Learning and Selection in the Sorting of Households Across Sectors

Jean-François Gauthier[*] Teresa Molina [†] Anant Nyshadham[‡][§]

Abstract

We study the role of imperfect information about relative productivity across sectors in explaining low productivity in developing countries. We estimate a generalized earnings equation with dynamic correlated random coefficients, allowing households to learn about their relative productivity across the agricultural and non-agricultural sectors. Estimates show that households sort across sectors on comparative advantage, but learn and converge slowly over time, with many households spending substantial time in their less productive sector. In our first wave of data, roughly 35% of households are in the less productive sector for them, earning 79% less on average than they could have if they were properly sorted across sectors. Our approach nests several alternative models which can be ruled out, including those without dynamics and/or heterogeneity in relative productivity across sectors. We also evaluate alternative interpretations for the dynamic sorting we observe in the data such as saving out of financial constraints and skill accumulation or learning by doing.

JEL Classification Codes: J24, J43, O14, O40

Keywords: sectoral choice, learning, sorting, comparative advantage, dynamic correlated random coefficients

[*] HEC Montréal; jean-francois.gauthier@hec.ca;

[†] University of Hawaii at Manoa and IZA; tmolina@hawaii.edu;

[‡] University of Michigan, Good Business Lab, JPAL, BREAD, & NBER; nyshadha@umich.edu;

[§] We would like to especially thank Mike Boozer, Penny Goldberg, Tavneet Suri, and Chris Udry for their invaluable support in the development of the dissertation upon which this paper builds. We would also like to thank Ach Adhvaryu, Joe Altonji, Francesco Amodio, A.V. Chari, Dave Donaldson, Andrew Foster, Guido Imbens, Joe Kaboski, Fabian Lange, Costas Meghir, Ted Miguel and participants in workshops and seminars at Yale, Stanford, UC Berkeley, Warwick, USC, UIUC, UC Riverside, Johns Hopkins, BREAD, and NBER for their helpful comments. All errors are our own.

1 Introduction

Productivity is much lower in developing countries than in developed countries (Bloom et al., 2010; Hall and Jones, 1999; Syverson, 2011). Hypothesized drivers of this gap have included managerial quality (Adhvaryu et al., 2019b; Bloom et al., 2013; Bloom and Van Reenen, 2007), trade relationships and costs (Adhvaryu et al., 2019a; Atkin and Donaldson, 2015; Atkin et al., 2017), and resource misallocation across sectors (Hsieh and Klenow, 2009). While much of this evidence has focused mainly on non-agricultural sectors and larger formal firms, related empirical work has documented that productivity gaps across developed and developing countries are particularly large in the agricultural sector (Gollin et al., 2014; Restuccia et al., 2008). Misallocation of capital and land has also been hypothesized as a driver of this latter pattern (Adamopoulos et al., 2017; Restuccia and Rogerson, 2013; Restuccia and Santaeulalia-Llopis, 2017), along with self-selection of households across sectors (Alvarez-Cuadrado et al., 2019; Lagakos and Waugh, 2013). Recent models of the process of structural transformation have used labor reallocation frictions to explain productivity patterns across agriculture and non-agriculture sectors (Porzio et al., 2020).

In this paper, we aim to build on this prior evidence, asking if inefficiency in the sorting of labor across sectors contributes to low productivity in developing countries in both the agricultural and non-agricultural sectors. We hypothesize that imperfect information about relative productivity might lead households to select into a less productive sector for them early on in their productive life cycles. Previous studies have modeled selection as a one-off sorting decision across sectors, limiting the ability to document inefficient sorting along households' productive life cycles. That is, these analyses can document sectoral sorting for a population at a given point in time, but cannot comment on whether this particular sorting decision is the most productive choice for each household. To the degree that households converge to their most productive sector over time as they learn about which sector best suits their skills, a dynamic approach is required to identify: i) for which sector each household ultimately appears best suited, ii) whether and for how long each household participates in an ill-matched sector, and iii) how much their earnings suffer along the way as a result.

We adapt the dynamic sectoral sorting framework in Gibbons et al. (2005) to the developing country household's decision to engage in non-agricultural work. This model of selection in which households learn about their relative productivity across sectors yields a generalized earnings equation with dynamic correlated random coefficients (DCRC). We use an extension of projection-based panel methods (Chamberlain, 1982, 1984; Islam, 1995; Suri, 2011) to estimate the model on the longitudinal Indonesia Family Life Survey (IFLS), which spans more than two decades.[1] We analytically link the interpretation of our structural estimates to the seminal formulation of the Roy (1951) model in Borjas (1987), which allows us to use our estimates to characterize the nature of sorting in our context as either positive selection, negative selection, or sorting on comparative advantage.

Results show that households sort across sectors on the basis of comparative advantage, consistent with findings from other recent studies (Adamopoulos et al., 2017; Lagakos and Waugh, 2013; Papageorgiou, 2014). We document substantial heterogeneity in the returns to engaging in non-agricultural work. While the average annual return is roughly 5.9 million rupiah (425 USD), the expected returns among households who actually switch in or stay in the non-agricultural sector are 2 to 3 times as large and the returns for households who switch out or stay out are negative.

We also document substantial churning along the sectoral margin, an empirical regularity across most developing countries that only a few papers have studied (Adhvaryu et al., 2020; Adhvaryu and Nyshadham, 2017; Calderon et al., 2020). Preliminary evidence from the raw data shows that this churning reduces with experience in a sector. That is, a household is less likely to switch the longer they have been in a particular sector, consistent with learning. Structural estimates confirm that the observed churning is at least in part a result of substantial learning and slow convergence such that many households spend substantial amounts of time in a sector which is less productive for them. At the start of the sample, roughly 35% of households are in their less productive sector, and these households are earning 79% less on average than they could have if they were properly sorted across sectors. After 14 years, 25% of households (and not necessarily the same households) remain in their less productive sector, sorting on persistently imprecise perceptions of relative productivity.

We recover structural estimates of both the household's latent relative ability

[1] The fundamentals of this approach to panel data are reviewed in Crépon and Mairesse (2008). We discuss later when we develop the methodology how we draw from extensions developed in Islam (1995) to allow for dynamics and Suri (2011) to allow for selection on comparative advantage.

across sectors and its evolving perceptions regarding it over time. We document that returns to participating in the non-agricultural sector are higher for households with members exhibiting higher cognitive ability and better physical health as well as more open-mindedness and extraversion. However, the full set of observable covariates still only explains 13% of the variation in returns across sectors, consistent with the observed prevalence and persistence of inefficient sorting.

Our approach nests several alternative models which can be ruled out. For example, we can estimate a model with comparative advantage but no dynamics as well as a model with neither dynamics nor heterogeneity in relative earnings across sectors. We find that dynamics are important and in fact that the heterogeneity in relative earnings across sectors is much more pronounced in estimates when allowing for dynamics.

We also evaluate alternative interpretations for the dynamic heterogeneity we observe in the data. One advantage of our projection-based approach to estimating the DCRC model is that it can obtain consistent estimates of both the average return and the latent heterogeneity under these alternative interpretations so long as the assumption of sequential exogeneity is preserved. Under these different models, however, the interpretation of the latent heterogeneity and the expected patterns of the estimated dynamics would differ. We evaluate whether land market frictions, saving out of financial constraints, or skill accumulation (i.e., learning by doing) could explain the patterns we observe in the raw data as well as the structural parameters we recover, and find each of these alternative interpretations to be less consistent with our findings than learning about comparative advantage.

Our study contributes to two strands of the literature on the causes of low productivity in developing countries (Bloom et al., 2010; Hall and Jones, 1999; Syverson, 2011). Several papers have investigated the role of the inefficient allocation of capital and other non-labor inputs across sectors due to various frictions (Adamopoulos et al., 2017; Hsieh and Klenow, 2009; Restuccia and Rogerson, 2013). The inefficient sorting of labor across sectors has also been hypothesized when documenting productivity gaps across sectors (Gollin et al., 2014); frictions in the movement of labor across sectors has been modeled in studies of sectoral sorting and structural transformation (Porzio et al., 2020; Pulido et al., 2018). We expand on this work by quantifying the degree of the inefficiency in labor sorting and identifying information frictions as a cause – leveraging a long panel to document in which sector each household's earnings are maximized and how often they deviate from this most productive sector. In this sense our paper is closest to the recent work by Adamopoulos et al. (2017) showing in China that labor selection reinforces the negative productivity effects of land and capital misallocation across sectors. We complement this work by documenting that labor selection can be imperfect due to information frictions, leading to substantial and costly inefficiency in the sorting of labor as well.[2]

In doing so, we also build on evidence of the sorting of households across sectors (Alvarez-Cuadrado et al., 2019; Lagakos and Waugh, 2013). We find strong evidence that households sort across sectors on the basis of perceived comparative advantage, but extend the approaches in previous papers to assess whether a household is sorting into the most productive sector for them each period. Static approaches interpret realized sorting as revealed preference; whereas our DCRC model allows for households to have imperfect information and make mistakes along the way as a result. This flexibility allows us to fit the observed sectoral churning in the data, common across contexts but often overlooked in empirical analyses of sorting. Our approach also allows us to recover consistent estimates of the average returns, latent heterogeneity, and correlations between current income realizations and future sectoral choices under several alternative interpretations including saving out of financial constraints and skill accumulation, and then to evaluate which of these interpretations is most consistent with the parameter estimates we recover. As mentioned above, we find the results to be most consistent with a learning about comparative advantage interpretation.

Our paper relates to recent work by Hicks et al. (2017) and Pulido et al. (2018), which use our same longitudinal dataset to explore reasons for the productivity gap between the agricultural and non-agricultural sectors. Unlike these two papers, our focus is on a phenomenon that can explain low productivity in both sectors – inefficient sorting driven by imperfect information. The model we use is an extension of the fixed effects approach used by Hicks et al. (2017) in which we allow for dynamic correlated random coefficients.[3] Our model nests both the fixed effects approach and

[2] Note that in our study we aim to explicitly cut past aggregate market level frictions by including community by year fixed effects to focus on information frictions at the household level. In this sense, we aim to complement prior evidence on land and capital market frictions. Those may very well still play a role in our setting in addition to the role of household-level information frictions we focus on, but they should not conflate the analysis we undertake, as discussed below, and are not the primary focus of our study.

[3] Note Hicks et al. (2017) also differ from us in that they perform their analysis at the individual

a model of sorting on comparative advantage without dynamics. This allows us to test and reject the ability of these simpler frameworks to match the patterns in the data. We are able to validate the importance of information frictions and learning, which are not considered in either of these two studies but which we find result in many households choosing a sector that is not the most productive one for them.[4]

2 Data and Motivation

2.1 IFLS

We use the Indonesian Family Life Survey (IFLS), a longitudinal household survey that began in 1993, with four follow-ups conducted in 1997, 2000, 2007, and 2014 (Strauss et al., 2016). The sample is representative of the 13 provinces that were selected to be included in the first survey wave (corresponding to over 80% of the Indonesian population). The IFLS collected detailed information about a wide array of household and individual characteristics, including basic demographics, educational attainment, physical health, cognitive ability, risk aversion, and most importantly for this paper, sectoral choice and income from various sources. Specifically, the main respondent for each household is asked about the household's ownership of and income from household enterprise (both farm and non-farm), and each household member aged 15 or older is asked to report their individual wage income as well as the sector of their primary and (if applicable) secondary occupation.

We are interested in total annual household income, which we calculate as the sum of profits from non-farm enterprise, profits from farm enterprise (both of which

level. In keeping with most other studies which focus on farm and non-farm enterprise in developing country contexts, we prefer to perform our analysis at the household level given the difficulty in measuring intrahousehold labor supply and the division of earnings from these enterprises which are very common in our data, but we show robustness of our results to individual level analysis below.

[4] Pulido et al. (2018) structurally estimate a macro model of sectoral sorting with restrictions to mobility across sectors, which like our approach leverages switching histories to better fit the data, but their estimates suggest that households who switch out of the non-farm sector realize income losses. They justify this either by taste or utility-based compensating differentials or with market frictions leading to switchers-out getting "stuck" in the agriculture sector despite greater earning potential in the non-agricultural sector. Our estimates, on the other hand, show for many households earnings are actually maximized ultimately in the agricultural sector, such that switching out is ultimately optimal, but convergence to this realization is slow due to information frictions. We argue this explanation better fits the bilateral, high-frequency switching which slowly reduces over time observed in the data. We evaluate alternative interpretations including those related to frictions studied by Pulido et al. (2018) and Adamopoulos et al. (2017) in detail below.

can be negative or positive), and all household members' wage income (from both the primary and secondary occupation).[5] After this, we restrict to households with non-missing non-agricultural profits, farm enterprise profits, and wage income in all five waves. This leaves us with 3875 households in a balanced panel sample.

This paper focuses on the household-level decision to exit the agricultural sector. We use the household as our unit of analysis, as other related work does (Adamopoulos et al., 2017; Alvarez-Cuadrado et al., 2019), because ownership of a household enterprise, which is arguably a household-level rather than an individual-level decision, is common in our sample.[6] In household surveys like the IFLS, it can be difficult or even impossible to allocate and value time use of household members across these household enterprises, let alone to divide profits among all members associated with the enterprises. Nevertheless, we demonstrate robustness of our results to individual level analysis below as well.

As our sectoral choice variable of interest, we generate an indicator equal to one for households who either have a non-agricultural enterprise or earn at least half of their income from non-agricultural wage work. We show, however, that our results are robust to variations of this definition (e.g., having more than half of household members working in the non-agricultural sector), which is not surprising as the learning structure of the model allows for households to learn about their relative productivity regardless of which sector they are currently working in. Over the five survey waves, between 55% to 66% of households worked in the non-agricultural sector according to this definition (as shown in Table 1).

In Table 1, we also report total annual household income in millions of 2015 Indonesian rupiahs. In 1993, average household income was approximately 9 million rupiahs (around 650 USD), but by 2014, this increased to approximately 24 million.

2.2 Preliminary Evidence

Basic descriptive exercises reveal substantial churning in and out of agriculture. In Figure 1, we illustrate the share of households in the agricultural and non-agricultural

[5] Given the importance of this income variable for our analysis, we first drop outliers in each wave (specifically, the top 1% and bottom 1% of the income distribution), which we suspect suffer from reporting errors – a common method for trimming self-reported incomes.

[6] In 1993, 39% of IFLS households own a farm business, while 34% own a non-farm business (62% own either). In 2014, the percent of households who own any enterprise is roughly the same (59%), though a larger share own non-farm businesses (38%) than farm businesses (32%) by this time.

Table 1: Summary Statistics

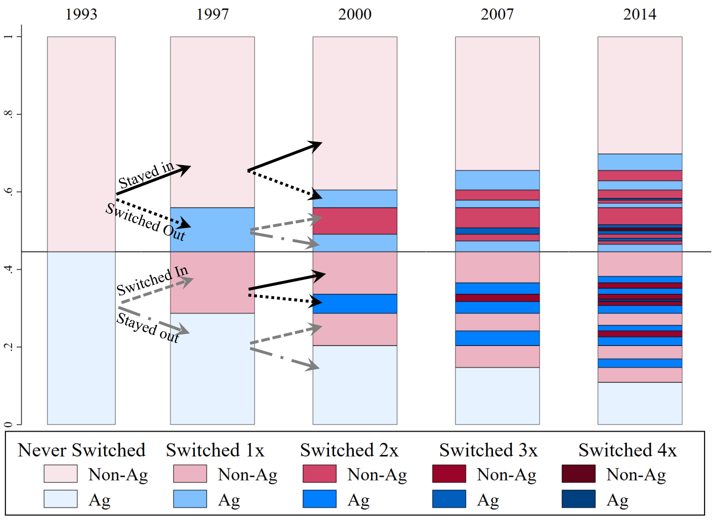
			Year		
	1993	**1997**	**2000**	**2007**	**2014**
Non-Ag Sector	0.55	0.60	0.66	0.64	0.65
	(0.50)	(0.49)	(0.48)	(0.48)	(0.48)
Total Household Income	9.06	11.3	13.5	17.8	23.9
	(13.3)	(14.3)	(16.2)	(21.4)	(31.1)
Household Size	4.69	4.61	4.59	4.14	3.84
	(2.01)	(1.91)	(1.92)	(1.87)	(1.91)
No. Females Aged 15-59	1.38	1.41	1.41	1.33	1.23
	(0.81)	(0.81)	(0.83)	(0.84)	(0.86)
No. Males Aged 15-59	1.27	1.27	1.31	1.26	1.11
	(0.88)	(0.89)	(0.92)	(0.94)	(0.93)
Observations	3875	3875	3875	3875	3875

Notes: Sample consists of IFLS households with non-missing income information in all five waves of the IFLS. Standard deviations reported in parentheses.

sectors, with five shades of red that represent non-agricultural households and five shades of blue that represent agricultural households. The darkness of a color indicates the number of times a household has switched. In 1993, when we do not have any previous information on sector, all households have never switched according to our data and are therefore represented by the lightest shades of red (for those currently in the non-agricultural sector) and blue (for those currently in agriculture). In 1997, however, 20% of the households who were in the non-agricultural sector in 1993 switched to agriculture in 1997 (represented by a slightly darker shade of blue because they switched once). At the same time, around 36% of the 1993 agricultural households switched into non-agricultural work in 1997 (represented by a slightly darker shade of red).

This switching behavior continues across the remaining 3 waves. By 2014, it is clear that over half of households have switched at least once (any color that is not the lightest red or blue represents a household that has switched). There are many households that have switched more than once, and even some that have switched four times. In short, switching sectors is common. We also note that every one of the 32 possible sectoral choice trajectories is represented in the the last wave, which is important for the identification of the model as we discuss later.

Figure 1: Churning Across Sectors Over Time



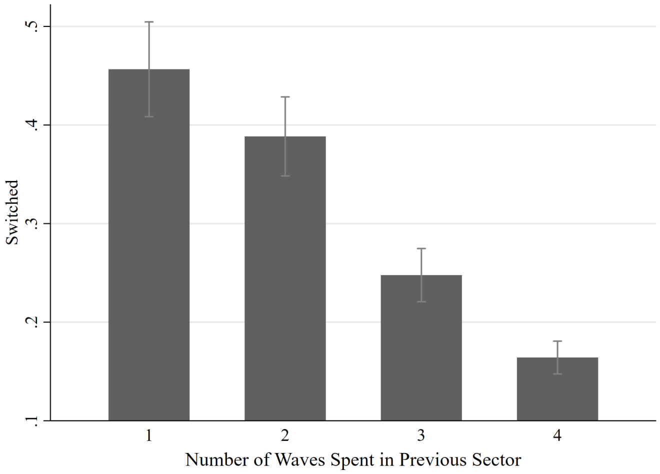
Notes: Sample consists of IFLS households with non-missing income information in all five waves of the IFLS. Shades of red represent households that are in the non-agricultural sector in the relevant wave, while shades of blue represent households that are not. Color darkness captures the number of times a household has switched prior to that wave.

We next ask whether switching declines with the amount of time a household spends in a sector. Figure 2 shows that it does. Between the fourth and fifth (the last) wave, among households that have been in their current sector for only one wave, 46% of households switched sectors. This share drops with the cumulative number of waves spent in the previous sector: only about 16% of households who have remained in their sector for 4 waves switch in the fifth wave.[7] This suggests that, though sectoral switching is common, households' switching decisions appear to exhibit convergence, such that longer time spent in a given sector yields a lower probability of switching. In the appendix, we show this pattern holds in both directions (i.e., for both agricultural and non-agricultural households (see Figure A1)). The patterns depicted in these figures motivate the model we develop in the next section, where workers learn about their sector-specific ability over time. The high-frequency, bidirectional switching and

[7] Patterns are similar when we generate these graphs using all waves of data, but we note that graphs restricting to the last wave provide a more accurate representation of the relationship between switching and time spent in a sector. Because we do not know how long a household has been in a given sector when we first observe them in 1993, we are underestimating the length of time a household has spent in a given sector and this is particularly problematic for early waves. By definition, no household is classified as having spent 4 waves in a sector until the last wave of the survey.

trend of reduced switching over time are also consistent with the stylized facts that motivate the model in Papageorgiou (2014), where workers also learn about their comparative advantage over time.

Figure 2: Switching by Number of Waves Spent in Previous Sector



Notes: Sample consists of IFLS households with non-missing income information in all five waves of the IFLS. Graph illustrates switching behavior from the fourth to fifth (and last) wave of the survey. Error bars denote 95% confidence intervals.

3 Model

3.1 Sectoral Choice

In this section, we outline a Roy (1951) model of sectoral choice, where household i in period t chooses whether to go into the non-agricultural sector (denoted by superscript N) or stay in the agricultural sector (denoted by superscript A). Sector-specific income Y_{it} is determined by the following equations:

$$\begin{aligned} Y_{it}^N &= \beta_t^N + \eta_i^N \\ Y_{it}^A &= \beta_t^A + \eta_i^A. \end{aligned} \tag{1}$$

β_t^N is average income in the non-agricultural sector and β_t^A is average income in the agricultural sector. η_i^N is the unobserved, heterogeneous component of productivity specific to the non-agricultural sector, while η_i^A is the corresponding component for the agricultural sector.

We can rewrite both η_i^N and η_i^A as a function of relative productivity $(\eta_i^N - \eta_i^A)$, and absolute advantage, τ_i, which we define as the component of the household-specific productivity that has the same effect on the household's productivity in both sectors. (Accordingly, τ_i does not affect the sectoral choice.) Specifically, we rewrite each sector-specific productivity term in the following way:

$$\begin{aligned} \eta_i^N &= (1+\phi)\eta_i + \tau_i \\ \eta_i^A &= \eta_i + \tau_i, \end{aligned} \tag{2}$$

where both ϕ and η_i depend on projection coefficients, b_A and b_N.[8] We define $\phi \equiv b_N/b_A - 1$, and $\eta_i \equiv b_A(\eta_i^N - \eta_i^A)$.

The equations in (2) show that a household's sector-specific productivity is a function of both relative productivity and absolute advantage. Importantly, the parameter ϕ depends on the covariance between non-agricultural and agricultural productivity in the population as a whole, $Cov(\eta_i^N, \eta_i^A)$, and therefore summarizes the nature of sorting in the population.

To explore how ϕ governs the nature of selection in the Roy model, we combine equations (1) and (2) and suppress t subscripts to express income in the non-agricultural and agricultural sectors as follows:

$$\begin{aligned} Y_i^N &= \beta^N + (1+\phi)\eta_i + \tau_i \\ Y_i^A &= \beta^A + \eta_i + \tau_i. \end{aligned}$$

Unconditional expected income (in the non-agricultural and agricultural sector) is

$$\begin{aligned} E[Y_i^N] &= \beta^N + (1+\phi)E[\eta_i] + E[\tau_i] \\ E[Y_i^A] &= \beta^A + E[\eta_i] + E[\tau_i]. \end{aligned}$$

Let D_i represent a dummy equal to one for households in the non-agricultural sector. Households sort across sectors based on their η_i; specifically, households with $\phi\eta_i >$

[8] Since with 2 sectors only the relative magnitude of η_i^A and η_i^N can be identified, we will define, following Lemieux (1998) and Suri (2011), η_i^A and η_i^N in terms of the household's relative productivity in non-agricultural over agricultural activity $(\eta_i^N - \eta_i^A)$ using the following projections: $\eta_i^A = b_A(\eta_i^N - \eta_i^A) + \tau_i$ and $\eta_i^N = b_N(\eta_i^N - \eta_i^A) + \tau_i$, where $b_N = (\sigma_N^2 - \sigma_{NA})/(\sigma_N^2 + \sigma_A^2 - 2\sigma_{NA})$, $b_A = (\sigma_{NA} - \sigma_A^2)/(\sigma_N^2 + \sigma_A^2 - 2\sigma_{NA})$, with $\sigma_{NA} \equiv Cov(\eta_i^N, \eta_i^A)$, $\sigma_N^2 \equiv Var(\eta_i^N)$, and $\sigma_A^2 \equiv Var(\eta_i^A)$.

$-\beta$ (where $\beta \equiv \beta^N - \beta^A$) will choose to go into non-agricultural work ($D_i = 1$). Therefore, conditional average non-agricultural and agricultural income, among those who select into the non-agricultural sector, is the following:

$$\begin{aligned}
E[Y_i^N|D_i = 1] &= E[Y_i^N|\phi\eta_i > -\beta] \\
&= \beta_t^N + (1+\phi)E[\eta_i|\phi\eta_i > -\beta] + E[\tau_i|\phi\eta_i > -\beta] \\
&= \beta_t^N + (1+\phi)E[\eta_i|\phi\eta_i > -\beta] + E[\tau_i] \\
E[Y_i^A|D_i = 1] &= E[Y_i^A|\phi\eta_i > -\beta] \\
&= \beta_t^A + E[\eta_i|\phi\eta_i > -\beta] + E[\tau_i|\phi\eta_i > -\beta] \\
&= \beta_t^A + E[\eta_i|\phi\eta_i > -\beta] + E[\tau_i],
\end{aligned}$$

where the last step is due to the independence of τ and η.

We focus on the same income differentials as Borjas (1987), who characterizes sorting by distinguishing between positive selection, negative selection, and "refugee sorting" or sorting on comparative advantage. The first differential of interest is the difference between average non-agricultural income among households that select into the non-agricultural sector and unconditional average non-agricultural income (labeled Q_1 in Borjas (1987) and defined by equation (3) below). The second differential of interest is the difference between average agricultural income among households that select into the non-agricultural sector and unconditional average agricultural income (labeled Q_0 in Borjas (1987) and defined by equation (4) below). Positive selection is defined as the case when $Q_1 > 0$ and $Q_0 > 0$, negative selection when $Q_1 < 0$ and $Q_0 < 0$, and sorting on comparative advantage when $Q_1 > 0$ and $Q_0 < 0$.

$$E[Y_i^N|D_i = 1] - E[Y_i^N] = (1+\phi)\left(E[\eta_i|\phi\eta_i > -\beta] - E[\eta_i]\right) \tag{3}$$

$$E[Y_i^A|D_i = 1] - E[Y_i^A] = \left(E[\eta_i|\phi\eta_i > -\beta] - E[\eta_i]\right). \tag{4}$$

3.1.1 Case 1: $\phi > 0$

When $\phi > 0$, average non-agricultural income among those who select into the non-agricultural sector is higher than the population average of non-agricultural income, as shown below. Average agricultural income is also higher among those who select into the non-agricultural sector. This means that non-agriculture households are positively selected.

$$E[Y_i^N|D_i=1]-E[Y_i^N]=\overbrace{(1+\phi)}^{>0}\overbrace{\left(E[\eta_i|\eta_i>-\frac{\beta}{\phi}]-E[\eta_i]\right)}^{>0}>0$$

$$E[Y_i^A|D_i=1]-E[Y_i^A]=\overbrace{\left(E[\eta_i|\eta_i>-\frac{\beta}{\phi}]-E[\eta_i]\right)}^{>0}>0.$$

3.1.2 Case 2: $-1<\phi<0$

When $-1<\phi<0$, we have negative selection. Both average non-agricultural income and average agricultural income among those who select into the non-agricultural sector are lower than population averages. Those who select into the non-agricultural sector tend to be less productive in both sectors.

$$E[Y_i^N|D_i=1]-E[Y_i^N]=\overbrace{(1+\phi)}^{>0}\overbrace{\left(E[\eta_i|\eta_i<-\frac{\beta}{\phi}]-E[\eta_i]\right)}^{<0}<0$$

$$E[Y_i^A|D_i=1]-E[Y_i^A]=\overbrace{\left(E[\eta_i|\eta_i<-\frac{\beta}{\phi}]-E[\eta_i]\right)}^{<0}<0.$$

3.1.3 Case 3: $\phi<-1$

Finally, when $\phi<-1$, average non-agricultural income among those who select into the non-agricultural sector is higher than the population average of non-agricultural income. However, average agricultural income is lower among those who select into the non-agricultural sector. This implies sorting based on comparative advantage: productive non-agricultural households would have low productivity in agriculture, while productive agricultural households would have low productivity in the non-agricultural sector.

$$E[Y_i^N|D_i=1]-E[Y_i^N]=\overbrace{(1+\phi)}^{<0}\overbrace{\left(E[\eta_i|\eta_i<-\frac{\beta}{\phi}]-E[\eta_i]\right)}^{<0}>0$$

$$E[Y_i^A|D_i=1]-E[Y_i^A]=\overbrace{\left(E[\eta_i|\eta_i<-\frac{\beta}{\phi}]-E[\eta_i]\right)}^{<0}<0.$$

3.1.4 Generalized Income Equation

Reintroducing t subscripts and combining equations (1) and (2), we arrive at the following generalized income equation:

$$Y_{it} = \alpha_t + \beta D_{it} + \eta_i(1 + \phi D_{it}) + \tau_i, \tag{5}$$

where $\alpha_t \equiv \beta_t^A$ and $\beta \equiv (\beta_t^N - \beta_t^A)$, which we assume to be constant over time.[9] Estimation of the parameters β and ϕ is complicated by the fact that D_{it} is endogenous. As described above, households will choose $D_{it} = 1$ if they expect higher earnings in the non-agricultural sector (that is, if $\phi\eta_i > -\beta$).

3.2 Learning

Having established that households make their sorting decision based on η_i, we now discuss what households know about their own η_i, and how this knowledge evolves over time. We assume that households know the population average earning in both sectors (α_t, β), their own absolute advantage (τ_i), and ϕ, but have imperfect information about their comparative advantage (η_i).[10] In particular, we introduce an additive productivity shock, ε_{it}, to η_i in equation (5) and assume that $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2 = 1/h_\varepsilon)$. That is, the household only observes the sum of η_i and ε_{it}, but not either individually. The generalized income equation then becomes:

$$Y_{it} = \alpha_t + \beta D_{it} + (\eta_i + \varepsilon_{it})(1 + \phi D_{it}) + \tau_i \tag{6}$$

Households hold the initial belief that $\eta_i \sim N(m_{i0}, \sigma^2 = 1/h)$; and this belief is refined each period using output observations, Y_{it}. That is, from Y_{it}, households can compute

$$l_{it} = \frac{Y_{it} - \alpha_t - \beta D_{it} - \tau_i}{(1 + \phi D_{it})} = \eta_i + \varepsilon_{it}, \tag{7}$$

a noisy signal of their relative productivity η_i, which is independent of the their period t sectoral choice. Let $l_i^t = (l_{i1}, ..., l_{it})$ denote the history of household i's normalized relative productivity observations through period t. Then, the posterior distribution of η_i given history l_i^t is distributed $N(m_t(l_i^t), 1/h_t)$, where

$$m_t(l_i^t) = \frac{hm_{i0} + h_\varepsilon(l_{i1} + ... + l_{it})}{h + th_\varepsilon}, \quad \text{and} \quad h_t = h + th_\varepsilon \tag{8}$$

[9] As we discuss later, when we estimate the model we will explicitly purge all outcome variables and regressors of variation in means across communities and within communities over time, using community fixed effects that vary across time periods (essentially, community-by-time dummies). These fixed effects will account for changes in relative output prices across sectors, as long as relative prices do not vary within a community in a single year. Under these conditions, extending the analysis to estimate a time-varying β seems of little empirical benefit.

[10] As we explain below, ϕ can be thought of as the value of skills in each sector, where the skills are captured by the comparative advantage component.

Note that the specific learning mechanism proposed here allows households to learn about returns to participating in the non-agricultural sector each period, irrespective of the sector the household has chosen that period. This learning structure is borrowed from Gibbons et al. (2005) who use it to study learning about comparative advantage in a model of occupational choice.[11] The bidirectional churning and convergence observed in the raw data motivates the use of this approach in our setting (see Figures 1 and 2).

The intuition behind this proposed mechanism is that relative productivity, η_i, is an index of fundamental skills which affect productivity in both sectors, but is valued differentially across the two sectors. Assuming that the household knows ϕ but not η_i corresponds to assuming the household knows how much each sector values these skills but not their own skill stock. Accordingly, households can learn about their stock through production in either sector.

For example, suppose that η_i represents the household's managerial skill and that while both sectors reward this skill, the non-agricultural sector rewards it more heavily. The assumptions of the model imply that the household recognizes that the non-agricultural sector rewards managerial ability more than the agricultural sector does; however, the household is unsure of its specific stock of managerial skill.

Of course, an excellent manager might still be able to earn more in the agricultural sector than someone with worse managerial skill (but who is similar in other ways). Therefore, a household that initially believes it is bad at management will operate in the agricultural sector to start, where this lack of managerial skill is less penalized; however, should this household find this period that it is better able to manage its agricultural inputs (for example) than it expected, it will decide to enter the non-agricultural sector next period, knowing that this would be lucrative for a household with strong managerial ability. The mechanism, of course, works in the opposite direction as well. We should note that, to the degree that both sectors reward some skills (e.g., work ethic) *equally*, these skills are represented by τ_i and will affect household income in both sectors, but will not affect the return to switching sectors.

Household i will choose the non-agricultural sector in period t if $E[Y_{it}^N - Y_{it}^A] > 0$, and choose the agricultural sector otherwise. That is, household i will choose the

[11] They, in turn, borrow heavily from the classic development in DeGroot (1970). Please see these previous works for more in depth discussion of this framework.

non-agricultural sector in period t (i.e., $D_{it} = 1$) if and only if $\phi m_i^{t-1} > -\beta$.

3.3 Estimation

Allowing for measurement error in equation (6), our estimating equation is the following:

$$Y_{it} = \alpha_t + \beta D_{it} + (\eta_i + \varepsilon_{it})(1 + \phi D_{it}) + \tau_i + \zeta_{it} \tag{9}$$

where measurement error ζ_{it} is assumed mean independent of sector and input decisions conditional on η_i and τ_i. That is, in particular, we will assume $E(D_{it}|\zeta_{it}, \eta_i, \tau_i) = E(D_{it}|\eta_i, \tau_i)$.

As discussed above, D_{it} will depend on the mean of the household's prior distribution on η_i coming into period t, $m_{i,t-1}$, which we cannot observe. Accordingly, OLS estimates of β will be biased. We now develop a strategy which allows us to consistently estimate β, recover ϕ, and validate the importance of learning dynamics in this empirical context.

In particular, in order to recover consistent estimates of β, we must purge the composite unobserved term, $(\eta_i + \varepsilon_{it})(1 + \phi D_{it}) + \tau_i + \zeta_{it}$, of its correlation with D_{it}. We know from section 3.2 that the portion of $(\eta_i + \varepsilon_{it})$ which correlates with sectoral choices is $m_{i,t-1}$. We will begin by decomposing $m_{i,t-1}$ into two components which have distinct effects on the household's history of sectoral choices. Note that the Bayesian updating of beliefs implies that the mean of the prior distribution is a martingale. That is, the law of motion for $m_{i,t}$ is

$$m_{i,t} = m_{i,t-1} + \xi_{it} \quad \Rightarrow \quad m_{i,t-1} = m_{i0} + \sum_{k=1}^{t-1} \xi_{ik}, \tag{10}$$

where ξ_{it} is a noise term orthogonal to $m_{i,t-1}$. Then, denoting $\tilde{m}_{i,t-1} \equiv \sum_{k=1}^{t-1} \xi_{ik}$ as the sum of the signals received up to period $t-1$, we have

$$Y_{it} = \alpha_t + \beta D_{it} + (m_{i0} + \tilde{m}_{i,t-1} + \omega_{it})(1 + \phi D_{it}) + \upsilon_{it}, \tag{11}$$

where $\upsilon_{it} \equiv \tau_i + \zeta_{it}$ is orthogonal to sectoral choice in period t, D_{it}, by construction and $\omega_{it} \equiv \eta_i + \varepsilon_{it} - (m_{i0} + \tilde{m}_{i,t-1})$ is orthogonal to D_{it} by nature of the martingale structure of $m_{i,t-1}$.

Extending the approaches developed by Chamberlain (1982, 1984), Islam (1995), and Suri (2011), we can overcome the endogeneity of D_{it} by projecting m_{i0} and $\tilde{m}_{i,t-1}$

onto the history of sectoral choices. In particular, the law of motion of the prior, as expressed in equation (10), suggests that the initial belief, m_{i0}, will affect sectoral choices in all periods. On the other hand, the cumulative update, $\tilde{m}_{i,t-1}$, will only affect sectoral choices in period t onwards.

We have five waves of data and therefore four cumulative updates. The projection of the initial belief, m_{i0}, which appears in the estimating equation for all periods, will include the entire history of sectoral choices as follows:[12]

$$m_{i0} = \lambda_0 + \prod_{k=1}^{5}(1 + \lambda_k D_{ik}) - 1 + \psi_{i0} \tag{12}$$

where ψ_{it} is projection error in period t. The projection of each cumulative update, $\tilde{m}_{it}$, includes only the sectoral choices in $t+1$ and onward:

$$\begin{aligned}
\tilde{m}_{i1} &= \theta_{20} + \theta_{22} D_{i2} + \theta_{23} D_{i3} + \theta_{24} D_{i4} + \theta_{25} D_{i5} + \psi_{i1} \\
\tilde{m}_{i2} &= \theta_{30} + \theta_{33} D_{i3} + \theta_{34} D_{i4} + \theta_{35} D_{i5} + \psi_{i2} \\
\tilde{m}_{i3} &= \theta_{40} + \theta_{44} D_{i4} + \theta_{45} D_{i5} + \psi_{i3} \\
\tilde{m}_{i4} &= \theta_{50} + \theta_{55} D_{i5} + \psi_{i4}.
\end{aligned} \tag{13}$$

Note that the martingale structure of the prior on η_i implies that learning is *efficient*; that is, all information the household will use to make its decision at time t is fully summarized in the initial condition m_{i0} and the sum of the orthogonal updates to period $t-1$, $\tilde{m}_{i,t-1}$. In other words, the path by which the prior reaches $m_{i,t-1}$ will not, conditional on $m_{i,t-1}$ itself, affect sectoral choice in period t, D_{it}. Most importantly, the path by which the sum of the updates reaches $\tilde{m}_{i,t-1}$ will not, conditional on both the initial belief m_{i0} and $\tilde{m}_{i,t-1}$ itself, affect D_{it}. Therefore, we need not include past sectoral choices nor the interactions of future sectoral choices in the update projections in (13).

Note also that the relative sizes of h and h_ϵ will determine the degree to which the initial condition, m_{i0}, or subsequent updates, $\tilde{m}_{i,t-1}$, correlate more strongly with choices across periods. We do not explicitly discuss this relationship further as the

[12] If we expand m_0, we get: $m_0 = \lambda_0 + \lambda_1 D_1 + \lambda_2 D_2 + \lambda_3 D_3 + \lambda_4 D_4 + \lambda_5 D_5 + \lambda_{12} D_1 D_2 + \lambda_{13} D_1 D_3 + \lambda_{14} D_1 D_4 + \lambda_{15} D_1 D_5 + \lambda_{23} D_2 D_3 + \lambda_{24} D_2 D_4 + \lambda_{25} D_2 D_5 + \lambda_{34} D_3 D_4 + \lambda_{35} D_3 D_5 + \lambda_{45} D_4 D_5 + \lambda_{123} D_1 D_2 D_3 + \lambda_{124} D_1 D_2 D_4 + \lambda_{125} D_1 D_2 D_5 + \lambda_{134} D_1 D_3 D_4 + \lambda_{135} D_1 D_3 D_5 + \lambda_{145} D_1 D_4 D_5 + \lambda_{234} D_2 D_3 D_4 + \lambda_{235} D_2 D_3 D_5 + \lambda_{245} D_2 D_4 D_5 + \lambda_{345} D_3 D_4 D_5 + \lambda_{1234} D_1 D_2 D_3 D_4 + \lambda_{1235} D_1 D_2 D_3 D_5 + \lambda_{1245} D_1 D_2 D_4 D_5 + \lambda_{1345} D_1 D_3 D_4 D_5 + \lambda_{2345} D_2 D_3 D_4 D_5 + \lambda_{12345} D_1 D_2 D_3 D_4 D_5 + \psi_{i0}$, where $\lambda_{ijklm} = \lambda_i \lambda_j \lambda_k \lambda_l \lambda_m$.

estimation will approach this issue agnostically. That is, the estimation will allow the data to show (in the projection coefficients) the degree to which initial conditions and subsequent updates affect choices without restricting *a priori* the relative magnitudes of these correlations. If, for example, a large dispersion in the initial conditions effectively makes their impact on production decisions negligible, the coefficients in equation (12) will be estimated as indistinguishable from 0, while those from the equations in (13) might be estimated with larger magnitudes and more precision.

Plugging projections (12) and (13) into equation (11), and grouping terms, we can now express each Y_{it} as a function of all sectoral choices as shown below.[13]

$$\begin{aligned}
Y_{i1} =& \alpha_1 + \beta D_{i1} + (\lambda_0 + \prod_{t=1}^{5}(1+\lambda_t D_{it}) - 1)(1+\phi D_{i1}) + \\
& (\omega_{i1} + \psi_{i0})(1+\phi D_{i1}) + \nu_{i1} \\
Y_{i2} =& \alpha_2 + \beta D_{i2} + (\lambda_0 + \prod_{t=1}^{5}(1+\lambda_t D_{it}) - 1 + \theta_{20} + \sum_{t=2}^{5}\theta_{2t}D_{it})(1+\phi D_{i2}) + \\
& (\omega_{i2} + \psi_{i0} + \psi_{i1})(1+\phi D_{i2}) + \nu_{i2} \\
Y_{i3} =& \alpha_3 + \beta D_{i3} + (\lambda_0 + \prod_{t=1}^{5}(1+\lambda_t D_{it}) - 1 + \theta_{30} + \sum_{t=3}^{5}\theta_{3t}D_{it})(1+\phi D_{i3}) + \\
& (\omega_{i3} + \psi_{i0} + \psi_{i1} + \psi_{i2})(1+\phi D_{i3}) + \nu_{i3} \\
Y_{i4} =& \alpha_4 + \beta D_{i4} + (\lambda_0 + \prod_{t=1}^{5}(1+\lambda_t D_{it}) - 1 + \theta_{40} + \sum_{t=4}^{5}\theta_{4t}D_{it})(1+\phi D_{i4}) + \\
& (\omega_{i4} + \psi_{i0} + \psi_{i1} + \psi_{i2} + \psi_{i3})(1+\phi D_{i4}) + \nu_{i4} \\
Y_{i5} =& \alpha_5 + \beta D_{i5} + (\lambda_0 + \prod_{t=1}^{5}(1+\lambda_t D_{it}) - 1 + \theta_{50} + \theta_{55}D_{i5})(1+\phi D_{i5}) + \\
& (\omega_{i5} + \psi_{i0} + \psi_{i1} + \psi_{i2} + \psi_{i3} + \psi_{i4})(1+\phi D_{i5}) + \nu_{i5} \qquad (14)
\end{aligned}$$

This results in the following reduced form regressions, where income in each period depends on all five D_{it} as well as their double, triple, quadruple, and quintuple interactions:

$$Y_{it} = \gamma_0^t + \prod_{k=1}^{5}(1+\gamma_k^t D_{ik}) - 1 + \nu_{it}. \qquad (15)$$

[13] It is important that we properly specify the projections in (12) and (13). That is, we must include all necessary elements of the history of sectoral choices in order to ensure that the projection errors (ψ) are, indeed, orthogonal to current choices.

If we define $\gamma^t_{ijklm} \equiv \gamma^t_i \gamma^t_j \gamma^t_k \gamma^t_l \gamma^t_m$, each equation has 32 reduced form coefficients to be estimated.[14] Following Chamberlain (1982, 1984), we will first estimate these reduced form coefficients by seemingly unrelated regressions (SUR) and then estimate from these coefficients the structural parameters of the model using minimum distance. After normalizing each of the intercepts in equations (12), (13), and (15),[15] there are 43 structural parameters of the model (31 λ coefficients, 10 θ coefficients, β, and ϕ), to be identified from the 155 reduced form coefficients using the minimum distance restrictions implied by the model. The minimum distance restrictions are reported in Appendix section B.1. Identification requires that every single possible trajectory of sectoral choice is represented in the data, which was shown to be the case in Figure 1. There are on average 121 households per switching history trajectory.

For simplicity, we have not included any covariates in the exposition above, although one could argue that there are household-level characteristics which are correlated with household income and also sectoral choice D_{it}. Though the inclusion of covariates will affect reduced form expressions (15), it will not affect the relationships between the reduced form coefficients on the choices and the structural parameters of interest. We control for community fixed effects and household composition variables (number of household members, number of women aged 15-59, and number of men aged 15-59) in each equation of the first stage SUR estimation. Controlling for household size is important because our income variable sums across all members. By

[14] Expanding, we obtain: $Y_{it} = \gamma^t_0 + \gamma^t_1 D_1 + \gamma^t_2 D_2 + \gamma^t_3 D_3 + \gamma^t_4 D_4 + \gamma^t_5 D_5 + \gamma^t_{12} D_1 D_2 + \gamma^t_{13} D_1 D_3 + \gamma^t_{14} D_1 D_4 + \gamma^t_{15} D_1 D_5 + \gamma^t_{23} D_2 D_3 + \gamma^t_{24} D_2 D_4 + \gamma^t_{25} D_2 D_5 + \gamma^t_{34} D_3 D_4 + \gamma^t_{35} D_3 D_5 + \gamma^t_{45} D_4 D_5 + \gamma^t_{123} D_1 D_2 D_3 + \gamma^t_{124} D_1 D_2 D_4 + \gamma^t_{125} D_1 D_2 D_5 + \gamma^t_{134} D_1 D_3 D_4 + \gamma^t_{135} D_1 D_3 D_5 + \gamma^t_{145} D_1 D_4 D_5 + \gamma^t_{234} D_2 D_3 D_4 + \gamma^t_{235} D_2 D_3 D_5 + \gamma^t_{245} D_2 D_4 D_5 + \gamma^t_{345} D_3 D_4 D_5 + \gamma^t_{1234} D_1 D_2 D_3 D_4 + \gamma^t_{1235} D_1 D_2 D_3 D_5 + \gamma^t_{1245} D_1 D_2 D_4 D_5 + \gamma^t_{1345} D_1 D_3 D_4 D_5 + \gamma^t_{2345} D_2 D_3 D_4 D_5 + \gamma^t_{12345} D_1 D_2 D_3 D_4 D_5 + \psi_{i0}$, where $\gamma^t_{ijklm} = \gamma^t_i \gamma^t_j \gamma^t_k \gamma^t_l \gamma^t_m$.

[15] We normalize the intercepts such that the estimates of the projection coefficients are mean zero, as follows:

$$\begin{aligned}
\lambda_0 &= -\bar{\mathbf{D}}\Lambda^T = -\lambda_1 \bar{D}_1 - \lambda_2 \bar{D}_2 - ... - \lambda_{12} \overline{D_1 D_2} - ... - \lambda_{12345} \overline{D_1 D_2 D_3 D_4 D_5} \\
\theta_{20} &= -\theta_{22} \bar{D}_2 - \theta_{23} \bar{D}_3 - \theta_{24} \bar{D}_4 - \theta_{25} \bar{D}_5 \\
\theta_{30} &= -\theta_{33} \bar{D}_3 - \theta_{34} \bar{D}_4 - \theta_{35} \bar{D}_5 \\
\theta_{40} &= -\theta_{44} \bar{D}_4 - \theta_{45} \bar{D}_5 \\
\theta_{50} &= -\theta_{55} \bar{D}_5,
\end{aligned}$$

where $\bar{D}_t$ is the sample mean of the non-agricultural dummy in period t. $\bar{\mathbf{D}}$ is a row vector of the sample mean of the dummies and the sample mean of all interactions of these dummies: $\bar{\mathbf{D}} = \begin{pmatrix} \bar{D}_1 & \bar{D}_2 & ... & \overline{D_1 D_2} & ... & \overline{D_1 D_2 D_3 D_4 D_5} \end{pmatrix}$. Λ^T is the column vector of the associated coefficients: $\Lambda^T = \begin{pmatrix} \lambda_1 & \lambda_2 & ... & \lambda_{12} & ... & \lambda_{12345} \end{pmatrix}^T$. Note that $\bar{D}_1 \bar{D}_2 \neq \overline{D_1 D_2}$ in general. An analogous normalization exercise is conducted for the reduced form regressions in (15).

allowing each community effect to vary across waves, we are also able to account for local community-level demand shocks and price fluctuations that may affect switching decisions but do not convey any information about household-level perceptions of relative ability across sectors. Communities correspond to Enumeration Areas (EAs) of the national population census, which are subdivisions of villages and cities. Each EA encompasses roughly 80-120 households as of 1993. The IFLS targeted a subset of households within more than 300 of these EAs, selected randomly. There are 310 communities in our sample with an average of 15 (median of 16) households surveyed per community. Hence, the community-by-wave fixed effects absorb very local shocks.

3.4 Identification

3.4.1 Identifying Assumptions

We obtain estimates of the structural parameters through the minimum distance restrictions, which map 43 structural parameters to 155 reduced form coefficients. When we plug in all of the projections into the generalized earnings equation to create equation (14), it can be seen that, in each period, the unobservable error term includes the product of D_{it} and $\omega_{it} + \sum_{k=0}^{t-1} \psi_{ik}$. We thus must assume that $(\omega_{i1} + \psi_{i0})$ is uncorrelated with D_{i1}, $(\omega_{i2} + \psi_{i0} + \psi_{i1})$ is uncorrelated with D_{i2}, and so on.

Given that the ψ_{it} terms are the projection error terms in (13), they are orthogonal to the relevant sectoral choice indicators by construction. However, we also require that the other component, $\omega_{it} \equiv \eta_i + \varepsilon_{it} - (m_{i0} + \tilde{m}_{i,t-1})$, is orthogonal to D_{it}. Recall that ε_{it} represents the productivity shock in period t. We are therefore assuming sequential exogeneity of the current period's productivity shock. Productivity shocks in a given period are allowed to influence decisions in future periods (as households use them to update their beliefs about η_i), but decisions in a given period cannot be influenced by productivity shocks in future periods. If households can predict future productivity shocks (e.g., good rains next year, infrastructure expansion in the village in the near future, rising demand for a specific good in village) and respond to them in their sector decisions, the update projection, as specified, will not fully account for the endogeneity in these choices. (Note, however, that these future predictions only matter if they are household-specific because community by time fixed effects are projected off in the first stage.) Specifically, there are no λ's and θ's included in the estimation to capture correlations between future idiosyncratic shocks and past household sectoral choices. These correlations are assumed to be zero in order to be able to identify the model with multiple endogenous choices and a small number of periods. Specifically, relaxing this assumption further in a model with heterogeneous returns would make the model not fully identified.[16]

In our theoretical model, the main source of endogeneity in the generalized income equation (6) is the fact that households sort into sectors based on their η_i and learn about their η_i over time. However, the empirical strategy outlined above will recover consistent estimates of β and ϕ under alternative models, as long as they satisfy sequential exogeneity. Suppose, for example, that households do not learn about their η_i over time, but need to save in order to overcome entry or switching costs before they can change sectors. Alternatively, households might not learn about their η_i over time but instead might be able to change their η_i through skill accumulation, as would be the case in a learning by doing model. In both of these examples, as long as sequential exogeneity holds, we can still recover consistent estimates of β and ϕ. However, estimates of θ (which govern how dynamics in relative earning potential η_i relate to future sectoral choices), along with the descriptive evidence from Figures 1 and 2, will allow us to detect whether one of these alternative models appears to be more plausible. We discuss this in more detail in section 4.5.

3.4.2 Identification Intuition

Identification of the structural parameters, such as β, ϕ, the λ's and θ's, comes from a comparison of the income trajectories across households with different sectoral choice histories. That is, we observe in the data the conditional sample mean of income for each sector choice history in each period (i.e. $E(Y_{it}|D_{i1}, D_{i2}, D_{i3}, D_{i4}, D_{i5})$). The econometric strategy uses variation in these means, as well as their evolution over time, across households with different sectoral histories, to recover the structural parameters of interest.

To help clarify the intuition behind the identification, we consider the simplified two-period version of the model described above, with generalized income equations:

[16] Though this paper contributes to the literature on panel data estimators of correlated random coefficients models by relaxing the strict exogeneity assumption to sequential exogeneity to allow for dynamics, we leave it to future work to relax the sequential exogeneity assumption further to allow for correlations of regressors with both past and future shocks.

$$\begin{aligned}
Y_{i1} &= \alpha_1 + \beta D_{i1} + (m_{i0} + \omega_{i1})(1 + \phi D_{i1}) + v_{i1} \\
Y_{i2} &= \alpha_2 + \beta D_{i2} + (m_{i0} + \tilde{m}_{i,1} + \omega_{i2})(1 + \phi D_{i2}) + v_{i2}.
\end{aligned}$$

The projections are then:

$$\begin{aligned}
m_{i0} &= \lambda_0 + \prod_{k=1}^{2}(1 - \lambda_k D_{ik}) - 1 + \psi_{i0} \\
m_{i0} &= \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_1 \lambda_2 D_{i1} D_{i2} + \psi_{i0} \\
m_{i0} &= \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_{12} D_{i1} D_{i2} + \psi_{i0} \\
\tilde{m}_{i1} &= \theta_{20} + \theta_{22} D_{i2} + \psi_{i1}. \qquad (16)
\end{aligned}$$

Replacing the projections in the income equations and grouping terms allows us to obtain the following reduced form equations:

$$\begin{aligned}
Y_{i1} =& \alpha_1 + \beta D_{i1} + (\lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_{12} D_{i1} D_{i2} + \psi_{i0} + \omega_{i1})(1 + \phi D_{i1}) + v_{i1} \\
Y_{i1} =& \overbrace{\alpha_1 + \lambda_0}^{\gamma_0^1} + \overbrace{[\beta + (1+\phi)\lambda_1 + \lambda_0 \phi]}^{\gamma_1^1} D_{i1} + \overbrace{[\lambda_2]}^{\gamma_2^1} D_{i2} + \overbrace{[(1+\phi)\lambda_{12} + \lambda_2 \phi]}^{\gamma_{12}^1} D_{i1} D_{i2} \\
& + \underbrace{(\psi_{i0} + \omega_{i1})}_{\perp D_{i1}} \phi D_{i1} + \underbrace{\psi_{i0} + \omega_{i1} + v_{i1}}_{u_{i1}} \\
Y_{i1} =& \gamma_0^1 + \gamma_1^1 D_{i1} + \gamma_2^1 D_{i2} + \gamma_{12}^1 D_{i1} D_{i2} + u_{i1} \qquad (17)
\end{aligned}$$

$$\begin{aligned}
Y_{i2} =& \alpha_2 + \beta D_{i2} + \\
& (\lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_{12} D_{i1} D_{i2} + \psi_{i0} + \theta_{20} + \theta_{22} D_{i2} + \psi_{i1} + \omega_{i2})(1 + \phi D_{i2}) + v_{i2} \\
Y_{i2} =& \overbrace{\alpha_2 + \lambda_0 + \theta_{20}}^{\gamma_0^2} + \overbrace{[\lambda_1]}^{\gamma_1^2} D_{i1} + \overbrace{[\beta + (1+\phi)(\lambda_2 + \theta_{22}) + \phi(\lambda_0 + \theta_{20})]}^{\gamma_2^2} D_{i2} + \\
& \underbrace{[(1+\phi)\lambda_{12} + \lambda_1 \phi]}_{\gamma_{12}^2} D_{i1} D_{i2} + \underbrace{(\psi_{i0} + \psi_{i1} + \omega_{i2})}_{\perp D_{i2}} \phi D_{i2} + \underbrace{\psi_{i0} + \psi_{i1} + \omega_{i2} + v_{i2}}_{u_{i2}} \\
Y_{i2} =& \gamma_0^2 + \gamma_1^2 D_{i1} + \gamma_2^2 D_{i2} + \gamma_{12}^2 D_{i1} D_{i2} + u_{i2}. \qquad (18)
\end{aligned}$$

These reduced form coefficients (γ's) represent differences in income between four different groups of households: those that stay in the non-agricultural sector in both periods ($D_{i1} = 1, D_{i2} = 1$), stay out of the non-agricultural sector in both periods

$(D_{i1}=0, D_{i2}=0)$, switch into the non-agricultural sector in period 2 $(D_{i1}=0, D_{i2}=1)$, and switch out of the non-agricultural sector in period 2 $(D_{i1}=1, D_{i2}=0)$. Specifically, it can be shown that

$$\begin{aligned}
\gamma_1^1 =& E(Y_{i1}|D_{i1}=1, D_{i2}=0) - E(Y_{i1}|D_{i1}=0, D_{i2}=0) \\
\gamma_2^1 =& E(Y_{i1}|D_{i1}=0, D_{i2}=1) - E(Y_{i1}|D_{i1}=0, D_{i2}=0) \\
\gamma_{12}^1 =& E(Y_{i1}|D_{i1}=1, D_{i2}=1) - E(Y_{i1}|D_{i1}=1, D_{i2}=0) \\
& - [E(Y_{i1}|D_{i1}=0, D_{i2}=1) - E(Y_{i1}|D_{i1}=0, D_{i2}=0)] \\
\gamma_1^2 =& E(Y_{i2}|D_{i1}=1, D_{i2}=0) - E(Y_{i2}|D_{i1}=0, D_{i2}=0) \\
\gamma_2^2 =& E(Y_{i2}|D_{i1}=0, D_{i2}=1) - E(Y_{i2}|D_{i1}=0, D_{i2}=0) \\
\gamma_{12}^2 =& E(Y_{i2}|D_{i1}=1, D_{i2}=1) - E(Y_{i2}|D_{i1}=1, D_{i2}=0) \\
& - [E(Y_{i2}|D_{i1}=0, D_{i2=1}) - E(Y_{i2}|D_{i1}=0, D_{i2}=0)] \, . \qquad (19)
\end{aligned}$$

As with the 5-period version of the model, equations (17) and (18) are estimated by a seemingly unrelated regression which allows us to recover estimates for the γ coefficients. We then estimate the structural parameters through minimum distance where the minimum distance restrictions are as follows.[17]

$$\begin{aligned}
\gamma_1^1 &= \beta + (1+\phi)\lambda_1 + \lambda_0\phi \\
\gamma_2^1 &= \lambda_2 \\
\gamma_{12}^1 &= (1+\phi)\lambda_{12} + \lambda_2\phi \\
\gamma_1^2 &= \lambda_1 \\
\gamma_2^2 &= \beta + (1+\phi)(\lambda_2 + \theta_{22}) + \phi(\lambda_0 + \theta_{20}) \\
\gamma_{12}^2 &= (1+\phi)\lambda_{12} + \lambda_1\phi. \qquad (20)
\end{aligned}$$

The minimum distance restrictions show how β, ϕ, the λ's, and the θ's are recov-

[17] Although it appears that there are 8 structural parameters to be estimated from 6 equations, we impose the following normalizations:

$$\begin{aligned}
\lambda_0 &= -\lambda_1\overline{D_{i1}} - \lambda_2\overline{D_{i2}} - \lambda_{12}\overline{D_{i1}D_{i2}} \\
\theta_0 &= -\theta_2\overline{D_{i2}} \quad ,
\end{aligned}$$

where $\overline{D_{ij}}$ is the average sectoral decision in period j and $\overline{D_{i1}D_{i2}}$ is the average of the interaction between the sectoral decisions in periods 1 and 2. These normalizations will make estimates of the projection coefficients mean zero and reduce the number of projection coefficients to be estimated by 2, improving efficiency at no real loss of generality or interpretation.

ered from the reduced form (γ) coefficients in equations (17) and (18). For example, the average return to non-agricultural work (β) is identified by the minimum distance restrictions for γ_1^1 (the difference in period 1 income between those who switch out and those who stay out) and γ_2^2 (the difference in period 2 income between those who switch in and those who stay out). Note that γ_1^1 and γ_2^2 are not solely determined by β. For instance, a large positive γ_1^1 could be due to a large positive β or a large positive $(1+\phi)\lambda_1$. Because λ_1 represents the difference in m_{i0} between those who switch out and those who stay out (see equation (16)), the latter could result from positive selection ($\phi > 0$), which would lead to the switch-out households (who are in the non-agricultural sector in period 1) having higher η_i than the stay-out households (who are in the agricultural sector in period 1) and therefore a positive $(1+\phi)\lambda_1$. Alternatively, selection based on comparative advantage ($\phi < -1$) would lead to the switch-out households having lower η_i than the stay-out households ($\lambda_1 < 0$) and once again a positive $(1+\phi)\lambda_1$.

To illustrate the intuition behind how ϕ is identified, we conduct three simulations using different values of ϕ (but the same values for β and the same distribution of η_i). To focus on the identification of ϕ, we shut down the learning mechanism by setting all θ coefficients to zero, which assumes a household's perception of η_i does not change across waves. For each simulation, we calculate the average income for each of the four groups described above (stay in, stay out, switch in, and switch out) in each period and plot the trajectory of average income, expressed as a deviation from the period-specific mean, for each of the four groups. In Figure 3, panel A illustrates the case of positive selection ($\phi > 0$), panel B illustrates negative selection ($-1 < \phi < 0$), and panel C illustrates sorting based on comparative advantage ($\phi < -1$). All panels assume a positive return to the non-agricultural sector (β).

Figure 3 demonstrates that different values of ϕ imply different patterns of income trajectories and income differences across the four groups. In panel A, when there is positive selection, those who stay in (yellow triangles) have higher period 1 income than those who switch out (green squares). This is because those who switch out are more marginal and have lower η_i on average. On the other hand, when there is negative selection (in panel B), those who switch out have (slightly) higher period 1 income than those who stay in. This is because negative selection implies that those with lower η_i are more likely to enter the non-agricultural sector, which means that the more marginal households (who switch out) should have higher η_i on average (and under negative selection the coefficient on η_i in the generalized income equation is positive for households in the non-agricultural sector). Finally, in panel C, we also see that those who switch out have lower period 1 income than those who stay in, similar to the case of positive selection. Under sorting based on comparative advantage, those with low η_i choose the non-agricultural sector, which means the more marginal switch-out households have higher η_i. Because $\phi < 1$, however, the coefficient on η_i is negative for those who are in the non-agricultural sector, which leads to higher income among the stay-in households that have lower η_i.

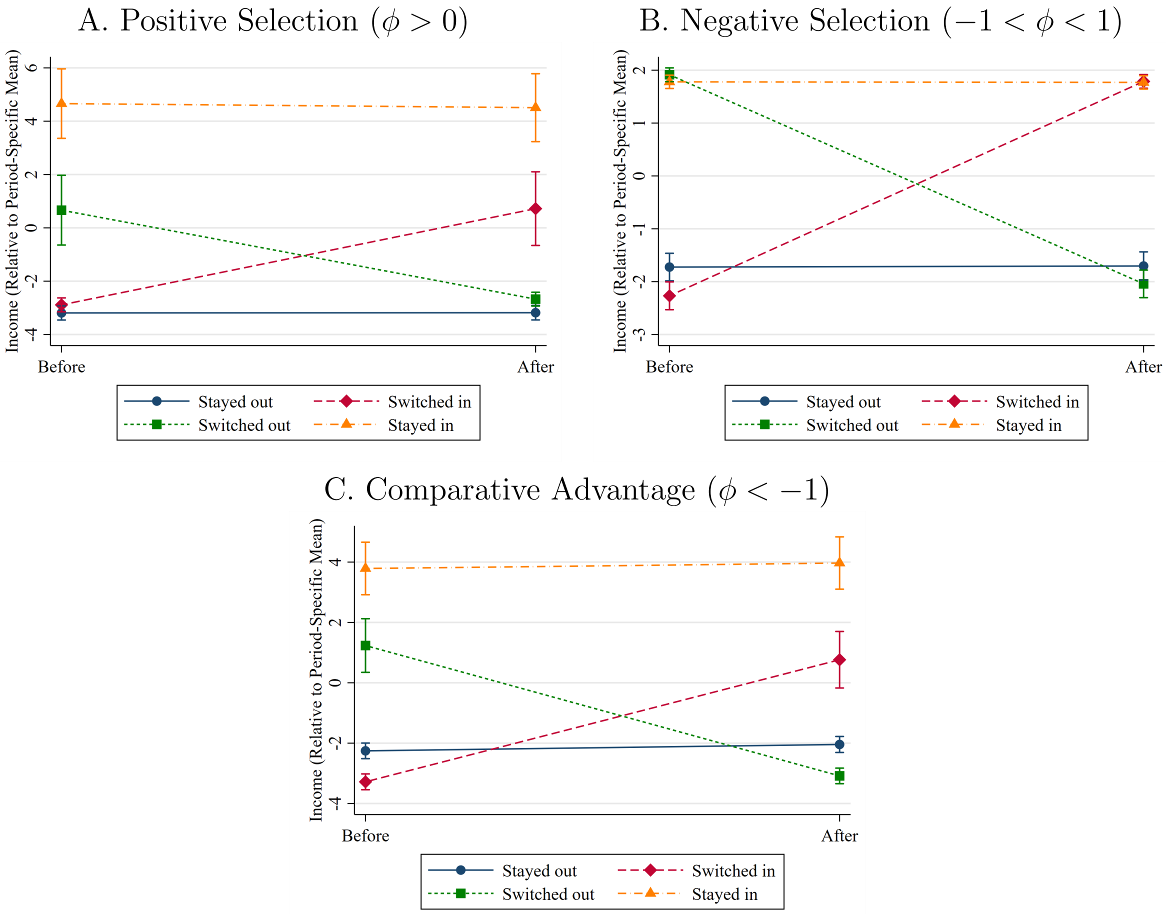
It is also important to compare the period 1 income of those who switch in (red diamonds) and those who stay out (blue circles). Under positive selection, those who switch in have higher period 1 income than those who stay out because they have higher η_i. Under negative selection and comparative advantage, the opposite is true, for reasons similar to those outlined in the previous paragraph.

Differences in period 2 income also contribute to the identification of ϕ. For example, comparing the period 2 income of those who switch in with those who stay in, we see in Panel A that period 2 income of the stay in group is higher. This is because those who stay in must have higher η_i on average than those who are more marginal and therefore switch in later. In panel B, these two groups have almost identical period 2 income, though that of the switch in group (who are more marginal and therefore have higher η_i under negative selection) is slightly higher. In panel C, period 2 income for those who switch in is lower than for those who stay in: those who switch in are more marginal and therefore have higher η_i on average under comparative advantage sorting, which translates into lower income due to $\phi < -1$. Similar reasoning can explain why period 2 income is higher for those who switch out than for those who stay out under positive selection, while the opposite is true under negative selection and comparative advantage.[18]

To present some preliminary analysis and preview what we find, we generate a version of Figure 3 that uses our actual data. Specifically, in Figure 4, we plot the evolution of realized incomes, after projecting off community-by-year fixed effects

[18] There are several group comparisons that cannot be signed solely based on the nature of the sorting process. For example, in panel A, positive ϕ does not necessarily determine whether the income differences between switch-out and switch-in households should be positive or negative in either period, but the specific values used in this simulation lead to the switch-out households having higher income in period 1 but lower income in period 2. The comparisons highlighted above, however, are what help identify ϕ.

Figure 3: Income by Switch Status, Simulations



A. Positive Selection ($\phi > 0$)

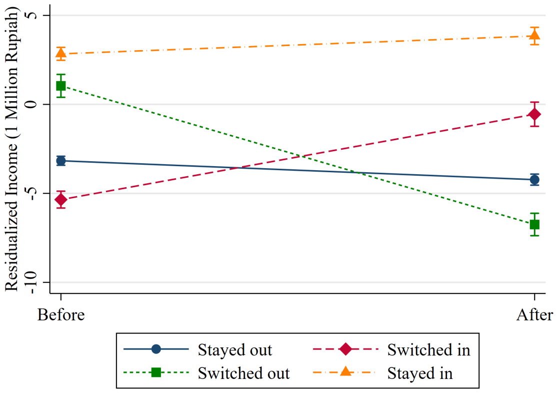
B. Negative Selection ($-1 < \phi < 1$)

C. Comparative Advantage ($\phi < -1$)

Notes: “Stayed out” includes households in agriculture in both period 1 and 2. “Switched In” includes households in agriculture in period 1 and the non-agricultural sector in period 2. “Switched Out” includes households in the non-agricultural sector in period 1 and agriculture in period 2. “Stayed In” includes households in the non-agricultural sector in both periods. Error bars denote 95% confidence intervals. We use $\beta = 4$ and a normally distributed η with mean 0 and standard deviation 3 for all cases, $\phi = 5$ for positive selection, $\phi = -0.9$ for negative selection, and $\phi = -5$ for selection based on comparative advantage.

and household composition controls, for four groups of households: period t non-agricultural households who stay in the non-agricultural sector in $t+1$, period t non-agricultural households who switch to agriculture in $t+1$, period t agricultural households who switch into the non-agricultural sector in $t+1$, and period t agricultural households who stay in agriculture in $t+1$. To generate this figure, we include all transitions between waves (such that each household appears multiple times, potentially in different groups), and calculate average residualized income across all households in each group, in the "before" period (t) and the "after" period ($t+1$). The patterns in the data are similar to those documented in panel C of Figure 3, the case of comparative advantage. As we discuss below, our estimated ϕ is indeed consistent with sorting based on comparative advantage, particularly once dynamics are allowed as in our preferred DCRC model.

Figure 4: Income by Switch Status, Data



Notes: Residualized income is calculated by taking the residuals of wave-by-wave regressions of income on community fixed effects and household composition controls. This figure treats each household transition as a separate observation, which means that each household has four observations (one for each transition: 1993-1997, 1997-2000, 2000-2007, and 2007-2014). "Stayed out" includes households in agriculture in both t and $t+1$. "Switched In" includes households in agriculture in t and the non-agricultural sector in $t+1$. "Switched Out" includes households in the non-agricultural sector in t and agriculture in $t+1$. "Stayed In" includes households in the non-agricultural sector in both t and $t+1$. Error bars denote 95% confidence intervals.

The minimum distance restrictions in the two-period case (20) also shed light on the identification of the λ and θ coefficients. For instance, two of the λ coefficients are simply equal to the reduced form coefficients γ_2^1 and γ_1^2. That is, the difference in m_{i0} across those who switch out and those who stay out (λ_2) is equal to the difference in period 1 income across those two groups (γ_2^1). Similarly, the difference in m_{i0} for those who switch in and those who stay out (λ_1) is equal to the period 2 income difference across those two groups (γ_1^2).

The learning coefficient is identified by the minimum distance restriction for γ_2^2 (the fifth equation in (20)), which captures the difference in period 2 income between those who switch in and those who stay out. The period 2 income of these two groups differs for several reasons. First, there is an average income gap between the non-agricultural and agricultural sectors (β). In addition, there are underlying differences in η_i across the two groups because those who switch in are closer to the sectoral choice cutoff. These differences in η_i imply there are differences in the m_{i0} (captured by the λ's) and differences in the learning update $\tilde{m}_{i1}$ (captured by the θ's), and the latter component is what informs us about the the learning process. If the magnitude of γ_2^2 is not equal to what we would predict based only on β and the underlying differences in m_{i0} (i.e., $\beta + (1+\phi)\lambda_2 + \phi\lambda_0$), this indicates that the relationship between latent heterogeneity in relative earnings and future sectoral choices is dynamic and the discrepancy generates our estimates of the θ coefficients.

While it is obviously more difficult to demonstrate the precise variation that identifies each of the structural coefficients in the 5-period model, the intuition remains the same: the coefficients are identified by comparing the income trajectories of households with different switching behavior.

3.4.3 Simulation to Check Identification

To demonstrate that the model is identified, we generate data following the learning model and estimate the model using the minimum distance procedure. We simulate a data set of 10,000 observations. We could chose any values, but for simplicity, we use the original data to parametrize β^A and β^N as the average income over all waves for household observations in agriculture and households that are not in agriculture, respectively. These two coefficients allow us to determine β, η_i, τ_i, and ϕ. We assume that the productivity shock is normally distributed with mean 0 and standard deviation 10 to generate enough observations for every possible switching histories. We assume that the initial beliefs of the households' own comparative advantage, $m_{i,0}$, are normally distributed and have a correlation of 0.5 with η_i in the population. With all this in hand, we let the equations of the model determine $E[Y_{i,t}^N|D_{i,t}]$, $E[Y_{i,t}^A|D_{i,t}]$, and $D_{i,t}$ for every household for every period.[19] Given this parametrization, the true $\beta = 9.19$ and true $\phi = -2.73$. Estimating the model by minimum distance using the generated data yields $\hat{\beta} = 8.91$ (SE 0.56) and $\hat{\phi} = -2.73$ (SE 0.11). Both estimated coefficients are not statistically different from the true population parameters confirming that the model is indeed identified (p-value=0.62 for $H_0 : \hat{\beta} = \beta$ vs $H_a : \hat{\beta} \neq \beta$ and p-value≈1 for $H_0 : \hat{\phi} = \phi$ vs $H_a : \hat{\phi} \neq \phi$).

3.5 Nested Models

The model described above is a DCRC model that allows for heterogeneous returns to the non-agricultural sector and dynamic relationships between income innovations in the current period and future sectoral sorting decisions. In addition to estimating this preferred model, we also estimate nested models which impose additional restrictions on the relationships between η_i and the endogenous choices, D_{it}. Specifically, we estimate a correlated random coefficients (CRC) model of heterogeneous returns to the non-agricultural sector with static relationships between income innovations and sectoral choices (i.e., strict exogeneity) and a simple fixed effects model with homogeneous returns and no dynamics, which is equivalent to a correlated random effects (CRE) model.

3.5.1 Heterogeneous Returns with Perfect Information: CRC

In the CRC model, households are assumed to have perfect information about their relative productivity η_i, which means there is no longer an additive productivity shock, ε_{it}, nor any updating of expectations about η_i. With perfect information, the model becomes a static CRC model. Models of this sort have been used to study agricultural technology adoption (Suri, 2011) and returns to schooling (Heckman and Vytlacil, 1998).

The estimating equation is nearly the same as in the DCRC model:

$$Y_{it} = \alpha_t + \beta D_{it} + \eta_i(1 + \phi D_{it}) + v_{it}.$$

However, now the household is assumed to have perfect information about its relative productivity, η_i; hence, there is no longer an additive productivity shock, ε_{it}. There-

[19] We simulate three periods for simplicity but the model only becomes more overidentified with a larger number of periods.

fore, the relationship between η_i and the history of sectoral choices is static. Note, however, that v_{it} could still include exogenous, transitory shocks that shift households from period to period above and below the cutoff for non-agricultural entry. That is, households will sort into a particular sectoral choice history on the basis of η_i and their expectations of Y_{it}^A and Y_{it}^N; however, these expectations will not evolve over time as they do in the imperfect information case.

Accordingly, we need only a single projection in which we project η_i onto the sectoral choice dummies and all of their interactions, as in equation (12):

$$\eta_i = \lambda_0 + \prod_{k=1}^{5}(1 + \lambda_k D_{ik}) - 1 + \psi_{i0}.$$

Because households no longer update their expectations over time, the cumulative updates $\tilde{m}_{it}$ are irrelevant, which means that the θ coefficients in equation (13) are all equal to zero. The CRC model is therefore a restricted version of the DCRC model where all θ coefficients are assumed to be zero. This model has 33 (instead of 43) structural parameters that we estimate from 155 reduced form coefficients (γ) using minimum distance.

3.5.2 Homogeneous Returns with Perfect Information: CRE

In the CRE model, in addition to perfect information about η_i, households are assumed to have homogeneous returns. Because a household's return to the non-agricultural sector no longer depends on their relative productivity η_i, ϕ is assumed to be zero. This amounts to assuming that the data generating process is a simple household fixed effects or CRE model. Conditionally mean independent productivity shocks are what drive sectoral switching and therefore the variation used for identification. Under these assumptions, the estimating equation becomes

$$Y_{it} = \alpha_t + \beta D_{it} + \eta_i + v_{it}.$$

We now need only a single projection of η_i on the five sectoral choice dummies:

$$\eta_i \quad = \quad \lambda_0 + \lambda_1 D_{i1} + \lambda_2 D_{i2} + \lambda_3 D_{i3} + \lambda_4 D_{i4} + \lambda_5 D_{i5} + \psi_{i0}.$$

Note that we have not included the interactions of sectoral choice dummies across periods. This is because, once we assume that η_i has no effect on the return to the non-agricultural sector, the changes in choices over time will no longer depend on the initial belief, though the choice in each period still will. As in the CRC model above, all θ coefficients are assumed to be equal to zero. Therefore, the CRE model is a restricted version of the DCRC model where ϕ, all θ coefficients in equation (13), and all λ coefficients in equation (15) – except for λ_1, λ_2, λ_3, λ_4, λ_5 – are assumed to be zero. This model has 6 structural parameters which we estimate from 25 reduced form coefficients using minimum distance.

In the existing literature, several studies identify the returns to a particular sector using sector switchers or households that participate in both sectors at once (Alvarez, 2020; Alvarez-Cuadrado et al., 2019; Herrendorf and Schoellman, 2018; Hicks et al., 2017). The assumed data generating process underlying these identification strategies is similar to the CRE model, in which all switchers have the same return. The CRC model relaxes this assumption by allowing heterogeneous returns across households, where households of the same type (defined by a sequence of sectoral choices) have the same type-specific return. Finally, the DCRC model that we use goes a step further and allows the relationship between type and returns to evolve over time.

4 Results

4.1 Structural Minimum Distance Estimates

In Table 2, we present the minimum distance estimates of β and ϕ. The first column displays estimates from our preferred DCRC model. We estimate an average return to the non-agricultural sector (β) of approximately 5.9 million rupiah, which is about two thirds of the average household income in 1993. ϕ is estimated to be -5.01. Significantly less than 1, this estimate implies that households sort based on comparative advantage in this context, consistent with the patterns shown in Figure 4. That is, households that are more productive in the non-agricultural sector tend to be less productive in agriculture and vice versa.

We next compare our preferred estimates of β and ϕ to those from the two nested models: the CRC model of heterogeneous returns and perfect information, and the CRE model of homogeneous returns and perfect information. Both restricted models substantially overestimate the average return to the non-agricultural sector. The CRC (column 2) and CRE model (column 3) estimate a return 37% and 15% larger than the DCRC model, respectively. The CRE model assumes that $\phi = 0$. Hence, both restricted models also greatly underestimate the degree of heterogeneity captured by ϕ.[20] Moreover, most of the estimates of the additional λ and θ parameters which appear in the DCRC model but are assumed 0 in the CRC and CRE are statistically significant, indicating a rejection of the nested models.

In short, ignoring heterogeneity in returns and dynamics results in an overestimation of the average return to the non-agricultural sector and the inability to capture the extent to which households sort based on comparative advantage. Notably, in the DCRC model, ϕ is significantly less than one, while it is forced to be zero in the CRC model and it is 50% smaller in magnitude in the CRC model. It is clear that the additional flexibility of the DCRC is needed in order to better fit the patterns in the data shown in Figure 4.

Table 2: Structural Estimates

	Specification		
	(1)	(2)	(3)
	DCRC	CRC	CRE
β	5.91***	8.07***	6.78***
	(0.49)	(0.38)	(0.28)
ϕ	-5.01***	-2.56***	
	(1.34)	(0.33)	

Notes: Structural parameters estimated using minimum distance. Standard errors (reported in parentheses) are calculated analytically for optimally weighted minimum distance for which the weight matrix is the inverse of the variance-covariance matrix from the SUR. * p< 0.1 ** p< 0.05 *** p< 0.01. Column 1 reports estimates from the full DCRC model (with heterogeneous returns and imperfect information), column 2 reports estimates from the CRC model (with heterogeneous returns and perfect information), and column 3 reports estimates from the CRE model (with homogeneous returns and perfect information).

4.2 Robustness Checks

Our main conclusions are robust to different definitions of the non-agricultural dummy variable, as we show in Appendix Table A1. In the first column we report again our baseline estimates, which are based on a non-agricultural dummy variable that equals 1 if a household owns a non-agricultural enterprise or earns more than half of its income from non-agricultural wage work. In column 2, non-agricultural households

[20] $H_0 : \beta^{DCRC} = \beta^{CRC}$ vs $H_a : \beta^{DCRC} < \beta^{CRC}$ yields p-value=0.0003. $H_0 : \phi^{DCRC} = \phi^{CRC}$ vs $H_a : \phi^{DCRC} < \phi^{CRC}$ yields p-value=0.04. $H_0 : \beta^{DCRC} = \beta^{CRE}$ vs $H_a : \beta^{DCRC} < \beta^{CRE}$ yields p-value=0.06. $H_0 : \phi^{DCRC} = \phi^{CRE}$ vs $H_a : \phi^{DCRC} < \phi^{CRE} = 0$ yields p-value=0.0001.

include those which own a non-agricultural enterprise or have at least one household member working outside of the agricultural sector. In column 3, we define non-agricultural households as those with a non-agricultural enterprise or more than half of the household working outside the agricultural sector. Across all columns, β is positive and ϕ is less than one.

In the last column of Appendix Table A1, we repeat our analysis using the individual-level dataset used in Hicks et al. (2017), which also relies on the IFLS. We restrict to individuals with non-missing earnings and sector data throughout the first four waves of the panel, use log earnings as our outcome variable, and define our non-agricultural dummy variable to be equal to 1 for individuals whose primary or secondary occupation is non-agricultural.[21] Using this individual-level dataset, we arrive at the same conclusions: the returns to non-agricultural work are positive, and individuals sort across sectors based on comparative advantage.[22]

4.3 Expected Returns

We next examine how sorting and switching behavior is governed by a household's expected returns to participating in the non-agricultural sector. The ability to recover and interpret these patterns is, perhaps, the main strength of our empirical approach. Other approaches to recovering β and even ϕ would not allow for the recovery of each household's expected returns at each decision point, or an analysis of whether these expectations correspond to subsequent choices in ways consistent with the model.[23]

First, we calculate $\beta + \phi m_{it}$ for each household, for periods $t = 1$ to 4. This represents a household's expected return to the non-agricultural sector, based on what they have learned up until the end of period t about their relative productivity η_i. In

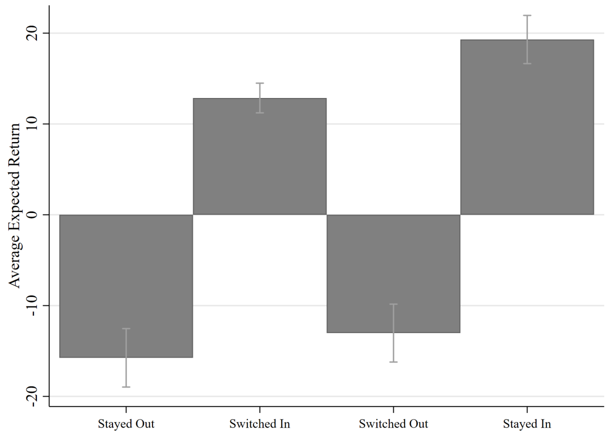
[21] We only use four waves because in the five-wave dataset, there were a few sectoral choice histories that were not experienced by anyone in the dataset (for example, the sequence involving switching in every period), which meant that some coefficients in the SUR could not be estimated.

[22] Note that the magnitudes of our household-level and individual-level β's cannot be compared because the household specifications use income in levels – due to the presence of negative business profits – while the individual specification uses log income, as is done in Hicks et al. (2017).

[23] Though approaches to estimating DCRC models are quite limited in the literature, instrumental variables approaches, for example, used to estimate CRC models (Heckman and Vytlacil, 1998) would not recover these additional parameters. Even to estimate static heterogeneous returns, it would likely be infeasible to find a rich enough set of instruments across such a large set of household types over such a long panel. That is, one would need instruments that predict switching in both directions across households with different relative abilities across different waves just to recover β and ϕ even in the absence of dynamics. For example, price fluctuations alone would not, in general, be enough.

Figure 5, we average these returns for households in four different groups: those who stay out of the non-agricultural sector in the next period, those who switch in to the non-agricultural sector, those who switch out of the non-agricultural sector, and those who stay in the non-agricultural sector. As expected, returns to the non-agricultural sector are higher for households in agriculture who switch into the non-agricultural sector compared to those who stay out. Returns are also higher for non-agricultural households who stay in the non-agricultural sector compared to those who switch out. In terms of magnitudes, the returns for those who switch into the non-agricultural sector is about twice the average return of 5.91 (reported in Table 2). The returns for those who stay in are three times the average. Figure A2 in the appendix calculates these returns by wave, and separately for current non-agricultural households and current agricultural households – both groups show similar patterns, consistent with both the patterns in the raw data and the learning structure assumed in the model.

Figure 5: Expected Returns by Switch Status



Notes: The figure reports the average return to the non-agricultural sector ($\beta + \phi m_{it}$) across $t = 1$ to 4 and all households in each category. “Stayed out” includes households in agriculture in both t and $t + 1$. “Switched In” includes households in agriculture in t and the non-agricultural sector in $t + 1$. “Switched Out” includes households in the non-agricultural sector in t and agriculture in $t + 1$. “Stayed In” includes households in the non-agricultural sector in both t and $t + 1$. Error bars denote 95% confidence intervals. Standard errors are calculated analytically (see Appendix C).

In short, the expected returns estimated by the model are consistent with households’ sorting behavior. Note that though the results are fully consistent with the model intuition, the estimated pattern is not mechanical. The estimation strategy does not restrict in any way these recovered correlations between income evolutions and the sequence of choices. For example, we could have found that only households that stayed in expected large gains while households that switched in expected substantially smaller or negligible gains, suggesting that productivity in the new sector accrues over time as in the case of learning by doing (Foster and Rosenzweig, 1995), an alternative model we discuss in more detail in section 4.5. As such, we interpret the internally consistent pattern of estimates here as a resounding confirmation of the intuition of the model and structure assumed.

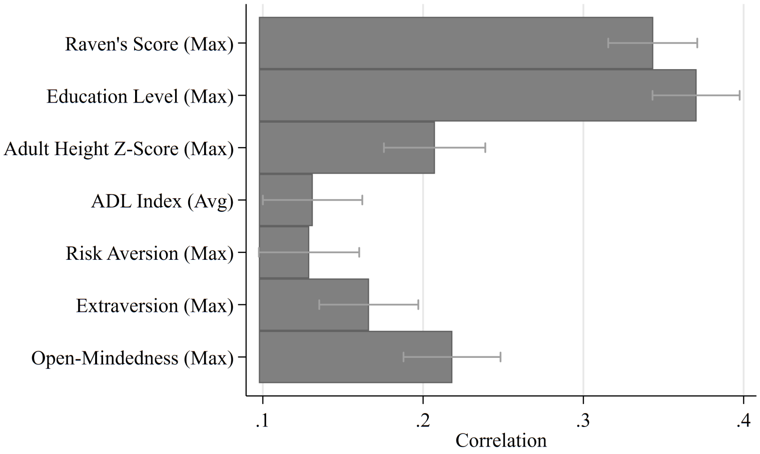
Using these estimated returns, we next explore what types of households tend to have high returns to the non-agricultural sector. To do this, we take each household's final return ($\beta + m_{i4}$) – which is the household's most informed or precise estimate of its return – and calculate its correlation with various household-level characteristics. We take these household characteristics from the 2014 wave of the IFLS because $\beta + m_{i4}$ is a household's perceived return going into this last wave and because this wave includes variables not found in the others (like personality traits). We first use LASSO to select predictors of final returns from a large set of household-level characteristics covering a wide range of areas: cognitive ability, educational attainment, physical health, risk aversion, mental health, and personality traits (see Appendix section D.1 for a description of all variables). Then, for each of the seven variables that were selected, we calculate its correlation with the estimated final return.

These correlations, reported in Figure 6, for the most part have the expected signs. Returns to the non-agricultural sector are positively correlated with cognitive ability (measured by Raven's test scores), educational attainment, height, physical functioning, extraversion, and open-mindedness. Although the correlation between risk aversion and returns is positive, it is the smallest in magnitude; in addition, in a multivariate regression that includes all selected variables, the coefficient on risk aversion is statistically insignificant. In fact, in a multivariate regression that includes all selected variables, only Raven's scores, education, adult height z-scores, and open-mindedness yield statistically significant coefficients.

It is important to note that these variables explain only a small percentage of the variation in returns. In a multivariate regression that includes these seven variables, the adjusted R-squared is 0.13.[24] In other words, returns to the non-agricultural sector are driven primarily by unobservables, which could explain why it is difficult

[24] The adjusted R-squared is roughly the same (and in fact, slightly smaller) for a multivariate regressions with all 27 variables originally included in the LASSO.

Figure 6: Expected Returns and Household Characteristics



Notes: Each bar illustrates the correlation between the listed household level characteristic, taken from the 2014 wave of the IFLS, and the final return to the non-agricultural sector ($\beta + m_{i4}$). Error bars denote 95% confidence intervals. These variables were selected from a larger set of variables (listed in Appendix D.1) using LASSO.

for households to calculate their returns to the non-agricultural sector and therefore why sorting on imperfect information is common, as we discuss below.

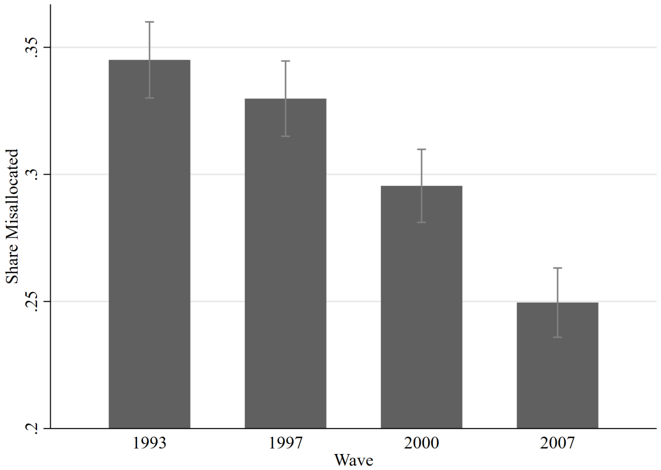
4.4 Sorting on Imperfect Information

Because households switch in and out of the non-agricultural sector as they learn more information about their η_i, many households spend time in a sector which is not the most productive sector for them. To identify households for whom this is the case, we use the household's beliefs about its relative productivity going into the final period (m_{i4}), and calculate its expected return to the non-agricultural sector using this value ($\beta + m_{i4}$). Households with a positive return should be in the non-agricultural sector, while households with a negative return should be in agriculture.[25] Based on this, we determine whether a household is in the most productive sector for them. Figure 7 shows that a large share of households are in their less productive sector in each wave. This share declines from 35% in 1993 to 25% in 2007, indicating that households are learning about their true η_i and becoming increasingly likely to

[25] Note that the underlying incomes and, as a result, these estimated returns are in terms of *net* earnings. As such, any costs of engaging in either activity are already accounted for.

select their most productive sector.[26]

Figure 7: Share of Households in their Less Productive Sector



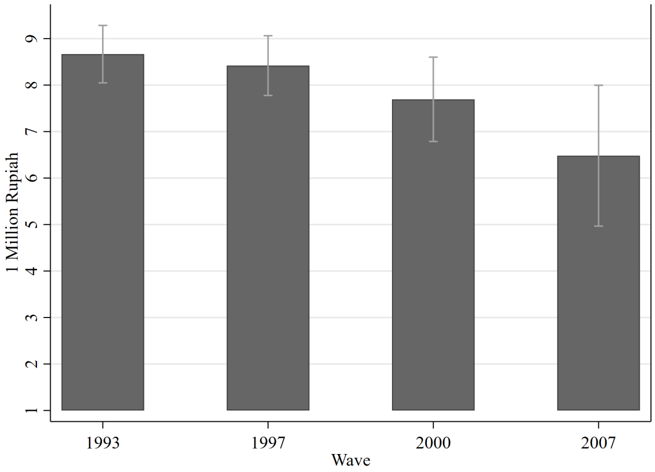
Notes: Households in the less productive sector for them are defined as those with final returns ($\beta + m_{i4}$) greater than zero but in the agricultural sector, or those with final returns less than zero but in the non-agricultural sector. Error bars denote 95% confidence intervals.

We next explore the costs of this inefficient sorting, represented by the absolute value of non-agricultural returns (calculated using final beliefs about η_i, as described above) among households in the less productive sector for them. Households who are currently in agriculture but should be in the non-agricultural sector have a positive return, which represents unrealized income gains due to this sorting decision. Similarly, households who are currently in the non-agricultural sector but should be in agriculture have a negative return, the absolute value of which represents how much more they could have earned if they had chosen the agricultural sector instead. We sum all of these amounts for each wave and divide by the total number of households in the sample. We plot these values in Figure 8. Sorting on imperfect information leads to losses of around 8.6 million rupiah (about 622 USD) per household in 1993. This declines over time, driven both by reductions in the share of households in a less productive sector and the magnitude of their losses. That is, as households converge over time and beliefs become more precise, fewer households are inefficiently sorted and the remaining households have smaller average forgone earnings per household.

[26] Recall that the sample is a balanced panel such that these patterns are not driven by the entry of new households.

We can also express these amounts as a fraction of total potential income (which is equal to a household's realized income plus their return). As we show in Appendix Figure A3, amounts lost due to imperfect information correspond to 79% of these households' potential income overall in 1993. Put differently, households who are inefficiently sorted earn 79% less than they could have had they been in their most productive sector. This figure decreases to around 68% in 2007.

Figure 8: Average Income Lost due to Inefficient Sorting



Notes: A household's lost income is equal to zero if they are in the most productive sector for them, and equal to the absolute value of their estimated final return ($\beta + m_{i4}$) if they are in the less productive sector for them. Standard errors are calculated analytically (see Appendix C).

4.5 Alternative Models

As described above, our empirical strategy can recover consistent estimates of β and ϕ (as long as sequential exogeneity still holds), even if the learning structure outlined above is not the main driver of the switching dynamics we observe in the data. In this section, we discuss some of these alternative models and evaluate whether our evidence is consistent with them.

4.5.1 Land Market Frictions

Frictions in land markets have been proposed as an important potential source of misallocation (Adamopoulos et al., 2017; Adamopoulos and Restuccia, 2020; Chen, 2017), but we argue that they are unlikely to be the primary driver of the inefficient sorting we document here for several reasons. First, the substantial, bilateral, high frequency churning in Figure 1 is inconsistent with the idea that land market frictions are driving the dynamic sorting patterns we attempt to explain in this paper, as such frictions should restrict switching out of and into agriculture substantially.

In addition, households in our sample do not appear to be substantially constrained in their ability to buy and sell land. For example, using IFLS survey questions on land ownership at the household level, we find that around half of households in our sample change land ownership status at least once in the study period (i.e., they go from owning no land to owning land or vice versa).[27] In spite of this, we acknowledge that some sort of land friction could still be a source of misallocation in our context. We explicitly aim to cut past these issues by absorbing community by year fixed effects. The fact that we find inefficient sorting of households even after controlling for these fixed effects suggests that something other than market level frictions must be driving this result.

4.5.2 Saving out of Financial Constraints

Households might save to relax financial constraints or overcome switching costs, and this could be a separate reason why households switch sectors and appear to have evolving (perceptions of) η_i. However, this explanation is at odds with Figure 2, which shows that switching declines with the amount of time spent in a given sector. If households were saving to overcome switching costs, we would expect to see the opposite pattern. In addition, because we absorb community by year fixed effects, our estimates are not picking up the effects of any formal or informal borrowing conditions that vary at the community by wave level (for example, the existence, strength, and/or aggregate resources of informal borrowing networks in a village).

[27] While one may worry that part of this could be due to measurement error, or the inclusion or departure of land-owning household members, we also find that 12% of households who owned land for a farm business at any point during the study period reported either buying or selling that land during this time. The IFLS does not ask about sales or purchases of land owned for a non-farm business after the 1997 wave, and does not ask about sales or purchases of other land owned (not for the purpose of any business) after the 1993 wave, which means we cannot calculate this statistics for the full sample. But if anything, this statistic we are able to obtain substantially underestimates the land transactions in our sample.

4.5.3 Learning by Doing

If households accumulate the skills that are more valuable in a sector while participating in that sector, this would generate evolutions in η_i over time. That is, with $\phi < 0$, η_i would go up with time spent in the agricultural sector and go down with time spent in the non-agricultural sector.[28] As long as the evolution process is a martingale such that sequential exogeneity is still valid, this would not prevent our strategy from obtaining consistent estimates of β and ϕ. However, this learning by doing process would result in a different pattern for the evolution of η_i (and therefore expected returns), and importantly would not imply households are failing to sort into the most productive sector for them.

To determine whether this learning mechanism appears consistent with the data, we examine how expected returns evolve for households from the end of period $t-1$ to the end of period t, separately for agricultural and non-agricultural households. Under a learning by doing model, we would expect returns to the non-agricultural sector to decrease from $t-1$ to t, for those who are in agriculture in period t (because they improve their skills in agriculture during that period). At the same time, we would expect returns to the non-agricultural sector to increase from $t-1$ to t for those in the non-agricultural sector, as they improve their non-agricultural skills.

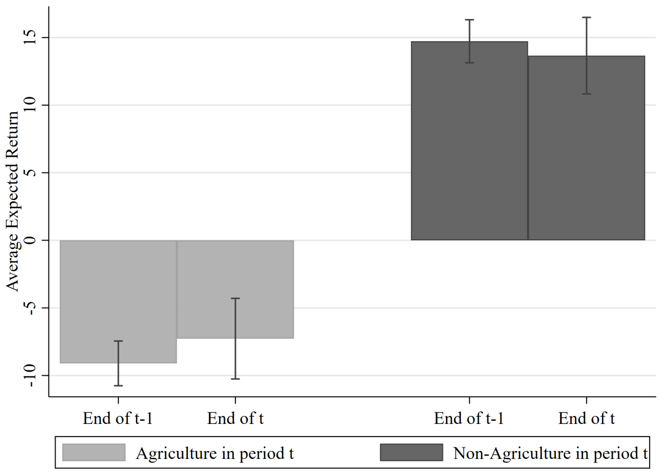
This is not what we find in Figure 9. The first pair of light gray bars shows that expected returns are statistically unchanged from the end of period $t-1$ to the end of period t for those in the agricultural sector in period t. The second pair of dark gray bars shows that expected returns are also unchanged for those in the non-agricultural sector from period period $t-1$ to t.

This flat pattern for each sector is precisely what our proposed learning process would predict. That is, the updates to η_i, unconditional on future decisions, are assumed to have a martingale structure such that further innovations should be mean 0 after switching. As such, this pattern is both inconsistent with a learning by doing interpretation and a strong confirmation of precisely the learning about comparative

[28] An alternative learning structure that could be relevant to our context is the multi-armed bandit problem. That is, households might choose in advance the optimal sector or even sequence of sectoral choices in order to learn about or invest in building their η_i. Under this scenario, households would choose to invest in the sector they believe is most likely to be best for them for several periods and hope to accumulate skill there. Only those who learn they have very low sector-specific skill in their chosen sector or who suffer a very large relative earnings shock would eventually switch, and would be very unlikely to ever switch back. This scenario is completely at odds with the high-frequency bilateral switching in Figure 1.

advantage model we propose.

Figure 9: Evolution of Expected Returns by Sector



Notes: The figure reports the average return to the non-agricultural sector ($\beta + \phi m_{it}$) in $t-1$ and t, separately for households in the agricultural and non-agricultural sector. Because returns can only be estimated for the first four periods and because we also calculate a one period lag, we restrict to the three middle waves (1997, 2000, and 2007). Error bars denote 95% confidence intervals. Standard errors are calculated analytically (see Appendix C).

5 Conclusion

We hypothesize that imperfect information about relative productivity across sectors might lead to inefficient labor sorting. We use a dynamic sectoral sorting framework to study the household's decision to participate in the non-agricultural sector. Previous studies have modeled selection as a one-off sorting decision across sectors, limiting the ability to document inefficient sorting along households' productive life cycles. We document substantial churning along the sectoral margin and show that this churning reduces with experience in a sector.

Using an extension of projection-based panel methods to estimate a generalized earnings equation with dynamic correlated random coefficients, we find many households spend substantial amounts of time in a sector which is less productive for them, earning 79% less on average than they could have if they were properly sorted across sectors. That is, structural estimates confirm that the sectoral churning is, at least in part, due to substantial learning about relative abilities across sectors and slow convergence to a household's most productive sector.

Our approach nests several alternative models which can be ruled out. For example, we can estimate a model with comparative advantage but no dynamics as well as a model with neither dynamics nor heterogeneity in relative earnings across sectors. We find that dynamics are important and in fact that the heterogeneity in relative earnings across sectors is more pronounced when allowing for dynamics. Finally, we also evaluate alternative interpretations for the dynamic heterogeneity we observe in the data. We consider whether land market frictions, saving out of financial constraints, or learning by doing could explain the patterns we observe in the raw data and the structural parameters we recover, and find each of these alternatives to be less consistent with our findings than learning about comparative advantage.

References

Adamopoulos, T., Brandt, L., Leight, J., and Restuccia, D. (2017). Misallocation, selection and productivity: A quantitative analysis with panel data from china. Technical report, National Bureau of Economic Research.

Adamopoulos, T. and Restuccia, D. (2020). Land reform and productivity: A quantitative analysis with micro data. *American Economic Journal: Macroeconomics*, 12(3):1–39.

Adhvaryu, A., Bassi, V., Nyshadham, A., and Tamayo, J. (2019a). No line left behind: Assortative matching inside the firm. *Available at SSRN 3462873*.

Adhvaryu, A., Kala, N., and Nyshadham, A. (2020). Booms, busts, and household enterprise: Evidence from coffee farmers in tanzania. *The World Bank Economic Review*.

Adhvaryu, A. and Nyshadham, A. (2017). Health, enterprise, and labor complementarity in the household. *Journal of development economics*, 126:91–111.

Adhvaryu, A., Nyshadham, A., and Tamayo, J. (2019b). Managerial quality and productivity dynamics. Technical report, Mimeo, University of Michigan, Boston College and Harvard Business School.

Alvarez, J. A. (2020). The agricultural wage gap: Evidence from brazilian micro-data. *American Economic Journal: Macroeconomics*, 12(1):153–73.

Alvarez-Cuadrado, F., Amodio, F., and Porschke, M. (2019). Selection and absolute advantage in farming and entrepreneurship: Microeconomic evidence and macroeconomic implications.

Atkin, D. and Donaldson, D. (2015). Who's getting globalized? the size and implications of intra-national trade costs. Technical report, National Bureau of Economic Research.

Atkin, D., Khandelwal, A. K., and Osman, A. (2017). Exporting and firm performance: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 132(2):551–615.

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013). Does management matter? evidence from india. *The Quarterly Journal of Economics*, 1(51):51.

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2010). Why do firms in developing countries have low productivity? *American Economic Review*, 100(2):619–23.

Bloom, N. and Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, pages 1351–1408.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. Technical report, National Bureau of Economic Research.

Calderon, G., Cunha, J. M., and Giorgi, G. D. (2020). Business literacy and development: Evidence from a randomized controlled trial in rural mexico. *Economic Development and Cultural Change*, 68(2):507–540.

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.

Chen, C. (2017). Untitled land, occupational choice, and agricultural productivity. *American Economic Journal: Macroeconomics*, 9(4):91–121.

Crépon, B. and Mairesse, J. (2008). The chamberlain approach to panel data: an overview and some simulations. In *The Econometrics of Panel Data*, pages 113–183. Springer.

DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw Hill, New York.

Foster, A. D. and Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of political Economy*, 103(6):1176–1209.

Frankenberg, E. and Karoly, L. A. (1995). The 1993 indonesian family life survey: Overview and field report.

Frankenberg, E. and Thomas, D. (2000). The indonesia family life survey (ifls): Study design and results from waves 1 and 2.

Gibbons, R., Katz, L. F., Lemieux, T., and Parent, D. (2005). Comparative advantage, learning, and sectoral wage determination. *Journal of labor economics*, 23(4):681–724.

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

Hall, R. E. and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, 114(1):83–116.

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. *Journal of Human Resources*, pages 974–987.

Herrendorf, B. and Schoellman, T. (2018). Wages, human capital, and barriers to structural transformation. *American Economic Journal: Macroeconomics*, 10(2):1–23.

Hicks, J. H., Kleemans, M., Li, N. Y., and Miguel, E. (2017). Reevaluating agricultural productivity gaps with longitudinal microdata. Technical report, National Bureau of Economic Research.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, 124(4):1403–1448.

Islam, N. (1995). Growth empirics: a panel data approach. *The quarterly journal of economics*, 110(4):1127–1170.

Lagakos, D. and Waugh, M. E. (2013). Selection, agriculture, and cross-country productivity differences. *American Economic Review*, 103(2):948–80.

Lemieux, T. (1998). Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and nonrandom selection. *Journal of Labor Economics*, 16(2):261–291.

Papageorgiou, T. (2014). Learning your comparative advantages. *Review of Economic Studies*, 81(3):1263–1295.

Porzio, T., Rossi, F., and Santangelo, G. (2020). The human side of structural transformation.

Pulido, J., Swiecki, T., et al. (2018). Barriers to mobility or sorting? sources and aggregate implications of income gaps across sectors and locations in indonesia. *Unpublished Working Paper, Vancouver School of Economics.*

Restuccia, D. and Rogerson, R. (2013). Misallocation and productivity. *Review of Economic Dynamics*, 1(16):1–10.

Restuccia, D. and Santaeulalia-Llopis, R. (2017). Land misallocation and productivity. Technical report, National Bureau of Economic Research.

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*, 3(2):135–146.

Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herawati, Y., and Witoelar, F. (2004). The third wave of the indonesia family life survey (ifls3): Overview and field report. *NIA/NICHD*.

Strauss, J., Witoelar, F., and Sikoki, B. (2016). The fifth wave of the indonesia family life survey: Overview and field report: Volume 1.

Strauss, J., Witoelar, F., Sikoki, B., and Wattie, A. M. (2009). The fourth wave of the indonesian family life survey (ifls4): overview and field report. *RAND Corporation*.

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica*, 79(1):159–209.

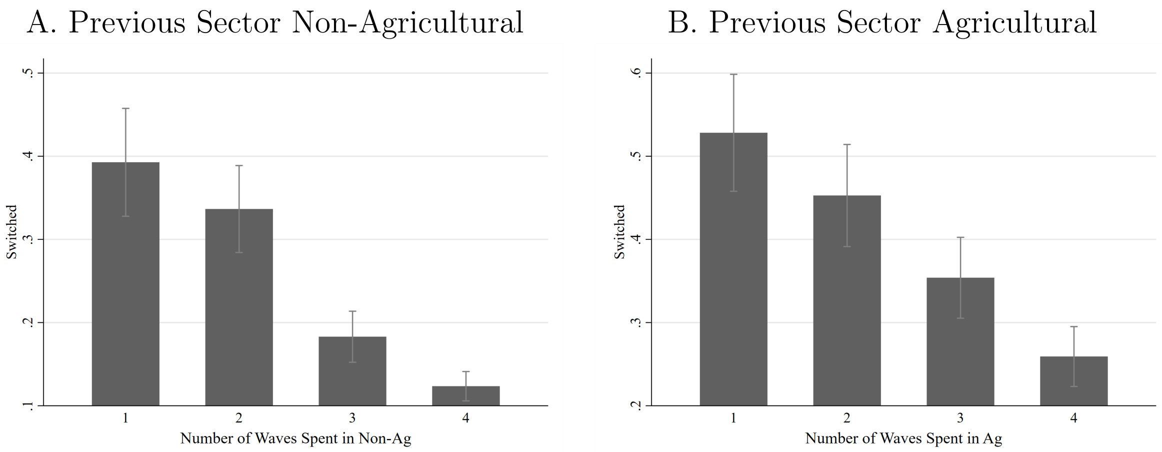
Syverson, C. (2011). What determines productivity? *Journal of Economic Literature*, 49(2):326–365.

Zhang, W., O'Brien, N., Forrest, J. I., Salters, K. A., Patterson, T. L., Montaner, J. S., Hogg, R. S., and Lima, V. D. (2012). Validating a shortened depression scale (10 item ces-d) among hiv-positive people in british columbia, canada. *PloS one*, 7(7):e40793.

Online Appendix

A Appendix Figures

Figure A1: Switching by Number of Waves Spent in Previous Sector: Non-Agricultural and Agricultural



A. Previous Sector Non-Agricultural

B. Previous Sector Agricultural

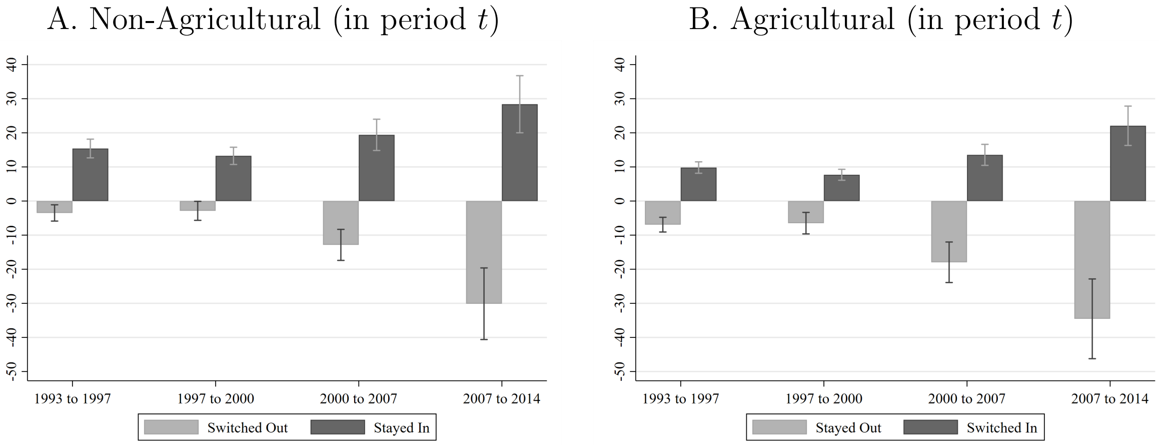
Notes: Sample consists of IFLS households with non-missing income information in all five waves of the IFLS. Graph illustrates switching behavior from the fourth to fifth (and last) wave of the survey. Error bars denote 95% confidence intervals.

Table A1: Structural Estimates (Robustness)

	Specification			
	(1)	(2)	(3)	(4)
	Baseline	Definition2	Definition3	Individual
β	5.91***	4.53***	4.15***	0.98***
	(0.49)	(0.51)	(0.73)	(0.35)
ϕ	-5.01***	-4.81***	-11.97	-1.28**
	(1.34)	(1.86)	(12.38)	(0.19)

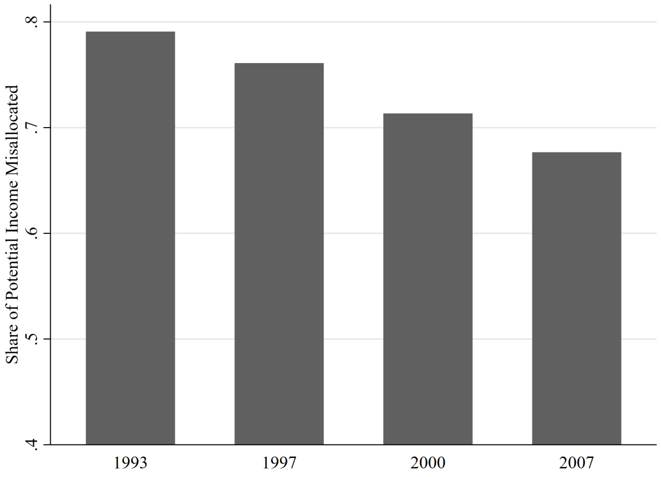
Notes: Structural parameters estimated using minimum distance. Standard errors (reported in parentheses) are calculated analytically for optimally weighted minimum distance for which the weight matrix is the inverse of the variance-covariance matrix from the SUR. * p< 0.1 ** p< 0.05 *** p< 0.01. Column 1 reports estimates from the baseline model in Table 2, which defines non-agricultural households as those who have a non-agricultural enterprise or earn at least half of their income from non-agricultural wage work. Column 2 requires that non-agricultural households own a non-agricultural enterprise or have any wage workers in the non-agricultural sector. Column 3 requires that non-agricultural households have a non-agricultural enterprise or at least half of the household working in the non-agricultural sector. Column 4 uses individual-level IFLS data from Hicks et al. (2017).

Figure A2: Expected Returns by Switch Status, Wave, and Current Sector



Notes: The figure reports the average return to the non-agricultural sector ($\beta + \phi m_{it}$), separately for each transition, across all households in each category. “Stayed out” includes households in agriculture in both t and $t+1$. “Switched In” includes households in agriculture in t and the non-agricultural sector in $t+1$. “Switched Out” includes households in the non-agricultural sector in t and agriculture in $t+1$. “Stayed In” includes households in the non-agricultural sector in both t and $t+1$. Error bars denote 95% confidence intervals. Standard errors are calculated analytically (see Appendix C).

Figure A3: Share of Income Lost due to Inefficient Sorting



Notes: Income lost due to inefficient sorting is defined as the absolute value of final returns $(\beta + m_{i4})$ among households in the less productive sector for them. The share of potential income lost is equal to the sum of all income lost due to imperfect information, divided by the potential income (realized income plus final return) among households who are in the less productive sector for them.

B Additional Equations

B.1 Minimum Distance Restrictions

The minimum distance restrictions are as follows.

$\gamma_1^1 = \beta + \phi\lambda_0 + \lambda_1 + \phi\lambda_1$

$\gamma_2^1 = \lambda_2$

$\gamma_3^1 = \lambda_3$

$\gamma_4^1 = \lambda_4$

$\gamma_5^1 = \lambda_5$

$\gamma_{12}^1 = \phi\lambda_2 + \lambda_{12} + \phi\lambda_{12}$

$\gamma_{13}^1 = \phi\lambda_3 + \lambda_{13} + \phi\lambda_{13}$

$\gamma_{14}^1 = \phi\lambda_4 + \lambda_{14} + \phi\lambda_{14}$

$\gamma_{15}^1 = \phi\lambda_5 + \lambda_{15} + \phi\lambda_{15}$

$\gamma_{23}^1 = \lambda_{23}$

$\gamma_{24}^1 = \lambda_{24}$

$\gamma_{25}^1 = \lambda_{25}$

$\gamma_{34}^1 = \lambda_{34}$

$\gamma_{35}^1 = \lambda_{35}$

$\gamma_{45}^1 = \lambda_{45}$

$\gamma_{123}^1 = \phi\lambda_{23} + \lambda_{123} + \phi\lambda_{123}$

$\gamma_{124}^1 = \phi\lambda_{24} + \lambda_{124} + \phi\lambda_{124}$

$\gamma_{125}^1 = \phi\lambda_{25} + \lambda_{125} + \phi\lambda_{125}$

$\gamma^1_{134} = \phi\lambda_{34} + \lambda_{134} + \phi\lambda_{134}$

$\gamma^1_{135} = \phi\lambda_{35} + \lambda_{135} + \phi\lambda_{135}$

$\gamma^1_{145} = \phi\lambda_{45} + \lambda_{145} + \phi\lambda_{145}$

$\gamma^1_{234} = \lambda_{234}$

$\gamma^1_{235} = \lambda_{235}$

$\gamma^1_{245} = \lambda_{245}$

$\gamma^1_{345} = \lambda_{345}$

$\gamma^1_{1234} = \phi\lambda_{234} + \lambda_{1234} + \phi\lambda_{1234}$

$\gamma^1_{1235} = \phi\lambda_{235} + \lambda_{1235} + \phi\lambda_{1235}$

$\gamma^1_{1245} = \phi\lambda_{245} + \lambda_{1245} + \phi\lambda_{1245}$

$\gamma^1_{1345} = \phi\lambda_{345} + \lambda_{1345} + \phi\lambda_{1345}$

$\gamma^1_{2345} = \lambda_{2345}$

$\gamma^1_{12345} = \phi\lambda_{2345} + \lambda_{12345} + \phi\lambda_{12345}$

$\gamma^2_{1} = \lambda_{1}$

$\gamma^2_{2} = \beta + \phi\theta_{20} + \theta_{22} + \phi\theta_{22} + \phi\lambda_{0} + \lambda_{2} + \phi\lambda_{2}$

$\gamma^2_{3} = \theta_{23} + \lambda_{3}$

$\gamma^2_{4} = \theta_{24} + \lambda_{4}$

$\gamma^2_{5} = \theta_{25} + \lambda_{5}$

$\gamma^2_{12} = \phi\lambda_{1} + \lambda_{12} + \phi\lambda_{12}$

$\gamma^2_{13} = \lambda_{13}$

$\gamma^2_{14} = \lambda_{14}$

$\gamma^2_{15} = \lambda_{15}$

$\gamma^2_{23} = \phi\theta_{23} + \phi\lambda_3 + \lambda_{23} + \phi\lambda_{23}$

$\gamma^2_{24} = \phi\theta_{24} + \phi\lambda_4 + \lambda_{24} + \phi\lambda_{24}$

$\gamma^2_{25} = \phi\theta_{25} + \phi\lambda_5 + \lambda_{25} + \phi\lambda_{25}$

$\gamma^2_{34} = \lambda_{34}$

$\gamma^2_{35} = \lambda_{35}$

$\gamma^2_{45} = \lambda_{45}$

$\gamma^2_{123} = \phi\lambda_{13} + \lambda_{123} + \phi\lambda_{123}$

$\gamma^2_{124} = \phi\lambda_{14} + \lambda_{124} + \phi\lambda_{124}$

$\gamma^2_{125} = \phi\lambda_{15} + \lambda_{125} + \phi\lambda_{125}$

$\gamma^2_{134} = \lambda_{134}$

$\gamma^2_{135} = \lambda_{135}$

$\gamma^2_{145} = \lambda_{145}$

$\gamma^2_{234} = \phi\lambda_{34} + \lambda_{234} + \phi\lambda_{234}$

$\gamma^2_{235} = \phi\lambda_{35} + \lambda_{235} + \phi\lambda_{235}$

$\gamma^2_{245} = \phi\lambda_{45} + \lambda_{245} + \phi\lambda_{245}$

$\gamma^2_{345} = \lambda_{345}$

$\gamma^2_{1234} = \phi\lambda_{134} + \lambda_{1234} + \phi\lambda_{1234}$

$\gamma^2_{1235} = \phi\lambda_{135} + \lambda_{1235} + \phi\lambda_{1235}$

$\gamma^2_{1245} = \phi\lambda_{145} + \lambda_{1245} + \phi\lambda_{1245}$

$\gamma^2_{1345} = \lambda_{1345}$

$\gamma^2_{2345} = \phi\lambda_{345} + \lambda_{2345} + \phi\lambda_{2345}$

$\gamma^2_{12345} = \phi\lambda_{1345} + \lambda_{12345} + \phi\lambda_{12345}$

$\gamma^3_1 = \lambda_1$

$\gamma^3_2 = \lambda_2$

$\gamma^3_3 = \beta + \phi\theta_{30} + \theta_{33} + \phi\theta_{33} + \phi\lambda_0 + \lambda_3 + \phi\lambda_3$

$\gamma^3_4 = \theta_{34} + \lambda_4$

$\gamma^3_5 = \theta_{35} + \lambda_5$

$\gamma^3_{12} = \lambda_{12}$

$\gamma^3_{13} = \phi\lambda_1 + \lambda_{13} + \phi\lambda_{13}$

$\gamma^3_{14} = \lambda_{14}$

$\gamma^3_{15} = \lambda_{15}$

$\gamma^3_{23} = \phi\lambda_2 + \lambda_{23} + \phi\lambda_{23}$

$\gamma^3_{24} = \lambda_{24}$

$\gamma^3_{25} = \lambda_{25}$

$\gamma^3_{34} = \phi\theta_{34} + \phi\lambda_4 + \lambda_{34} + \phi\lambda_{34}$

$\gamma^3_{35} = \phi\theta_{35} + \phi\lambda_5 + \lambda_{35} + \phi\lambda_{35}$

$\gamma^3_{45} = \lambda_{45}$

$\gamma^3_{123} = \phi\lambda_{12} + \lambda_{123} + \phi\lambda_{123}$

$\gamma^3_{124} = \lambda_{124}$

$\gamma^3_{125} = \lambda_{125}$

$\gamma^3_{134} = \phi\lambda_{14} + \lambda_{134} + \phi\lambda_{134}$

$\gamma^3_{135} = \phi\lambda_{15} + \lambda_{135} + \phi\lambda_{135}$

$\gamma^3_{145} = \lambda_{145}$

$\gamma^3_{234} = \phi\lambda_{24} + \lambda_{234} + \phi\lambda_{234}$

$\gamma^3_{235} = \phi\lambda_{25} + \lambda_{235} + \phi\lambda_{235}$

$\gamma^3_{245} = \lambda_{245}$

$\gamma^3_{345} = \phi\lambda_{45} + \lambda_{345} + \phi\lambda_{345}$

$\gamma^3_{1234} = \phi\lambda_{124} + \lambda_{1234} + \phi\lambda_{1234}$

$\gamma^3_{1235} = \phi\lambda_{125} + \lambda_{1235} + \phi\lambda_{1235}$

$\gamma^3_{1245} = \lambda_{1245}$

$\gamma^3_{1345} = \phi\lambda_{145} + \lambda_{1345} + \phi\lambda_{1345}$

$\gamma^3_{2345} = \phi\lambda_{245} + \lambda_{2345} + \phi\lambda_{2345}$

$\gamma^3_{12345} = \phi\lambda_{1245} + \lambda_{12345} + \phi\lambda_{12345}$

$\gamma^4_1 = \lambda_1$

$\gamma^4_2 = \lambda_2$

$\gamma^4_3 = \lambda_3$

$\gamma^4_4 = \beta + \phi\theta_{40} + \theta_{44} + \phi\theta_{44} + \phi\lambda_0 + \lambda_4 + \phi\lambda_4$

$\gamma^4_5 = \theta_{45} + \lambda_5$

$\gamma^4_{12} = \lambda_{12}$

$\gamma^4_{13} = \lambda_{13}$

$\gamma^4_{14} = \phi\lambda_1 + \lambda_{14} + \phi\lambda_{14}$

$\gamma^4_{15} = \lambda_{15}$

$\gamma^4_{23} = \lambda_{23}$

$\gamma^4_{24} = \phi\lambda_2 + \lambda_{24} + \phi\lambda_{24}$

$\gamma^4_{25} = \lambda_{25}$

$\gamma^4_{34} = \phi\lambda_3 + \lambda_{34} + \phi\lambda_{34}$

$\gamma^4_{35} = \lambda_{35}$

$\gamma^4_{45} = \phi\theta_{45} + \phi\lambda_5 + \lambda_{45} + \phi\lambda_{45}$

$\gamma^4_{123} = \lambda_{123}$

$\gamma^4_{124} = \phi\lambda_{12} + \lambda_{124} + \phi\lambda_{124}$

$\gamma^4_{125} = \lambda_{125}$

$\gamma^4_{134} = \phi\lambda_{13} + \lambda_{134} + \phi\lambda_{134}$

$\gamma^4_{135} = \lambda_{135}$

$\gamma^4_{145} = \phi\lambda_{15} + \lambda_{145} + \phi\lambda_{145}$

$\gamma^4_{234} = \phi\lambda_{23} + \lambda_{234} + \phi\lambda_{234}$

$\gamma^4_{235} = \lambda_{235}$

$\gamma^4_{245} = \phi\lambda_{25} + \lambda_{245} + \phi\lambda_{245}$

$\gamma^4_{345} = \phi\lambda_{35} + \lambda_{345} + \phi\lambda_{345}$

$\gamma^4_{1234} = \phi\lambda_{123} + \lambda_{1234} + \phi\lambda_{1234}$

$\gamma^4_{1235} = \lambda_{1235}$

$\gamma^4_{1245} = \phi\lambda_{125} + \lambda_{1245} + \phi\lambda_{1245}$

$\gamma^4_{1345} = \phi\lambda_{135} + \lambda_{1345} + \phi\lambda_{1345}$

$\gamma^4_{2345} = \phi\lambda_{235} + \lambda_{2345} + \phi\lambda_{2345}$

$\gamma^4_{12345} = \phi\lambda_{2345} + \lambda_{12345} + \phi\lambda_{12345}$

$\gamma^5_1 = \lambda_1$

$\gamma^5_2 = \lambda_2$

$\gamma^5_3 = \lambda_3$

$\gamma^5_4 = \lambda_4$

$\gamma^5_5 = \beta + \phi\theta_{50} + \theta_{55} + \phi\theta_{55} + \phi\lambda_0 + \lambda_5 + \phi\lambda_5$

$\gamma^5_{12} = \lambda_{12}$

$\gamma^5_{13} = \lambda_{13}$

$\gamma^5_{14} = \lambda_{14}$

$\gamma^5_{15} = \phi\lambda_1 + \lambda_{15} + \phi\lambda_{15}$

$\gamma^5_{23} = \lambda_{23}$

$\gamma^5_{24} = \lambda_{24}$

$\gamma^5_{25} = \phi\lambda_2 + \lambda_{25} + \phi\lambda_{25}$

$\gamma^5_{34} = \lambda_{34}$

$\gamma^5_{35} = \phi\lambda_3 + \lambda_{35} + \phi\lambda_{35}$

$\gamma^5_{45} = \phi\lambda_4 + \lambda_{45} + \phi\lambda_{45}$

$\gamma^5_{123} = \lambda_{123}$

$\gamma^5_{124} = \lambda_{124}$

$\gamma^5_{125} = \phi\lambda_{12} + \lambda_{125} + \phi\lambda_{125}$

$\gamma^5_{134} = \lambda_{134}$

$\gamma^5_{135} = \phi\lambda_{13} + \lambda_{135} + \phi\lambda_{135}$

$\gamma^5_{145} = \phi\lambda_{14} + \lambda_{145} + \phi\lambda_{145}$

$\gamma^5_{234} = \lambda_{234}$

$\gamma^5_{235} = \phi\lambda_{23} + \lambda_{235} + \phi\lambda_{235}$

$\gamma^5_{245} = \phi\lambda_{24} + \lambda_{245} + \phi\lambda_{245}$

$\gamma^5_{345} = \phi\lambda_{34} + \lambda_{345} + \phi\lambda_{345}$

$\gamma^5_{1234} = \lambda_{1234}$

$\gamma^5_{1235} = \phi\lambda_{123} + \lambda_{1235} + \phi\lambda_{1235}$

$\gamma^5_{1245} = \phi\lambda_{124} + \lambda_{1245} + \phi\lambda_{1245}$

$\gamma^5_{1345} = \phi\lambda_{134} + \lambda_{1345} + \phi\lambda_{1345}$

$\gamma^5_{2345} = \phi\lambda_{234} + \lambda_{2345} + \phi\lambda_{2345}$

$\gamma^5_{12345} = \phi\lambda_{1234} + \lambda_{12345} + \phi\lambda_{12345}$

C Standard Errors

In Figures 5 and A2, we report error bars for average returns ($\beta + \phi m_{it}$) across various combinations of household types and waves. In this section, we describe how we obtain the required standard errors.

We denote estimated average returns for a particular group of households in a particular wave as $\hat{f}$. To estimate $\hat{f}$, we use estimates of the parameters β, ϕ, and some combination of the λ and θ parameters that are required to estimate m_{it}. In short, $\hat{f}$ is a non-linear function of estimated parameters and household decisions D_{it}. We define

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{A} h(X_i, \hat{\rho}),$$

where $\hat{\rho}$ represents a vector of the estimated structural parameters, X_i is vector of household i's sectoral decisions, and $h(.)$ is a continuous and differentiable function. We can define $\tilde{f}$ as the sample average return calculated using the true parameter vector (ρ_0):

$$\tilde{f} = \frac{1}{N}\sum_{i=1}^{A} h(X_i, \rho_0),$$

and the population average return as

$$f = E[h(X, \rho_0)],$$

where the expectation is over the joint distribution of X.

If we decompose the difference between the estimated $\hat{f}$ and the population parameter f into two parts:

$$(\hat{f} - f) = (\hat{f} - \tilde{f}) + (\tilde{f} - f),$$

then it can be shown that the variance of $(\hat{f} - f)$ is the sum of two terms: the variance of $(\tilde{f} - f)$ and $(\hat{f} - \tilde{f})$. Specifically,

$$\text{Var}(\hat{f} - f) = \frac{\sigma^2}{N} + \frac{s^2}{N},$$

where (using the delta method)

$$\frac{\sigma^2}{N} = \frac{1}{N} E\left[\nabla h(\rho_0)\right]' V E\left[\nabla h(\rho_0)\right]$$

and

$$\frac{s^2}{N} = \frac{1}{N} \mathrm{Var}(h(X, \rho_0)).$$

D Data Appendix

D.1 Selecting Household Characteristics

As described in section 4.3, we use LASSO to select a set of household-level predictors of returns to the non-agricultural sector from a wide range of variables. Below, we describe all 27 variables included in the LASSO.

- Years of educational attainment (average and maximum): We calculate both average and maximum educational attainment across all household members.

- Raven's test z-score (average and maximum): The IFLS administered a test of cognitive ability (which included questions from the Raven's test of fluid intelligence as well as a few math questions). Different versions of the test were given to respondents aged 7-14 and 15-59. We calculate the version-specific z-score for each respondent and average across all household members. We also calculate the maximum.

- Risk aversion score (average and maximum): This is a five-point score generated from a set of five questions asked of those aged 15 and older, where a score of 5 represents the highest level of risk aversion. Each question offers two hypothetical options: receiving 4 million rupiah for certain, or a lottery with a higher expected value. We calculate the average and maximum score across all household members.

- Height (average and maximum): The IFLS measures height for all household members. Restricting to adults aged 20-65, we standardize height separately for men and women. We calculate the average and maximum z-score across all adults.

- Self-reported health (average and maximum): All respondents aged 15 and older are asked whether they consider themselves very healthy, somewhat healthy, somewhat unhealthy, or unhealthy. We assign a 4 to very healthy and 1 to unhealthy, and calculate both the average and maximum.

- Share of very healthy adults: We calculate the share of household members aged 15 and older who consider themselves very healthy.

- Share of somewhat healthy adults: We calculate the share of household members aged 15 and older who consider themselves very healthy or somewhat healthy.

- Physical functioning (average and maximum): The IFLS asks all respondents aged 15 and older whether they can easily, can with difficulty, or cannot at all do 23 physical activity tasks (including activities of daily living, instrumental activities of daily living, and other physical tasks). We calculate the share of activities a respondent "can easily" do. We then calculate the average share and maximum share for each household.

- Mental health score (average and maximum): To measure mental health (for respondents aged 15 and older), the IFLS includes a 10-question version of the CES-D questionnaire designed to help identify clinical depression. We sum the responses to all 10 questions, which generates a score ranging from 0 to 30 points, where higher numbers are associated with a higher severity of depressive symptoms. We calculate the average and maximum score for each household.

- Share of members with depressive symptoms: Using the 10-question CES-D questionnaire described above, we calculate the share of (adult) household members with a score of 10 or greater, a cutoff that is used as an indicator of significant depressive symptoms (Zhang et al., 2012).

- Big 5 personality traits (open-mindedness, conscientiousness, extraversion, agreeableness, negative emotionality – average and maximum): The IFLS includes the Big Five Index 15 (BFI 15), a set of 15 questions about the respondents' personality, three for each of the five personality traits. We use these to create a five-point score for each of the five personality traits. We calculate the average and maximum score for each household, for each of the five personality traits.