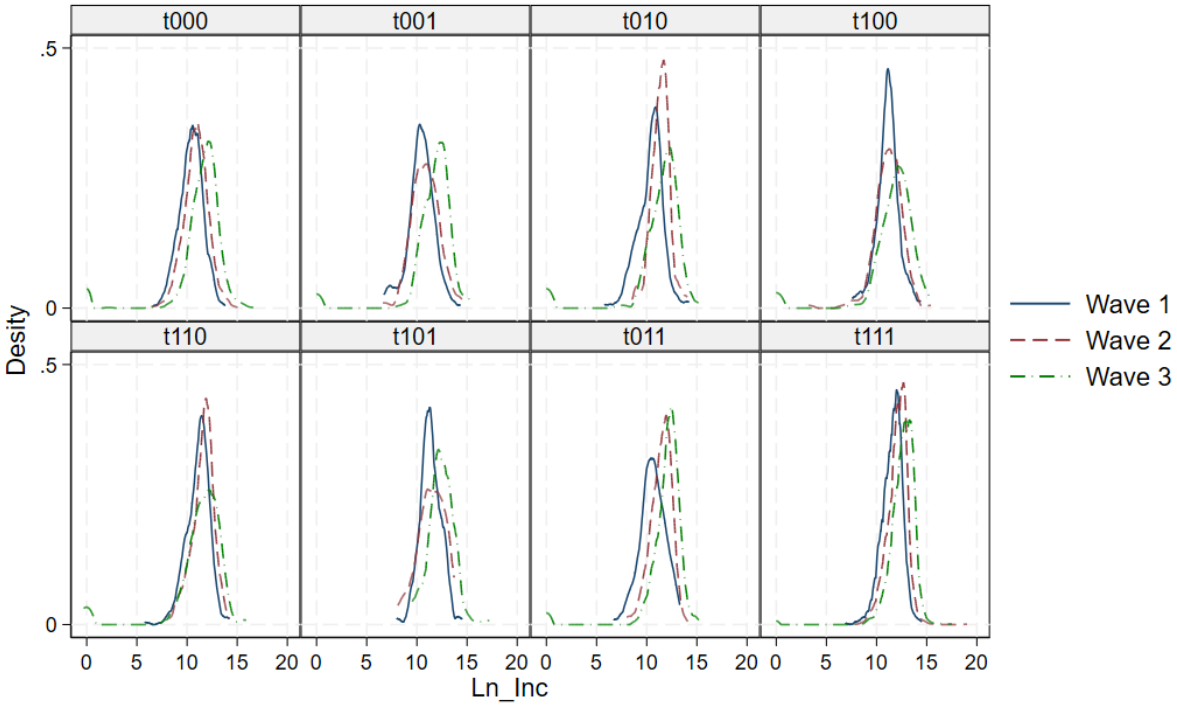

Figure 15: Log Earnings Distribution in 3 Waves by Choice Trajectory

Figure 16 further shows that hours worked remain relatively stable for stayers but fluctuate for switchers. Taken together, these results suggest that trajectory groups capture systematic heterogeneity consistent with comparative advantages on which sectoral choices are made.

ht

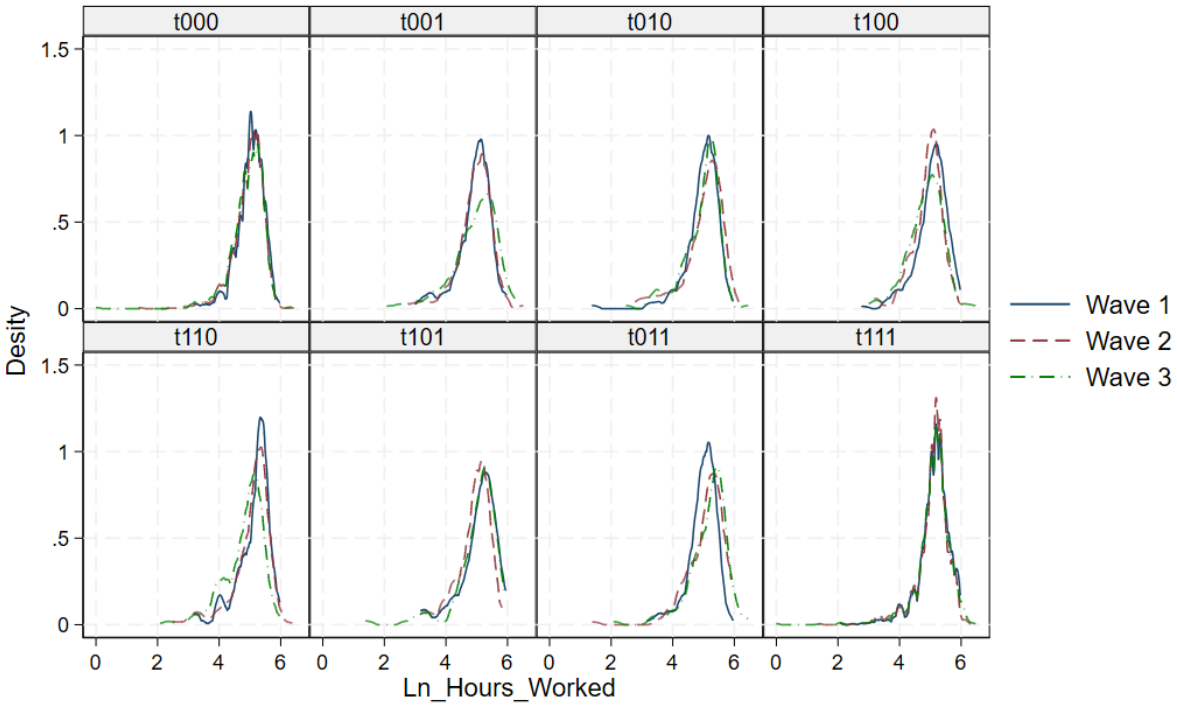

Figure 16: Log Hours Worked Distribution in 3 Waves by Choice Trajectory

H Appendix: Model Structures and Estimation Details

H.1 Heckman Two-step Estimation (Pooled Panel)

The canonical Heckman two-step estimator is designed to correct for selection bias in cross-sectional data. To evaluate the implications of parametric assumptions on the selection effect in the APG context, I estimate the selection-corrected wage equations under the assumption of joint normality in unobserved sectoral abilities using the Heckman two-step method (Heckman, 1979).

This approach, when applied to pooled panel data from the first three waves of IFLS (used in this paper), reveals a statistically significant selection effect. Recall the main result in the present paper: the selection effect due to comparative advantages at the individual level does not impact the sectoral productivity gap significantly when avoiding such a functional form assumption. By imposing a joint normal distribution on the latent skills, the pool sample from the same dataset finds that individual comparative advantages explain a significant portion of the sectoral productivity gap, which is consistent with the finding in Pulido and Świecki (2019) with the same distributional assumptions.

Abstracting from the time dimension, the Roy model framework is based on the assumption that an individual i has potential earnings y_i^n and y_i^a in the non-agricultural sector (n) and the agricultural sector (a), respectively.

$$y_i^n = X_i\gamma^n + \epsilon_i^n \tag{81}$$

$$y_i^a = X_i\gamma^a + \epsilon_i^a \tag{82}$$

This setup mirrors the Roy model and is structurally comparable to the potential outcomes framework in Section 3. However, the estimation strategy differs: Heckman's method treats sectoral choice as endogenous and estimates the selection correction term explicitly, while the main framework in this paper treats sectoral choice as informative of comparative advantage and avoids strong distributional assumptions

In Heckman's estimation, for an individual who chooses to work in the non-agricultural sector, the selection equation governing sectoral choice is given by:

$$D_i = \begin{cases} 1 & \text{if} \quad Pr(y_i^n > y_i^a | X_i = x) \\ 0 & \text{otherwise} \end{cases} \tag{83}$$

Therefore, the probability of choosing the nonagricultural sector (n) follows:

$$\begin{aligned} Pr(D_i = 1 | X_i = x) &= Pr(y_i^n > y_i^a | X_i = x) \\ &= Pr(X_i\gamma^n + \epsilon_i^n > X_i\gamma^a + \epsilon_i^a) \\ &= Pr(\epsilon_i^a - \epsilon_i^n < X_i(\gamma^n - \gamma^a)) \end{aligned} \tag{84}$$

Assuming sector-specific unobserved abilities ϵ_i^n and ϵ_i^a are jointly normally distributed, the difference in unobserved abilities $\epsilon_i^n - \epsilon_i^a$ also follows a normal distribution. The resulting selection equation can be estimated as a probit model. Let $\hat{\gamma}$ denote the estimated coefficients from this probit regression of the sectoral choice equation. Then, IMR for individual i, denoted λ_i), is given by:

$$\lambda_i = \frac{\phi(X_i'\hat{\gamma})}{\Phi(X_i'\hat{\gamma})} \tag{85}$$

where $\phi(.)$ and $\Phi(.)$ refer to the standard normal density and Cumulative Density Function (CDF), respectively. Note that the derivation of IMR is well known and can refer to Heckman's seminal paper (1979). The IMR enters the outcome equation to correct for selection bias as follows:

$$y_i = X_i\gamma^n + \lambda_i\beta + u_i \quad \text{for} \quad D_i = 1 \tag{86}$$

In this outcome equation (86), β captures the direction and magnitude of selection bias, and y_i is an individual's observed wage working in the non-agricultural sectors, which is only observed for those with $D_i = 1$. X_i is a vector of observed characteristics, and λ_i is the estimated IMRs coming out of the selection equation. β captures direction and magnitude of selection bias.

To bolster identification, I incorporate exclusion restrictions Z_i in the selection equations but omit them from the outcome equation. This yields the augmented selection equation (87):

$$\begin{aligned} Pr(D_i = 1|X_i = x) &= Pr(y_i^n > y_i^a|X_i = x, Z_i = z) \\ &= Pr(X_i\gamma^n + Z_i^n\gamma_z^n + \epsilon_i^n > X_i\gamma^a + Z_i^a\gamma_z^a + \epsilon_i^a) \\ &= Pr(\epsilon_i^a - \epsilon_i^n < Z_i^n\gamma_z^n - Z_i^a\gamma_z^a + X_i(\gamma^n - \gamma^a)) \end{aligned} \tag{87}$$

In practice, the first-stage probit regression uses Z_i as exclusion restrictions, the variables that affect sectoral choice but are assumed not to directly influence wages. For non-agricultural wage estimation, valid exclusion restrictions include age and non-farm business ownership. For the agricultural sector, the exclusion restrictions are rural-born, marital status, and farm business ownership. Valid exclusion restrictions are supported empirically (see Tables 13 and 14). The three regressions on the left of Table 13 show that the two variables, age and non-farm business, are not significant in the outcome equation for the non-agricultural workers; the left regression on Table 14 indicate they are both highly relevant to the sectoral decision in the selection equation. The evidence for exclusion restrictions for agriculture is on the right side of Tables 13 and 14.

Columns (1) to (3) in Table 15 present the Heckman two-step estimates for the non-agricultural sector. The table is divided into three blocks: (i) outcome equation estimates, (ii) selection equation estimates, and (iii) IMR coefficients. Column (1) includes basic covariates; column (2) adds log CPI; and column (3) includes both log CPI and a shock variable. All regressions include time fixed effects. Across all specifications, the IMR coefficient is negative and statistically significant, indicating adverse selection into the non-agricultural sector. The estimated selection effects are -0.138, -0.142, and -0.143, respectively. Selection effects for agricultural workers are estimated in a similar manner. The right block of Table 15, labelled as agriculture, shows corresponding regressions for agricultural workers under specifications (1) to (3), mirroring those used in the non-agricultural sector. In all three specifications, the IMR coefficients are positive and statistically significant, ranging from 0.202 to 0.245.

The wage for the agricultural workers can also be observed in the data. Similarly, the selection effect for the agricultural workers can be estimated. The right side of Table 13 and

Table 14 demonstrate that three variables, marital status, farm business, and rural born, are suitable exclusion restrictions for the agricultural sector. Table 15 presents the estimation of the Heckman selection correction for the agricultural workers. The three columns on the right side of the table correspond to three specifications used in the non-agriculture sector.

On average, the IMR coefficient for agricultural workers is positive and statistically significant, with estimated values of 0.245, 0.226, and 0.202 across specifications (Table 15). Taken together, under the assumption of bivariate normality in unobserved abilities, the selection effect in the non-agricultural sector is negative and statistically significant (ranging from -0.138 to -0.143), while that in the agricultural sector is positive and significant (ranging from 0.202 to 0.245). These magnitudes remain stable across model specifications, both with and without additional controls (log CPI and shock).

To benchmark these magnitudes, Column (3) of Table 10 estimates an average productivity gap of 0.654 using a pooled OLS regression with the same covariates as column (1) of Table 15. This implies that selection effects, as captured by the Heckman two-step estimator, explain approximately 21.1% to 37.5% of the observed APG in the baseline specification, and between 22% and 31% in the most complete specification (see Table 16).

This finding echoes the conclusion in Pulido and Świecki (2019), who report substantial selection effects under joint normality. In contrast, the empirical strategy employed in this paper imposes no functional form assumptions on unobserved heterogeneity and finds no significant selection effect. This divergence underscores the sensitivity of selection estimates to distributional assumptions.

H.2 Heckman Selection Estimation with Panel Structure (xtheckman)

To evaluate whether panel structure affects the magnitude or direction of selection estimates under joint normality, I estimate the selection-corrected wage equation using Stata's **xtheckman** command. This approach extends the Heckman framework to panel data and fits a full maximum likelihood model accounting for both unobserved individual heterogeneity and time-varying error components.

Under the panel Roy model, individual i's potential log earnings in sector $s \in \{n, a\}$ at time t are given by:

$$y_{it}^n = X_{it}\gamma^n + \theta_i^n + \epsilon_{it}^n \tag{88}$$

$$y_{it}^a = X_{it}\gamma^a + \theta_i^a + \epsilon_{it}^a \tag{89}$$

where (X_{it}) are observed characteristics, (θ_i^s) are unobserved time-invariant individual sector-specific abilities, and (ϵ_{it}^s) are time-varying unobserved shocks. The model assumes that both (θ_i^n, θ_i^a) and $(\epsilon_{it}^n, \epsilon_{it}^a)$ follow a joint normal distribution across sectors.

Sectoral choice at each period is modelled by a latent selection equation:

$$D_{it} = \begin{cases} 1 & \text{if } y_{it}^n > y_{it}^a \\ 0 & \text{otherwise} \end{cases} \tag{90}$$

Observed wages for those working in the non-agricultural sector ($D_{it} = 1$) are:

$$y_{it} = X_{it}\gamma^n + \theta_i^n + \epsilon_{it}^n \tag{91}$$

Unlike the pooled two-step Heckman estimator, **xtheckman** does not report estimated Inverse Mills Ratios (IMRs) or directly quantify the magnitude of selection effects. Instead, it provides estimated correlations across unobserved components of the outcome and selection equations, which imply the presence and direction of selection bias.

Table 17 reports results from the **xtheckman** estimation for non-agricultural workers. Two specifications are presented, mirroring columns (1) and (2) of Table 15. The model includes valid exclusion restrictions in the selection equation: age and non-farm business ownership. While estimation is computationally demanding, and convergence was not achieved in either specification, the results still reveal significant correlation in unobserved components, consistent with selection.

Despite limitations—long run times, convergence issues, and lack of explicit IMR estimates—the results from **xtheckman** reinforce the key message: under the joint-normality assumption, significant selection effects are recovered even in a panel framework. These findings align with those from the pooled Heckman model and underscore the influence of

distributional assumptions on empirical conclusions regarding comparative advantage and selection in the agricultural productivity gap.

H.3 Control Function Approach for Panel Data

To further understand the impact of the distributional assumptions on the estimation results, I deploy the panel selection bias correction developed by Wooldridge (1995) to the first three waves of the IFLS dataset. This method takes into account the panel structure and uses a control function with a weaker assumption regarding the unobserved heterogeneity. Under the same Roy model framework in the panel structure, the key departure of Wooldridge (1995) from **xtheckman** is that the distribution of unobserved individual abilities remains unspecified; at the same time, IMRs are calculated for individual i and each period t. Under the assumption that the error term in the outcome equation satisfies the conditional mean assumption (see Equation (96), Wooldridge (1995) exploits the panel structure by regressing the de-meaned log earnings on de-meaned regressors (the mathematical form expressed in (99)), including the transformed IMRs for selection bias correction.

In this method, each period latent choice variable D_{it}^* follows (92). Instead, assuming a bivariate normal distribution of unobserved abilities between sectors for individuals, the selection equation (92) assumes that the error term ν_{it} is independent of $\mathbf{x_i}$ and normally distributed, which allows for calculating IMRs for each period. In Equation (93), $\mathbf{x_i'}$ refers to a vector of observed characteristics, including 1 for the intercept, δ_{t0} in Equation (92). Therefore, Equation (93) is a shorthand expression of Equation (92) using a vector form.

$$D_{it}^* = \delta_{t0} + x_{i1}\delta_{t1} + ... + x_{iT}\delta_{tT} + \nu_{it} \tag{92}$$

$$D_{it}^* = \mathbf{x_i'}\delta_\mathbf{t} + \nu_{it}, \quad t = 1, 2, ..., T \tag{93}$$

In the case that $D_{it} = 1$ represents individuals who choose the non-agricultural sector, Wooldridge (1995) uses the observed characteristics in all periods for a person's probability of choosing a non-agricultural sector. This method assumes ν_{it} is conditional mean zero and normally distribute, $\nu_{it} \sim$ Normal $(0, \sigma_t^2)$. Only when the latent choice variable $D_{it}^* > 0$, it will be observed, expressed in (94).

$$D_{it} = \begin{cases} 1 & \text{if} \quad D_{it}^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{94}$$

Subsequently, in the outcome equation, when $D_{it} = 1$, the log earnings will be observed. Hence, for both agricultural and non-agricultural workers, we can observe their earnings in the data; however, we don't know what they would have earned if they had chosen the other sector instead. For non-agricultural workers, the outcome equation is

$$y_{it} = \theta_i + X_{it}\gamma + u_{it} \quad \text{when} \quad D_{it} = 1 \tag{95}$$

In Equation (95), y_{it} is observed earning for non-agricultural workers, and the assumption is that u_{it} is strictly exogenous conditional on unobserved ability θ_i and observed characteristics X_{it}, see Equation (96). Let ρ be the correlation between u_{it} in the outcome equation and ν_{it} in the selection equation, representing the selection bias in the dataset (refer to Equation (97). Hence, the outcome equation in the expectation has the expression in Equation (98).

$$E(u_{it}|\theta_i, \mathbf{x_i}) = 0 \tag{96}$$

$$E(u_{it}|\theta_i, \mathbf{x_i}, \nu_\mathbf{i}) = E(u_{it}|\nu_{it}) = \rho\nu_{it} \tag{97}$$

$$E(y_{it}|\theta_i, \mathbf{x_i}, \nu_\mathbf{i}, \mathbf{D_i}) = \theta_i + X_{it}\gamma + \rho\nu_{it} \tag{98}$$

When estimating the outcome equation, Wooldridge (1995) first calculates the IMR for each period in the selection equation and then includes the IMRs as a control function in the outcome equation to correct for selection bias (mathematical form expressed in (99)). In estimating the outcome equation, all variables are transformed into demeaned variables (see Equations (100) and (101)), including IMRs (Equation (102)). This method can deliver consistent estimates while imposing much weaker assumptions on the unobserved components. By applying this method to the same dataset, the results will provide an assessment of the selection effect in the sample with a weaker distributional assumption.

$$\ddot{y}_{it} = \ddot{X}_{it}\gamma + \rho\ddot{\nu}_{it} + e_{it} \tag{99}$$

$$\ddot{y}_{it} \equiv y_{it} - \frac{1}{T_i}\sum_{r=1}^{T} D_{ir}y_{ir} \tag{100}$$

$$\ddot{X}_{it} \equiv X_{it} - \frac{1}{T_i}\sum_{r=1}^{T} D_{ir}X_{ir} \tag{101}$$

$$\ddot{\nu}_{it} \equiv \nu_{it} - \frac{1}{T_i}\sum_{r=1}^{T} D_{ir}\nu_{ir} \tag{102}$$

Table 19 represents three different sets of probit estimation as the first stage of Wooldridge's selection correction for panel data. The primary purpose of this step is to obtain the IMRs for each period by assuming the error terms are distributed as a normal distribution. As sector choice is a binary variable, i.e., either working in the non-agricultural or agricultural sector, the IMRs can be calculated for both agricultural and non-agricultural workers under the normal distribution of the error terms in the probit estimation. As Equation (93) shows, the regressors include all the history of observed characteristics. The exclusion restrictions are not required in this Wooldridge (1995) approach; however, they help improve the estimation if available. Let $\lambda(.)$ represent the Inverse Mills Ratio (IMR) and $\hat{\delta}_{\mathbf{t}}$ be estimated coefficients in vector form. Then, IMR for the non-agriculture sector is expressed in Equation (103), and IMR for agriculture workers is calculated as in Equation (104).

$$\lambda(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}}) = \frac{\phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})}{\Phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})} \tag{103}$$

$$\lambda(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}}) = \frac{-\phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})}{1 - \Phi(\mathbf{x}_{\mathbf{i}}'\hat{\delta}_{\mathbf{t}})} \tag{104}$$

The first block in Table 19 does not include either log CPI or shock; the second and third blocks add shock and log CPI, respectively. Each block estimates the selection probability for each period, which enables the calculation of IMRs for each individual for each period. Note that shock does not have any predictive power on the sectoral selection (shown in the second block). This estimation result provides evidential support for the earlier claim that individuals do not switch sectors in response to shocks. However, shocks will affect the

earnings in the outcome equation. Since the outcome equation only includes time-varying variables, the time-invariant variables, such as rural born and gender, can serve as exclusion restrictions in the selection equation.

With the IMRs acquired, a demeaned IMR will be calculated for each individual and included in the outcome equation estimation as a control function to correct selection bias. Table 20 uses IMRs calculated in the first block of Table 19, and the dependent variables are demeaned log income in the primary job. The left three regressions of Table 20 are for non-agriculture workers, and the right three ones are for agriculture. For each sector, three specifications are estimated, from the basic numbers of regressors in Column (1) to gradually adding log CPI in Column (2) and shock in Column (3); all the variables in the outcome equations are demeaned, as expressed in Equations (99) to (102). Table 20 shows that the selection effect is not statistically significant in both sectors, as indicated by the estimated coefficient of λ (in the first line of the table). Moreover, the signs of the estimated coefficients are opposed to the results in the pooled dataset using the Heckman two-step method.

Wooldridge (1995) avoids making explicit the joint-normal distribution of individual unobserved heterogeneity across sectors; instead, this method uses the control function approach and exploits the panel structure. When applying this method to the three waves of the IFLS balanced dataset, selection effects are not statistically significant in either sector. This finding is aligned with the main analysis results in this paper by using Suri (2011)'s approach. Importantly, this method provides evidence on the consequences of the joint-normal distribution assumption on unobserved abilities. Pulido and Świecki (2019) use the same IFLS dataset and find a significant selection effect when imposing a joint-distributional assumption on unobserved components. Examining the first three waves of the IFLS dataset, the pooled Heckman and **xtheckman** with the same assumption as Pulido and Świecki (2019) produce a significant selection effect. Using the same dataset, Wooldridge's (1995) method relaxes the assumption of the bivariate normal and applies a weaker control function on the outcome equation; the estimation results show an insignificant selection effect. The selection effect varies significantly with the distributional assumptions imposed on the unobserved heterogeneity. Hence, it is more desirable to impose weaker assumptions on the individual unobserved abilities when estimating the magnitude of the selection effect. Heck-

man and Honore (1990) have warned of the consequences of such normal distribution in the empirical studies, even though this assumption provides meaningful insight in the theoretical studies.

This paper adopts the empirical approach by Suri (2011) to directly model individual latent skills in each sector while avoiding explicitly imposing functional forms; instead, this method exploits the information of each individual's sectoral choices over time. Unlike the Wooldridge (1995) approach, the methodology used in this paper estimates the individual latent comparative advantages through the choice trajectories and panel structure, which is a more desirable approach when the primary goal is to study the magnitude of the selection effect empirically. However, this method has its limitations. First, data on earnings and choices are required for each individual in each period. Second, the variations of earnings among different groups of choice trajectories need to be sufficiently large for the solutions to be stable when solving a system of equations. Tjernström et al. (2023) thoroughly discuss this limitation and provide evaluations. Despite those limitations, this method offers an attractive solution to the two primary challenges that the research question in this paper faces.

In this section, I first estimate the selection effect by using one of the prevailing approaches in the APG literature, TWFE on panel data, on the same dataset for the primary analysis of the present paper. The estimation results show a significant individual selection effect on sectoral productivity gaps, which reconciles with the findings in the APG literature applying this method. While controlling for individual fixed effects on panel data can remove unobserved heterogeneity, this method is inadequate to model comparative advantages due to the inability to model sector-specific unobserved individual abilities. As a result, this method confounds the effect relevant to sector choice with those that are irrelevant. Then, I present the consequences of the joint-normal distributions on unobserved components in the estimation results. With the same dataset, under the conventional joint-normal assumption, canonical Heckman two-step on pooled data and **xtheckman** both produce significant selection effect, which is aligned with the finding in Pulido and Świecki (2019). Once relaxing the joint normal distribution and using a weaker control function approach, Wooldridge's (1995) finds no significant selection effect in the same dataset. The empirical approach used in the

present paper avoids making such an assumption and exploits the information embedded in the trajectories of choices, which is a more desirable method to study the selection effect on the APG empirically.

nonag_main	(1)			(1) + shock			(1) + ln_cpi		
	t = 1	*t = 2*	*t = 3*	*t = 1*	*t = 2*	*t = 3*	*t = 1*	*t = 2*	*t = 3*
nfarmbiz1	0.798 **	0.514 **	0.357 **	0.801 **	0.513 **	0.355 **	0.809 **	0.522 **	0.361 **
	0.065	0.064	0.063	0.065	0.064	0.063	0.065	0.064	0.063
farmbiz1	-0.702 **	-0.351 **	-0.374 **	-0.708 **	-0.344 **	-0.374 **	-0.681 **	-0.321 **	-0.356 **
	0.068	0.068	0.067	0.069	0.068	0.068	0.069	0.069	0.068
ruralborn1	-0.155 *	0.009	-0.142	-0.154 *	0.009	-0.148 *	-0.156 *	0.013	-0.136
	0.091	0.086	0.088	0.091	0.087	0.088	0.092	0.087	0.088
shock1				-0.006	-0.023	0.027			
				0.058	0.057	0.057			
age1	0.003	-0.001	-0.024 **	0.003	-0.001	-0.024 **	0.006	0.001	-0.023 **
	0.011	0.011	0.01	0.011	0.011	0.01	0.011	0.011	0.01
gender1	-4.919	0.763	0.739	-4.953	0.77	0.732	-4.986	0.642	0.673
	85.969	0.894	0.874	85.321	0.891	0.863	99.891	0.88	0.87
educlevel21	0.215 **	0.149 **	0.211 **	0.217 **	0.149 **	0.211 **	0.228 **	0.163 **	0.216 **
	0.063	0.062	0.06	0.063	0.062	0.06	0.063	0.063	0.06
marital_status1	-0.213 **	-0.213 **	-0.168 **	-0.206 **	-0.214 **	-0.175 *	-0.203 **	-0.203 **	-0.164 **
	0.068	0.066	0.065	0.069	0.066	0.065	0.069	0.066	0.065
urban1	0.14	-0.1	0.111	0.139	-0.1	0.117	0.157	-0.077	0.122
	0.177	0.177	0.173	0.176	0.177	0.174	0.176	0.177	0.174
wagedwork_main1	0.43 **	0.054	-0.063	0.434 **	0.051	-0.069	0.411 **	0.04	-0.069
	0.074	0.074	0.074	0.074	0.074	0.075	0.074	0.074	0.074
ln_hrsworked1_m1	0.186 **	0.033	-0.048	0.184 **	0.034	-0.045	0.193 **	0.043	-0.044
	0.054	0.053	0.054	0.055	0.053	0.054	0.055	0.053	0.054
ln_cpi1							-3.22 **	-1.755	-0.951
							1.323	1.318	1.291
nfarmbiz2	0.387 **	0.657 **	0.275 **	0.385 **	0.658 **	0.278 **	-0.228 **	0.648 **	0.273 **
	0.065	0.064	0.064	0.065	0.064	0.064	0.068	0.065	0.065
farmbiz2	-0.215 **	-0.585 **	-0.26 **	-0.227 **	-0.581 **	-0.256 **	-0.228 **	-0.6 **	-0.265 **
	0.067	0.065	0.066	0.068	0.066	0.067	0.068	0.066	0.067
ruralborn2	-0.038	-0.184	-0.091	-0.033	-0.186	-0.095	-0.052	-0.193	-0.095
	0.162	0.151	0.149	0.162	0.151	0.149	0.163	0.152	0.149
shock2				0.074	0.017	0.003			
				0.056	0.055	0.055			
age2	-0.001	0.004	0.028 *	-0.002	0.004	0.027 *	-0.003	0.003	0.027 *
	0.016	0.016	0.015	0.016	0.016	0.015	0.016	0.016	0.015
gender2	4.156	-1.52 *	-1.528 *	4.187	-1.526 *	-1.521 *	4.21	-1.409	-1.465 *
	85.929	0.898	0.877	85.322	0.894	0.867	99.891	0.883	0.873
educlevel22	0.2 **	0.186 **	0.063	0.197 **	0.168 **	0.064	0.215 **	0.197 **	0.068
	0.062	0.063	0.06	0.063	0.06	0.06	0.063	0.063	0.06
marital_status2	0.097	0.118	0.021	0.092	0.117	0.013	0.096	0.118	0.023
	0.081	0.08	0.078	0.082	0.079	0.078	0.082	0.08	0.078
urban2	-0.555 **	-0.564 **	-0.714 **	-0.561 **	-0.561 **	-0.717 **	-0.521 **	-0.53 **	-0.706 **
	0.205	0.203	0.202	0.204	0.203	0.202	0.206	0.204	0.203
wagedwork_main2	0.193 **	0.561 **	0.116	0.193 **	0.561 **	0.119	0.174 **	0.536 **	0.109
	0.078	0.075	0.078	0.078	0.075	0.076	0.078	0.076	0.076
ln_hrsworked1_m2	-0.012	0.129 **	0.015	-0.011 **	0.128 **	0.018	-0.016	0.125 **	0.013
	0.051	0.049	0.049	0.051	0.049	0.049	0.051	0.049	0.049
ln_cpi2							2.57 **	2.04 *	1.291
							1.281	1.248	1.223
nfarmbiz3	0.418 **	0.375 **	0.73 **	0.418 **	0.376 **	-0.729 **	0.399 **	0.358 **	0.719 **
	0.062	0.061	0.062	0.062	0.061	0.062	0.062	0.061	0.063
farmbiz3	-0.314 **	-0.336 **	-0.72 **	-0.314 **	-0.336 **	-0.709 **	-0.304 **	-0.326 **	-0.716 **
	0.069	0.068	0.066	0.069	0.069	0.066	0.069	0.068	0.066
ruralborn3	-0.085	-0.047	0.062	-0.089	-0.046	0.068	-0.075	-0.038	0.068
	0.097	0.095	0.093	0.097	0.095	0.093	0.098	0.096	0.094
shock3				0.013	-0.0146	-0.132 **			
				0.057	0.056	0.055			
age3	-0.004	-0.011	-0.012	-0.004	-0.011	-0.012	-0.005	-0.012	-0.013
	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
gender3	0	0	0	0	0	0	0	0	0
	(omitted)	(omotted)	(omitted)	(omitted)	(omitted)	(omitted)	(omitted)	(omitted)	(omitted)
educlevel23	-0.026	-0.017	0.092 **	-0.027	-0.017	0.093 **	-0.034	-0.023	0.09 **
	0.034	0.034	0.033	0.034	0.034	0.033	0.034	0.034	0.033
marital_status3	-0.055	-0.079	-0.017	-0.059	-0.076	-0.003	-0.062	-0.086	-0.02
	0.063	0.063	0.063	0.063	0.063	0.063	0.063	0.063	0.063
urban3	1.031 **	1.142 **	1.079 **	1.041 **	1.139 **	1.073 **	0.999 **	1.106 **	1.069 **
	0.123	0.123	0.121	0.124	0.123	0.122	0.125	0.125	0.123
wagedwork_main3	0.17 **	0.162 **	0.556 **	0.172 **	0.16 **	0.551 **	0.164 **	0.162 **	0.553 **
	0.076	0.075	0.075	0.076	0.075	0.075	0.076	0.075	0.075
ln_hrsworked1_m3	-0.015	0.071	0.208 **	-0.011	0.07	0.204 **	-0.01	0.072	0.207 **
	0.047	0.046	0.044	0.047	0.046	0.045	0.047	0.046	0.044
ln_cpi3							-3.217 **	-2.934 **	-1.334 **
							0.673	0.66	0.649
cons	-0.582	-0.499	-0.378	-0.622 **	-0.485 **	0.353 **	19.087 **	13.148	4.674
	0.439	0.426	0.42	0.44	0.428	0.422	5.596	5.561	5.447
N	4,513	4,513	4,513	4,510	4,510	4,510	4,513	4,513	4,513
LR chi2	2,885.33	2,773.61	2,859.02	2,884.73	2,771.46	2,862.19	2,909.10	2,793.80	2,863.38
Pseudo R2	0.494	0.475	0.479	0.494	0.475	0.479	0.498	0.478	0.479

Table 19: Wooldridge Probit Estimation (IFLS 1-3 Waves)

dm_ln_inc1_m	*nonag_main*			ag_main		
	(1)	(2)	(3)	(1)	(2)	(3)
dm_lambda	0.502 **	0.334	0.33	-0.208	-0.3	-0.26
	0.236	0.227	0.227	0.432	0.425	0.426
dm_age	0.126 **	0.058 **	0.057 **	0.112 **	0.045 **	0.042 **
	0.005	0.01	0.01	0.008	0.011	0.011
dm_educ	0.098 **	0.092 **	0.093 **	0.076	0.071	0.07
	0.023	0.022	0.022	0.065	0.064	0.064
dm_marital	-0.055	-0.077 *	-0.072	0.176 **	0.145 *	0.162 **
	0.048	0.047	0.047	0.078	0.078	0.08
dm_urban	0.013	0.013	0.011	-0.044	0.005	0.026
	0.057	0.055	0.055	0.168	0.165	0.168
dm_wagedwork	0.033	0.022	0.02	0.394 **	0.371 **	0.363 **
	0.067	0.066	0.066	0.128	0.125	0.125
dm_ln_hrsworked	0.234 **	0.231 **	0.23 **	0.243 **	0.254 **	0.247 **
	0.048	0.047	0.047	0.077	0.076	0.077
dm_farmbiz	-0.049	-0.011	-0.005			
	0.045	0.044	0.044			
dm_nfarmbiz				-0.031	-0.08	-0.074
				0.117	0.115	0.115
dm_ln_cpi		1.488 **	1.499 **		1.652 **	1.746 **
		0.168	0.167		0.21	0.207
dm_shock			-0.048 *			-0.301 **
			0.027			0.067
control for individual fixed effect	Y	Y	Y	Y	Y	Y
clustered	Y	Y	Y	Y	Y	Y
cons	4.05E-08 **	-4.11E-09	0.000011	3.75E-08 **	-3.82E-08 **	-4.49E-08 **
	1.72E-09	5.65E-09	7.87E-06	1.99E-09	9.60E-09	9.73E-09
sigma_u	2.78E-07	3.32E-07	0.002	2.67E-07	3.44E-07	3.53E-07
sigma_epsilon	0.828	0.816	0.816	1.598	1.588	1.58E+00
ICC	1.13E-13	1.66E-13	6.67E-06	2.79E-14	4.68E-14	4.97E-14
corr(u_i, X_it)	0	0	0.007	0	0	0
N	8,691	8,691	8,688	4,848	4,848	4,848
Selected	3,343	3,343	3,343	2,063	2,063	2,063
F	117.91 **	273.16 **	259.14	36.65 **	42.09 **	42.62 **
	(8, 3,342)	(9, 3,342)	(10, 3,342)	(8, 2,062)	(9, 2062)	(10, 2,062)

Table 20: Wooldridge Outcome Equation Estimation 1 (IFLS 1-3 Waves)

dm_ln_inc1_m	*nonag_main*			ag_main		
	(1)	(2)	(3)	(1)	(2)	(3)
dm_lambda	0.505 **	0.386 *	0.386 *	-0.143	-0.161	-0.116
	0.245	0.237	0.237	0.423	0.419	0.42
dm_age	0.126 **	0.058 **	0.057 **	0.111 **	0.044 **	0.042 **
	0.005	0.01	0.01	0.008	0.011	0.011
dm_educ	0.098 **	0.094 **	0.095 **	0.079	0.076	0.075
	0.023	0.022	0.022	0.065	0.065	0.064
dm_marital	-0.055	-0.077 *	-0.072	0.176 **	0.147 *	0.164 **
	0.048	0.047	0.047	0.078	0.078	0.08
dm_urban	0.015	0.015	0.014	-0.04	0.01	0.031
	0.057	0.055	0.055	0.168	0.165	0.168
dm_wagedwork	0.033	0.026	0.024	0.411 **	0.407 **	0.401 **
	0.067	0.066	0.066	0.126	0.124	0.123
dm_ln_hrsworked	0.234 **	0.234 **	0.233 **	0.248 **	0.264 **	0.258 **
	0.048	0.047	0.047	0.077	0.076	0.076
dm_farmbiz	-0.049	-0.019	-0.013			
	0.045	0.044	0.044			
dm_nfarmbiz				-0.021	-0.058	-0.051
				0.116	0.114	0.114
dm_ln_cpi		1.49 **	1.501 **		1.642 **	1.737 **
		0.167	0.166		0.209	0.207
dm_shock			-0.049 *			-0.301 **
			0.028			0.067
control for individual fixed effect	Y	Y	Y	Y	Y	Y
clustered	Y	Y	Y	Y	Y	Y
cons	4.03E-08 **	-4.38E-09	0.000011	3.78E-08 **	-3.73E-08 **	-4.41E-08 **
	1.74E-09	5.62E-09	7.90E-06	1.86E-09	9.66E-09	9.61E-09
sigma_u	2.78E-07	3.33E-07	0.002	2.67E-07	3.43E-07	3.52E-07
sigma_epsilon	0.828	0.816	0.816	1.598	1.588	1.583
ICC	1.13E-13	1.66E-13	6.54E-06	2.79E-14	4.66E-14	4.95E-14
corr(u_i, X_it)	0	0	0.007	0	0	0
N	8,691	8,691	8,688	4,848	4,848	4,848
Selected	3,343	3,343	3,343	2,063	2,063	2,063
F	116.48 **	272.41 **	258.68	36.67 **	41.89 **	42.62 **
	(8, 3,342)	(9, 3,342)	(10, 3,342)	(8, 2,062)	(9, 2,062)	(10, 2,062)

Table 21: Wooldridge Outcome Equation Estimation 2 (IFLS 1-3 Waves)

References

Adamopoulos, T., Brandt, L., Leight, J., and Restuccia, D. (2022). Misallocation, selection and productivity: A quantitative analysis with panel data from china. *Econometrica*, 90:1261–1282.

Adamopoulos, T. and Restuccia, D. (2014). The size distribution of farms and intenational productivity differences. *American Economic Review*, 104(6):1667–1697.

Alvarez, J. A. (2020). The agricultural wage gap: Evidence from brazilian microdata. *American Economic Journal: Macroeconomics*, 12:153–173.

Alvarez-Cuadrado, F., Amodio, F., and Poschke, M. (2020). Selection, absolute advantage, and the agricultural productivity gap. *Centre for Economic Policy Research*, 4:1–82.

Alvarez-Cuadrado, F., Long, N. V., and Poschke, M. (2017). Capital-labour substitution, structural change and growth. *Theoretical Economics*, 12(3):1229–1266.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *The American Economic Review*, 77(4):531–553.

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.

Cabanillas, O. B., Michler, J. D., Michuda, A., and Tjernstrom, E. (2018). Fitting and interpreting correlated random-coefficient models using stata. *The Stata Journal*, 18(1):159–173.

Caselli, F. (2005). Accounting for cross-country income differences. In Aghion, P. and N., D. S., editors, *Handbook of Economics Growth, Vol. 1A*, pages 679–741. Elsevier, Amsterdam.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, 18(1):5–46.

Chamberlain, G. (1984). Panel data. *Handbook of Econometrics*, 2:1247–1318.

Chanda, A. and Dalgaard, C.-J. (2008). Dual economies and international total factor productivity differences: channelling the impact from institutions, trade, and geography. *Economica*, 75(300):629–661.

Chen, C., Restuccia, D., and Santaeulalia-Llopis, R. (2023). Land misallocation and productivity. *American Economic Journal: Macroeconomics*, 15(2):441–465.

Frankerberg, E., Karoly, L. A., Gertler, P., Achmad, S., Agung, I. G. N., Hatmadji, S. H., and Sudharto, P. (1995). The 1993 indonesia family life survey: overview and field report. *RAND*, 1(DRU-1195/1-NICHD/AID).

Gai, Q., Guo, N., Li, B., Shi, Q., and Zhu, X. (2021). Migration costs, sorting, and the agriculture productivity gap. *University of Toronto*, Working Paper(693).

Gollin, D. and Kaboski, J. p. (2023). New views of structural transformation: insights from recent literature. *Oxford Development Studies*, 51(4):339–361.

Gollin, D., Lagakos, D., and Waugh, M. E. (2014). The agricultural productivity gap. *The Quarterly Journal of Economics*, 129(2):939–993.

Gollin, D., Parente, S., and Rogerson, R. (2002). The role of agriculture in development. *American Economic Review*, 92(2).

Gouriéroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118.

Group, W. B. (2025). Databank: World development indicators, ind.

Hamory, J., Keelmans, M., Li, N. Y., and Miguel, E. (2021). Reevaluating agricultural gaps with longitudinal microdata. *Journal of the European Economic Association*, 19(3):1522–1555.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Heckman, J. and Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. *The Journal of Human Resources*, 33(4):974–987.

Heckman, J. J. and Honore, B. E. (1990). The empirical content of roy model. *Econometrica*, 58(5):1121–1149.

Herrendorf, B. and Schoellman, T. (2018). Wages, human capital, and barriers to structural transformation. *American Economic Journal: Macroeconomics*, 10:1–23.

Hofman, B. and Kaiser, K. (2002). The making of the big bang and its aftermath: A political economy perspective. In *Can Decentralization Help Rebuild Indonesia?*, volume Working Paper 02-25 of *International Studies Program*, Atlanta, Georgia. Andrew Young School of Policy Studies, Georgia State University.

House, F. (1998). Freedom in the world 1998 - indonesia.

Kuznets, S. (1971). *Economic Growth of Nations: Total Output and Production Structure*. Harvard University Press, Cambridge, MA.

Lagakos, D. (2020). Urban-rural gaps in the developing world: does internal migration offer opportunities. *Journal of Economic Perspectives*, 34(3):174–192.

Lagakos, D., 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.

Lagakos, D. and Waugh, M. E. (2013). Agriculture, and cross-country productivity differences. *The American Economic Review*, 103(2):948–980.

Lemieux, T. (1998). Estimating the effects of unions on wage inequality in two-sector model with comparative advantage and non-random selection. *Journal of Labour Economics*, 16:261–291.

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The Manchester School*, 22(2):115–227.

McMillan, M. and Rodrik, D. (2014). Globalization, structural change, and productivity growth, with an update on africa. *World Development*, 63:11–32.

Munshi, K. and Rosenzweig, M. (2016). Networks and misallocation: insurance, migration, and the rural-urban wage gap. *American Economic Review*, 106(1):46–98.

Pulido, J. and Świecki, T. (2019). Barriers to mobility or sorting? sources and aggregate implications of income gaps across sectors and locations in indonesia. *Working Paper*.

Restuccia, D. and Rogerson, R. (2017). The causes and costs of misallocation. *Journal of Economic Perpectives*, 31(3):151–174.

Restuccia, D., Yang, D. T., and Zhu, X. (2008). Agriculture and aggregate productivity: a quantitative cross-country analysis. *Journal of Monetary Economics*, 55(2):234–250.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, New Series*, 3(2):135–146.

Samuelson, P. A. (1938). A note on pure theory of consumer's behaviour. *Economica*, 5(17):61–71.

Samuelson, P. A. (1948). Consumer theory in terms of revealed preference. *Economica*, 15(60):243–253.

Strauss, J., Witoelar, F., and Sikoki, B. (2016). The fifth wave of indonesia family life survey (ifls4): overview and field report. *RAND*, 5(WR-1143/1-NIA/NICHD).

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica*, 79:159–209.

Tjernström, E., Ghanem, D., Cabanillas, O. B., Lybbert, T., Michuda, A., and Michler, J. (2023). Comment on suri (2011) "selection and comparative advantage in technology adoption". Working paper.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68:115–132.