Selection and Absolute Advantage in Farming and Entrepreneurship*

Francisco Alvarez-Cuadrado† Francesco Amodio‡ Markus Poschke§

October 29, 2021

Abstract

Output per worker is lower in agriculture than in other sectors, and relatively more so in poor countries. Sorting of workers can explain this fact if comparative and absolute advantage in agriculture are positively correlated. We investigate this correlation using representative household-level panel data from four African countries. We exploit information on households who engage in both agriculture and non-farm entrepreneurship – about one third of the population. More productive farming households are more likely to pursue entrepreneurship, allocate more hours to it, and are more likely to enter over time. This suggests that agricultural comparative and absolute advantage are negatively correlated.

Keywords: agricultural productivity gap, selection, entrepreneurship, Africa.

JEL Codes: J24, J31, J43, L26, O11, O13, O40.

*Yifan Li, Lesley Johnson and Masaya Takano provided excellent research assistance. We are thankful to Mark Bils, Paco Buera, Lorenzo Casaburi, Rui Castro, Julieta Caunedo, Stefan Dercon, Kevin Donovan, Jean-Marie Dufour, Raphael Godefroy, Doug Gollin, Clément Imbert, Fabian Lange, Sonia Laszlo, Todd Schoellman, Mike Waugh, and seminar participants at Carleton, CEMFI, Laval, McGill, National University of Singapore, Oxford, Universidad Autónoma de Madrid, Warwick, York, the 2019 Canadian Macroeconomics Study Group Meeting, the 2019 Firms, Markets and Development Workshop at Ortygia Business School, SAEe 2019, the 2020 STEG Workshop, the 2020 WUSTL Macro Development Mini-Conference and the 2021 Meeting of the Society of Economic Dynamics. Errors remain our own. We gratefully acknowledge the International Growth Centre for financial support.

†francisco.alvarez-cuadrado@mcgill.ca, McGill University and CIREQ, Leacock Building, 855 Sherbrooke St. West, Montreal QC H3A 2T7, Canada. Declarations of interest: none.

‡francesco.amodio@mcgill.ca, McGill University and CIREQ, Leacock Building, 855 Sherbrooke St. West, Montreal QC H3A 2T7, Canada. Declarations of interest: none.

§markus.poschke@mcgill.ca , McGill University and CIREQ, Leacock Building, 855 Sherbrooke St. West, Montreal QC H3A 2T7, Canada. Declarations of interest: none.

1 Introduction

It is well known that productivity is lower in agriculture than in other sectors in almost all countries. Poor countries have a particularly large *agricultural productivity gap*.[1] Only part of this gap can be accounted for by observables.[2] Since agriculture accounts for the majority of employment in the poorest countries, this gap has important implications for aggregate differences in output per capita across countries.

A recent influential literature argues that an important source of the agricultural productivity gap is worker *self-selection* or *sorting* according to comparative advantage (Lagakos and Waugh 2013). The intuition is simple: if the distributions of abilities in the population are similar across countries, and the best potential farmers choose to engage in farming, then only the very best farmers are active in countries with few farmers. In countries with more farmers, those same top farmers are still active, but they are accompanied by a group of less productive farmers. It follows that the average ability of active farmers is lower in countries with more farmers, reducing productivity of the sector. In the language of the literature, selection widens the agricultural productivity gap if comparative advantage – which determines individuals' sectoral choice – and absolute advantage – their ability or productivity in a sector – are positively correlated or "aligned": those who choose farming are also the best farmers overall.

Providing evidence on this hypothesis is challenging because selection itself shapes what is observable in the data (Heckman and Honoré 1990). In a typical cross-section, a farmer's non-agricultural productivity is unknown. The same is true for the farming ability of non-agricultural workers. Starting with the seminal paper by Roy (1951), the literature therefore imposes strict distributional assumptions or relies on the information revealed by those switching sector, since they can be observed in both activities. A disadvantage of the latter approach is that it only focuses on those at the margin between activities and is therefore not informative of the alignment of comparative and absolute advantage in the population.

We take a new, more direct approach to investigating the alignment of comparative and absolute advantage using household-level panel data from four African countries: Ethiopia, Malawi, Nigeria, and Uganda. The data we use come from the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) project, which provides uniquely rich data on agricultural production and non-farm entrepreneurship (de Magalhaes and Santaeulalia-Llopis 2018). The four countries are all poor, have low agricultural productivity, and large shares of employment in agriculture. At the same time, rural households in these countries engage in non-farm entrepreneurship at high rates – between 27% in Malawi and 51% in Nigeria – and occasionally switch activities. Moreover, around a third of households is active in *both* sectors. We can thus make several useful comparisons across households between and within groups, as well as over time.

Our empirical approach is motivated by an extension of the Roy model in which we allow

[1] See e.g. Gollin et al. (2002), Caselli (2005), and Restuccia et al. (2008).
[2] See Young (2013), Gollin et al. (2014), and Herrendorf and Schoellman (2015; 2018).

households to either be active in a single sector or to split their time between the two sectors. The model predicts that households with a strong comparative advantage in a sector will choose to only pursue that activity, while those with a weaker comparative advantage will pursue both. In the data, we then compare the agricultural productivity of households only engaged in farming to those also engaged in non-farm entrepreneurship. This reveals the correlation of comparative advantage in agriculture – which is weaker for those engaged in both activities – with absolute advantage in agriculture – agricultural productivity. Crucially, we can measure the latter in the data for both groups.

We find that, among those households in a village who produce some agricultural output, it is the more productive ones, i.e. those with high absolute advantage, who also engage in entrepreneurship, revealing that their comparative advantage in agriculture is weak. This suggests that comparative and absolute advantage are negatively correlated, or "misaligned", in agriculture. This is true in three out of the four African countries we analyze and suggests that self-selection may not be central in explaining productivity differences in agriculture across countries.

Where does the misalignment between comparative and absolute advantages in agriculture come from? It is a core prediction of the classical Roy model that higher ability households will tend to choose the activity with higher return dispersion (Roy 1951; Young 2014). In our setting, this implies that if household productive abilities in agriculture and non-agriculture are strongly positively correlated, and returns to non-agriculture are more dispersed, then the best farmers can on average reap higher returns outside agriculture, and therefore tend to specialize there. Intermediate ability farmers can still reap relatively high returns outside agriculture, and therefore pursue both activities. The lowest ability farmers, in contrast, tend to face very low returns outside agriculture, and therefore only pursue agriculture. This is consistent with the activity choice patterns we observe for farmers.

Our interpretation of households' activity choices along the extensive margin as reflecting selection based on comparative advantage could be confounded in the presence of frictions like entry barriers or fixed operating costs. The fact that wealthier farming households are more likely to also engage in entrepreneurship provides suggestive evidence that such barriers are present. To address this issue, we analyze the activity choices of households along the intensive margin, which are unaffected by such barriers. Among households pursuing both activities, those with higher productivity in agriculture work fewer hours in agriculture relative to non-agriculture, revealing weak comparative advantage in agriculture. This provides further evidence that comparative and absolute advantages are misaligned in agriculture. Households with higher productivity in non-agriculture instead work more hours in this sector relative to agriculture, suggesting that comparative and absolute advantages are aligned in non-agriculture. These findings are consistent with the scenario with strongly correlated abilities: when being a good farmer is associated with even higher returns outside farming, better farmers spend less time farming.

We also exploit the panel dimension of the data and look at patterns of sectoral choice over time. We find that, over time, among households that in the first wave of data are only active

in agriculture, it is the more productive ones who are most likely to start a non-agricultural enterprise in subsequent waves. This is consistent with our interpretation of the cross-sectional evidence. It also provides evidence that individuals and households sort into sectors in similar ways.

Our model implies that explaining the observed patterns of sectoral choices along the extensive and intensive margin requires a strong positive correlation of abilities in the two sectors, combined with barriers to entering non-farm entrepreneurship.

Finally, we consider a plethora of alternative mechanisms other than selection on ability, including the presence of frictions that distort the activity of households along the intensive margin. We discuss the extent to which these are consistent with the empirical findings. In general, we find little support for these alternative explanations.

To summarize, the fact that a large fraction of households in rural Africa engages in both agricultural and non-agricultural work allows us to sign the correlation of comparative and absolute advantages in agriculture and non-agriculture, as well as the correlation of absolute advantages. The fact that the best farmers are more likely to also engage in non-agriculture suggests a negative correlation, or misalignment, of comparative and absolute advantage in agriculture. The same conclusion can be drawn from the fact that, among those engaged in both activities, the more productive farmers spend fewer hours farming, and more hours in non-agricultural activities; and that, over time, the most productive farming households are systematically more likely to start a non-agricultural enterprise. Taken together, results from both cross-sectional and panel data analysis indicate that a strong positive correlation of productive abilities in the two sectors is responsible for the misalignment of advantages in agriculture.

Our paper is not the first one to analyze the correlation of comparative and absolute advantage or the correlation of productive abilities across sectors. To identify these correlations, earlier work exploits sector-level evidence combined with strict distributional assumptions or information from sector switchers. Lagakos and Waugh (2013) impose dependent Frechet distributions of abilities in the two sectors and calibrate them using average wages across sectors in the US. Their findings imply a positive correlation of advantages in both sectors. Adamopoulos, Brandt, Leight, and Restuccia (2021) calibrate the same joint distribution of abilities using information from sector switchers in Chinese panel data. Through the lens of their model, the observed weak correlation between the agricultural and non-agricultural incomes of switchers implies a positive correlation of advantages. Using data from Brazil, Alvarez (2019) shows that formal workers who transition out of agriculture experience limited compensation gains when compared to the large overall gap in mean wages between agriculture and other sectors. Hicks, Kleemans, Li, and Miguel (2017) use individual-level panel data from Indonesia and Kenya to estimate wage gains from sector switches, conditional on individual fixed effects. They find that wage gains for switchers from agriculture to non-agriculture are much smaller than average earnings differences between the two activities. Using different information from the same data set, Pulido and Święcki (2018) find income gains of over 20% for workers who move out of agriculture. They also conduct a structural estimation exercise that suggests that, while self-selection is

important, there are also barriers that significantly misallocate workers across sectors. These results are obtained under the identifying assumption of uncorrelated shocks to households' sectoral productivity. What these findings highlight is that, unless combined with information on infra-marginal individuals, selection itself makes observational returns to switching sectors or rural-to-urban migration uninformative about the correlation of advantages, the role of sorting, and the scope for worker reallocation (see also Herrendorf and Schoellman 2018; Lagakos, Mobarak, and Waugh 2018; Lagakos, Marshall, Mobarak, Vernot, and Waugh 2020).

Our approach differs from the previous literature in that it requires weaker distributional assumptions and exploits the presence of a large group of households that are simultaneously active in both sectors to sign the correlation of advantages. In several specifications, we consistently find misaligned advantages in agriculture. This is inconsistent with self-selection on unobserved ability as a major determinant of large agricultural productivity gaps in poor countries. The theoretical restrictions that our estimates place on the joint distribution of abilities also provide valuable information about the sign and size of the correlation coefficient of sectoral abilities and their relative dispersion.[3]

The remainder of the paper is organized as follows. Section 2 sketches a simple, general theory that motivates our analysis. Section 3 presents data sources and summary statistics. Section 4 contains the main results on patterns of selection along the extensive margin. Section 5 discusses the relationship between the observed correlation of advantages and the underlying joint distribution of abilities. It also illustrates the role of entry costs as confounders, motivating Section 6, which shows the empirical results on selection along the intensive margin. Section 7 discusses the relationship between household-level and individual-level results while providing additional results from changes in sectoral choice over time. Section 8 discusses possible alternative explanations for the empirical patterns that we document. Section 9 concludes.

2 A Simple Model of Selection

This section describes a simple, general model that motivates the empirical analysis that follows. Consider an economy with two sectors, agriculture and non-farming entrepreneurship, denoted by a and n respectively. There is a mass 1 of households indexed by i. These households are heterogeneous in terms of their abilities in the two sectors. In particular, each household is endowed with a vector of sector-specific abilities $\{z_i^a, z_i^n\}$. These abilities are drawn from a joint distribution $G(z^a, z^n)$ with support on the positive reals and finite mean μ_j and variance σ_j^2, where $j = \{a, n\}$. We define a household's comparative advantage in agriculture as the ratio of agricultural to non-agricultural abilities, z_i^a/z_i^n, while absolute advantage in agriculture is given by agricultural ability z_i^a. Similarly, entrepreneurial comparative and absolute advantage are given by z_i^n/z_i^a and z_i^n, respectively.[4]

[3] In this respect, our focus on the distribution of advantages rather than abilities is similar to Adão (2016), who assumes constant-elasticity schedules for comparative and absolute advantages to fully characterize the distribution of advantages.

[4] We conduct our analysis at the household level due to data availability. In Section 7, we present results indicating

We restrict the ability distributions to be such that $E\left(z^a \middle| z^a/z^n > x\right)$ and $E\left(z^n \middle| z^n/z^a > x\right)$ are monotone in x. This ensures that the correlation between comparative and absolute advantage in each sector maintains the same sign across the whole ability distribution. For instance, if $E\left(z^a \middle| z^a/z^n > x\right)$ is always increasing in x, stronger agricultural comparative advantage is associated, on average, with higher agricultural absolute advantage. As a result, the correlation of advantages in agriculture is positive. In contrast, if the conditional expectation monotonically decreases in x, the correlation of advantages is negative in agriculture. Notice that, differently from Lagakos and Waugh (2013), we only require monotonicity. We do not impose restrictions on whether these two objects increase or decrease with x, since this is determined by the correlation of advantages in each sector, which is what we are ultimately interested in.[5]

Each household i is endowed with one unit of time that it allocates between agriculture l_i^a and non-agricultural entrepreneurship $l_i^n = 1 - l_i^a$. The value added of household i in each sector, y_i^a and y_i^n, is produced combining hours of work with sector-specific abilities as given by

$$\begin{aligned} y_i^a &= \kappa z_i^a f\left(l_i^a\right) \\ y_i^n &= z_i^n g\left(l_i^n\right) = z_i^n g\left(1 - l_i^a\right) \end{aligned} \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ are increasing and strictly concave functions with bounded derivatives at the origin, and κ captures sectoral productivity differences and, in particular, the relative price of the agricultural good. It follows that agricultural value added y_i^a is expressed in units of non-agricultural value added y_i^n, which is the numéraire. Households take the relative price as given and allocate labor to maximize income

$$y_i = \kappa z_i^a f\left(l_i^a\right) + z_i^n g\left(1 - l_i^a\right). \tag{2}$$

In order to fix ideas, we start by focusing on the standard case considered in the selection literature (Roy 1951) in which households operate only in one of the two sectors, i.e. $l_i^j = \{0, 1\}$. The i-th household compares the payoffs of operating in each sector and decides accordingly. This household will be active in farming if and only if

$$\kappa z_i^a f(1) \geqslant z_i^n g(1) \tag{3}$$

As a result, sectoral choices are fully determined by comparative advantage: households with a strong comparative advantage in agriculture, i.e. $z_i^a/z_i^n \geqslant g(1)/(\kappa f(1))$, will engage in farming, while those with a strong entrepreneurial comparative advantage, i.e. $z_i^n/z_i^a > \kappa f(1)/g(1)$, will operate in the non-farm entrepreneurship sector. Combining equation 3 with 1 and the joint

that selection patterns at the individual level mirror those we find at the household level.

[5] Young (2014) and Adão (2016) place similar, although more restrictive, conditions on the distribution of abilities. The former requires sectoral abilities to be independent and the elasticity of the cumulative distribution function for each of the abilities to be decreasing in the level of the ability draw. The later imposes constant elasticity schedules for both comparative and absolute advantage.

density function $g(z^a, z^n)$, we derive mean sectoral labor productivity

$$\begin{aligned}
\mathbb{E}\left(y_i^a | z_i^a/z_i^n \geq g(1)/(\kappa f(1))\right) &= \frac{\kappa f(1) \int_{z_i^a/z_i^n \geqslant g(1)/(\kappa f(1))} z_i^a dGi}{\int_{z_i^a/z_i^n \geqslant g(1)/(\kappa f(1))} dGi} \\
\mathbb{E}\left(y_i^n | z_i^a/z_i^n < g(1)/(\kappa f(1))\right) &= \frac{g(1) \int_{z_i^a/z_i^n < g(1)/(\kappa f(1))} z_i^n dGi}{\int_{z_i^a/z_i^n < g(1)/(\kappa f(1))} dGi}.
\end{aligned} \tag{4}$$

These expressions increase in absolute advantage.

Although comparative advantage determines sectoral allocations, absolute advantage determines sectoral productivities. It follows that the relation between sectoral employment shares and labor productivities is determined by the correlation between comparative and absolute advantage in each sector. To understand this, consider first a situation where comparative and absolute advantage are positively correlated – aligned – in both sectors. In this case, an increase in the threshold of comparative advantage required to operate in a sector leads to an increase in the absolute advantage of those that remain active in the sector. It follows that average productivity increases as a sector shrinks, since the least productive leave the sector. The converse is true in expanding sectors: incoming workers have not only lower comparative advantage, but also, on average, lower absolute advantage than those already in the sector. As a result, average productivity declines in expanding sectors. This is the intuition developed by Lagakos and Waugh (2013) to rationalize the larger agricultural productivity gap in poor countries and by Young (2014) to understand the lower measured growth in labor productivity in the expanding service sector.

Figure 1 illustrates this reasoning. In each panel, the left figure shows a scatter plot of the abilities z_i^a and z_i^n in a simulated population of households. In panel (a), these are generated such that comparative and absolute advantage are aligned in both sectors. This is the case just discussed. In panel (b), they are generated such that advantages are aligned in entrepreneurship, but misaligned in agriculture.[6] In each of these figures, the lines emanating from the origin are lines of constant comparative advantage, indicating the threshold that determines selection across sectors. To show the impact of changes in such a threshold, we draw two lines: $z_i^a/z_i^n = g(1)/(\kappa_t f(1))$, $t = 0, 1$, with $\kappa_1 < \kappa_0$. This illustrates that fewer households find farming optimal when the comparative advantage threshold to engage in farming is higher (with κ_1).

The central figure in each panel shows a scatter plot of households' comparative advantage in agriculture, z_i^a/z_i^n, against their absolute advantage in agriculture, z_i^a. Naturally, the lines of constant comparative advantage determining selection are horizontal. Finally, the right figure in each panel shows a scatter plot of households' comparative advantage in non-farming entrepreneurship, z_i^n/z_i^a, against their absolute advantage in the same sector, z_i^n. Again, the lines of constant comparative advantage are horizontal.

Panel (a) illustrates the first case discussed above. As the agricultural sector shrinks and some households switch to non-agriculture, the average absolute advantage of those who remain in

[6] We discuss the link between the joint distribution of abilities and the correlation of advantages in more detail in Section 5 below.

agriculture ($\bar{z}_1^a$ in the central figure) exceeds that of those who switch sector ($\bar{z}_S^a$). That is, average agricultural productivity increases as the sector shrinks. Panel (b) shows that the opposite is true if advantages are misaligned in agriculture. In this case, the average agricultural absolute advantage of those leaving agriculture exceeds that of those staying. That is, average agricultural productivity decreases as the sector shrinks. Hence, average productivity in agriculture increases as the sector shrinks, or sectoral size and average productivity in agriculture are negatively correlated, only if comparative and absolute advantage in agriculture are aligned.

Having illustrated the distinct roles of comparative and absolute advantage in a simple case with full specialization, we return to the general case where households can operate in both sectors simultaneously. First consider the households who actually do so. These households split their time to equate the marginal value products of labor across the two activities. As a result, their optimal labor allocation $\tilde{l}_i^a$ is implicitly defined by

$$\frac{z_i^a}{z_i^n} = \frac{1}{\kappa}\frac{g'(1-\tilde{l}_i^a)}{f'(\tilde{l}_i^a)}. \tag{5}$$

The fraction of time devoted to farming is an increasing function of their agricultural comparative advantage, i.e.

$$\frac{\partial \tilde{l}_i^a}{\partial (z_i^a/z_i^n)} = -\frac{\kappa f'\left(l_i^a\right)}{\frac{z_i^a}{z_i^n}\kappa f''\left(l_i^a\right) + g''\left(1-l_i^a\right)} > 0. \tag{6}$$

We can also use condition 5 to evaluate sectoral choices. Households with a strong agricultural comparative advantage will engage in farming only. These are households for which

$$\frac{z_i^a}{z_i^n} \geqslant \frac{1}{\kappa}\frac{g'\left(0\right)}{f'\left(1\right)}. \tag{7}$$

They are at a corner solution of their hours allocation. Households with a high comparative advantage in non-farm entrepreneurship will fully specialize in that sector. For these households we have

$$\frac{z_i^n}{z_i^a} \geqslant \frac{\kappa f'\left(0\right)}{g'\left(1\right)}. \tag{8}$$

Finally, households with intermediate levels of comparative advantage will operate in both sectors. These households have

$$\frac{z_i^a}{z_i^n} \in \left(\frac{1}{\kappa}\frac{g'\left(1\right)}{f'\left(0\right)}, \frac{1}{\kappa}\frac{g'\left(0\right)}{f'\left(1\right)}\right). \tag{9}$$

The equations above show that when a household is endowed with a pair of relatively similar abilities and as a result its comparative advantage is intermediate, diminishing returns to labor at the sectoral level make it optimal to split the time endowment between the two activities. Nonetheless, it is worth noticing that this intermediate comparative advantage is not informative of absolute advantages, z_i^a and z_i^n. Households operating in both sectors could be high in the marginal distributions of each ability, or could, equally well, be low. In the same fashion, selection is not informative about the absolute advantage of those who fully specialize in either

sector: selection is only informative about the fact that these households have fairly different abilities across sectors, but not about the level of these abilities.

Figure 2 makes this point graphically. Its structure is identical to that of Figure 1. Again, panel (a) shows the case where advantages are aligned in both sectors, and panel (b) the case where advantages are aligned in entrepreneurship but misaligned in agriculture. In each panel, the left figure shows a scatter plot of abilities, the central figure comparative advantage in agriculture against absolute advantage in agriculture, and the right figure comparative advantage in non-farming entrepreneurship against absolute advantage in the same sector. The lines of constant comparative advantage now split the population in three groups: those with strong comparative advantage in agriculture, those with intermediate comparative advantage, and those with strong comparative advantage in non-agricultural entrepreneurship.

This figure illustrates our identification strategy. Consider in particular the central figures. These show that the fact that we observe absolute advantage in agriculture not only for specialized farmers, but also for households who are active in both sectors, allows us to sign the correlation of advantages in agriculture. Panel (a) shows that a positive correlation of advantages in agriculture implies that specialized farmers have *higher* absolute advantage in agriculture than households engaged in both activities, whereas panel (b) shows that a negative correlation of advantages in agriculture implies that those who specialize in farming have *lower* absolute advantage in agriculture than those who do both. A similar reasoning applies to non-farming entrepreneurship.

As a consequence, a simple assessment of how comparative advantage – revealed through activity choices – varies with absolute advantage will reveal the sign of the correlation of advantages. If absolute and comparative advantages are positively correlated, absolute advantage in one sector should be negatively correlated with the likelihood of being active in the other sector. The opposite holds if absolute and comparative advantages are negatively correlated.[7]

The simple model we present can be extended by allowing for complementary inputs besides labor. Consider the production functions

$$\begin{aligned} y_i^a &= \kappa z_i^a F(l_i^a, k_i^a, T_i) \\ y_i^n &= z_i^n G(l_i^n, k_i^n) \end{aligned} \tag{10}$$

where k_i^j denotes capital used by household i in activity j, and T_i denotes land. In this case, a household specializes in agriculture if

$$\frac{z_i^a}{z_i^n} \geqslant \frac{G'(0, k_i^n)}{\kappa F'(1, k_i^a, T_i)} \tag{11}$$

In this more general setting, an assessment of how comparative advantage varies with absolute advantage still reveals the sign of the correlation of advantages. In addition, the comparative

[7] A simple comparison of means of absolute advantage across the two groups – specialized farmers and those doing both activities – can also reveal the sign of the correlation of advantages in agriculture. However, this approach is less amenable to empirical analysis in the presence of other, observable dimensions of household heterogeneity.

advantage threshold for specializing in agriculture is not common across households here, but depends on the quantities of land and capital that could be used in production. Use of more or better land – an increase in T – reduces the threshold. Capital affects the threshold if the two sectors differ in capital intensity. In the empirical analysis, we take these effects of complementary inputs on sectoral choice into account using controls for land size and wealth.[8] We also use village fixed effects to net out differences across villages in land quality and in relative sectoral productivity κ. These can arise from differences in factors like crop choices, market access, and weather.

To summarize, selection together with the correlations between absolute and comparative advantage at the sectoral level determines the link between sector size and sectoral productivity. A sector's productivity grows as the sector shrinks only if absolute and comparative advantage in the sector are aligned. Selection on the basis of comparative advantage places no restrictions on the sectoral correlations of advantages, which are ultimately determined by the underlying distribution of abilities. The empirical analysis that follows aims to sign the correlation between comparative and absolute advantage in each sector using the identification strategy outlined above.

3 Data

The data we use belong to the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) project. This is led by the World Bank in collaboration with several national statistical offices. In this section, we describe the main features of the dataset and the variables we use while referring to Appendix C for detailed information on the project, sampling frame, survey design, and variable definitions.[9]

In each partner country, the LSMS-ISA project supports multiple rounds of a nationally representative panel survey designed to gather information on agriculture, non-farm income activities, and socioeconomic status. Our final dataset combines the information on four countries – Ethiopia, Malawi, Nigeria, and Uganda – for which we can retrieve consistent information on all variables we use. The number of survey rounds or waves per country varies from 2 (Malawi) to 4 (Uganda), covering the years from 2009 to 2016.

3.1 Measurement

Value Added and Hours Worked First, we compute for each household in each wave a measure of value added in agriculture.[10] We follow Gollin, Lagakos, and Waugh (2014), and obtain

[8] As explained in Section 3, while we have information on both owned and rented land, we do not have consistent information on the value of assets and their specific use across activities. We proxy for it using a measure of assets owned by the household. In the presence of financial frictions, we expect this to proxy productive capital closely.

[9] See also http://surveys.worldbank.org/lsms [consulted on October 9, 2018].

[10] Both agricultural and non-agricultural activities are typically household enterprises, which implies that value added can only be measured at the household level.

it by adding value added from non-permanent crops, permanent crops, livestock, livestock products, and fishery.[11] In doing so, we follow closely de Magalhaes and Santaeulalia-Llopis (2018) who provide a thorough discussion of how to derive farming income from the LSMS-ISA data. A common issue in using this data set is that of assigning a monetary value to unsold agricultural production, which represents the majority of total household production. To determine the market value of production that was not sold, we either use the price at which the household sold that same crop or, if not available, the one reported by households in the same location that sold that crop, or the price recorded in the community-level survey.[12] We then calculate value added for each product category as the sum across seasons of each household's revenues from selling each product plus the market value of the product that was not sold minus the associated production costs.[13]

Second, we calculate profits from non-farm entrepreneurship. We identify all enterprises owned by any household member in the 12 months before the interview. For each one of them, we calculate profits as the difference between total annual sales and associated costs. We then aggregate these figures to derive total profits from non-farm entrepreneurship for each household in each wave.

To describe the activity of households along the intensive margin, we rely on the information provided on the number of hours allocated to each activity. Each household member is asked about the hours worked in the last 7 days on the household farm, in any of the household non-farming enterprises, and outside the household, in the form of paid or unpaid work, temporary or not, apprenticeship, etc. We calculate the total number of hours worked in farming and non-farm entrepreneurship at the household level by aggregating the hours worked in each activity across all household members.

We derive measures of value added in agriculture and profits from non-farm entrepreneurship using information on production, sales, and costs over the entire year. This contrasts with the information on hours worked, which pertains to the last 7 days before the interview is conducted. This information belongs to the time use section of the household questionnaire, which is typically administered together with the post-harvest one.[14] The seasonality of farming and non-farming activities may induce measurement error in these records of time use and their relationship with value added and profits across sectors. Notice however that this would be problematic for our empirical analysis only insofar as such measurement error correlates systematically with the variables of interest, a possibility that we discuss in detail in Section 8.

[11] As we show later, the empirical results are robust to excluding livestock and livestock products in the definition of farming activity and value added in agriculture.

[12] As reported later in Section 4.1, according to our calculations the fraction of market revenues from agriculture over the total value of agricultural output ranges from 20% in Ethiopia to 37% in Uganda.

[13] de Magalhaes and Santaeulalia-Llopis (2018) also discuss other measurement issues such as income underreporting and seasonality of reported consumption. Upon investigating the former, they conclude that this is not a major issue in the LSMS-ISA data as for instance the reported agricultural production and the reported annualized self-farmed consumption yield very similar quantities. They further conclude that seasonality affects measures of consumption – for which data are collected with recalls of the past 3 months, past month, or even past week – but not income – for which the recall period is 1 year. Seasonality may still matter for information on hours worked, an issue we discuss below and in Section 8.

[14] See Appendix for detailed information on the timing of such questionnaire in each country and wave.

Moreover, 77% (86%) of households for which we can derive value added in agriculture (profits from non-farm entrepreneurship) report a positive number of hours worked in that sector in the last 7 days. Perhaps more importantly, as we show later, our results are not sensitive to the choice of using value added or hours worked in the definition of households' activity along the extensive margin.

Measuring Absolute Advantage We measure absolute advantage in agriculture using value added and value added per hour. Similarly, we measure absolute advantage in non-farm entrepreneurship using profits and profits per hour.

We use two measures per sector to correct for the fact that hours choices naturally differ systematically between households active in both sectors and specialized households. With a production function that is increasing and strictly concave in hours, the larger hours input of households who specialize in a sector implies that comparing value added overstates their absolute advantage relative to households active in both sectors, while comparing value added per hour understates their absolute advantage. Both measures taken together bound the absolute advantage of specialized households relative to those active in both sector.

To make these measures comparable across countries and waves, we compute for each measure the percentile the household belongs to in the corresponding country-wave distribution. These are our preferred measures of absolute advantage. They are comparable across countries and waves, even in the presence of differences in currency used and inflation rates over time. [15]

For robustness, we also conduct our analysis directly using value added and value added per hour as measures of absolute advantage, coupled with country-wave fixed effects. Finally, we also estimate the production function in both sectors and use the estimated productivity term as a measure of absolute advantage.

Measuring Comparative Advantage Section 2 shows that the activity of each household along the extensive margin can be informative of its comparative advantage. We use the information on value added described above to also define the activity of each household along the extensive margin. That is, we say that a household is active in farming if we can derive information on value added in agriculture. Similarly, we say that a household is active in non-farm entrepreneurship if we can derive information on profits from that sector. Through the lens of the model, households that only do farming have high comparative advantage in agriculture; households that only do non-farm entrepreneurship have high comparative advantage in this sector; households that are active in both sectors have weak comparative advantage in both sectors.

For those households that are active in both farming and non-farm entrepreneurship, we can derive an additional measure of comparative advantage that is informed by their activity along

[15] Another advantage of using percentiles is that they are robust to the lack of information on hired labor: if the amount of hired labor on a household's farm increases with farming productivity, using value added per se would overestimate the level of absolute advantage in hiring households. But, this does not affect the corresponding ranking of households, leaving the percentile measure unaltered.

the intensive margin. Given that the production function is strictly concave in both sectors, equation 5 shows that households that have a comparative advantage in one sector also work relatively more hours in that sector. We can thus use the ratio between hours worked in the two sectors as a continuous measure of comparative advantage.

Additional Variables The data provide detailed information on each land plot operated by the household, from which we derive the total area of cultivated land. We complement this with information on ownership. The survey asks whether each plot of land is owned or assigned by decision of the local leader, inherited, or rented. We calculate the fraction of land that is rented, which we also consider as a proxy for local development of land markets. The survey also asks a number of questions about asset ownership. Household members are given a list of durable goods, and asked whether they possess any. This module is not always consistent across countries. We combine the available information in an asset index that counts the number of assets the household reports to have, which is specific of each country.[16] Finally, we derive information on the total number of household members and the total number of female household members, which we use as controls to evaluate the robustness of the empirical results.

3.2 Summary Statistics

Table 1 shows the summary statistics for the variables we employ in the empirical analysis.[17] For each variable, the table reports the sample average, its estimated standard error, and the number of observations. It does so separately across three groups of households: those active in farming only, those active in non-farm entrepreneurship only, and those active in both sectors. The final dataset counts around 35,000 household observations across all countries and waves. Overall, 59% of households are active in farming only and 12% do only non-farm entrepreneurship. The remaining 30% of households in the sample are active in both sectors. This number is large in all countries, ranging from 24% in Ethiopia to 38% in Nigeria.

For Ethiopia and Malawi, household-run enterprises are further classified into industries. We can use this information to get a better sense of the kind of non-farming enterprises run by households in these countries. Among the most represented, 28% of household enterprises in Ethiopia provide a non-agricultural service from home or a household-owned shop (such as carwash, metal processing, mechanic, carpenter, tailor, barber, etc.); 25% process or sell agricultural by-products (flour, local beer, seed, etc., but excluding livestock by-products and fish); 15% of enterprises belong to the category of trading business on a street or market, while 12% offer services or sell anything on a street or market (including firewood, home-made charcoal, construction timber, woodpoles, traditional medicine, mats, bricks, cane furniture, weave baskets, thatch grass, etc.). These numbers are quite similar in Malawi, where 25% of household enterprises provide a non-agricultural service from home or a household-owned shop, 15% pro-

[16] For this reason, in our empirical analysis, we allow the correlations of this asset index with the variables of interest to vary flexibly across countries.

[17] Table A.1 in Appendix A shows the summary statistics of main variables by country.

cess or sell agricultural by-products, 29% are trading businesses, and 16% offer services or sell anything on a street or in a market.

Households that are active in both farming and non-farm entrepreneurship differ from the others along a number of characteristics. First, they are significantly larger, counting 0.6 more members than households doing only farming and 1.2 more than households doing only entrepreneurship. Second, the total number of hours worked cumulatively by all members is larger in these households than those in the other two groups. The number of hours worked in a week in total is 90 for households that are active in both sectors as compared to 75 for those doing only entrepreneurship and 66 for those doing only farming. Yet, the number of hours allocated to each activity by households engaged in both sectors is lower than the one allocated by households active in one of the two sectors only. Table 1 also shows that households for which we cannot derive profits from non-farm entrepreneurship – which we classify as active in farming only – still report an average of 4 hours a week in total of work in that sector. The opposite is also true, as households for which we cannot derive value added in agriculture – which we classify as active in non-farm entrepreneurship only – report positive hours worked in agriculture on average. Our classification of households, which is based on the sectors in which they produce output, thus differs slightly from an alternative one based on the sectors to which they supply inputs, i.e. hours. Yet, as we show later, our results are not sensitive to the choice of using value added or hours worked in the definition of households' activity along the extensive margin.

Among those households that are active in both sectors, the fraction of those where at least one member reports to work a positive number of hours in both sectors is 50%, and the fraction where more than one member reports to work in both sectors is 23%. The average number of members reporting positive hours in both sector is one. Hence, in general, there is not full specialization across household members. We discuss in Section 7 the extent to which our empirical results at the household level are informative of the correlation of advantages and abilities at the individual level.

Table 1 also shows that the size of cultivated land is significantly higher for households active in both sectors than for households active in farming only, and that only 10% of households active in non-farm entrepreneurship have land. The asset index value suggests that households in this last group have on average more assets than others.

Evidence so far shows that around one third of households in our sample is active in both farming and non-farm entrepreneurship. It also shows that significant differences exist between these households and those active only in one sector. In the analysis that follows, we report unconditional estimates of the parameters that capture the relationships between the main variables of interest, but also evaluate the robustness of results by including additional household characteristics as controls and changing the definitions of sectoral activity whenever appropriate.

4 Selection Along the Extensive Margin

Our objective is to sign the correlation between absolute and comparative advantage in each sector. As discussed in Section 3, we measure absolute advantage in each sector using information on value added and profits while using household's activity along the extensive margin as a revealed measure of comparative advantage. According to theory, households active in both sectors have weak comparative advantage. It follows that, if absolute and comparative advantages are positively correlated, value added or value added per hour in one sector should be negatively correlated with the likelihood of being active in the other sector. The opposite holds if absolute and comparative advantages are negatively correlated.

Agriculture We first analyze households that are active in the agricultural sector. We restrict the sample to households that do any farming. Among those, we identify with a dummy equal to 1 those that are also active in non-farm entrepreneurship. The top left graph in Figure 3 illustrates the unconditional relationship between the two variables. It reports the fraction of households involved in non-farm entrepreneurship per bin of 5 percentiles of the distribution of value added in agriculture. The relationship is negative. The bottom left graph in Figure 3 is drawn in a similar fashion. It plots the relationship between entrepreneurship rates and value added per hour in agriculture. We observe no correlation between the two.

As discussed in Section 3, value added is a downward-biased measure of absolute advantage for those households that are engaged in both activities, while value added per hour is an upward-biased measure of absolute advantage for those same households. The two lines in the top and bottom left graphs in Figure 3 thus serve as bounds for the true relationship between comparative and absolute advantage in agriculture. They indicate that those with stronger absolute advantage are more likely to be specialized, indicating stronger comparative advantage. These graphs would thus suggest that the correlation between comparative and absolute advantages is positive in the agricultural sector.

Notice that we produced these figures comparing farmers surveyed in the same country and wave, but across locations (enumeration areas). Average returns from both activities may vary across space. The top and bottom right graphs in Figure 3 address this issue by considering the *residual* likelihood to be active in non-farm entrepreneurship after netting out average differences across locations.[18] In contrast with the left graphs, the relationship between agricultural value added and entrepreneurship becomes positive. The likelihood of engaging in non-farm entrepreneurship is higher for households at the top of the agricultural value added distribution, compared to those at the bottom. This suggests that the correlation between comparative and absolute advantages is negative in the agricultural sector.

Contrasting the right and the left graphs in the Figure 3, we infer that average differences across locations confound the relationship between agricultural value added and entrepreneurship at

[18] We regress a dummy equal to one if the household is active in non-farm entrepreneurship over the full set of location (enumeration area) fixed effects, and plot the corresponding estimated residuals.

the household level. In particular, evidence shows that entrepreneurship rates are systematically higher in those locations where agricultural value added is lower. To sign the correlation between absolute and comparative advantage in agriculture, comparing households across the entire distribution of agricultural value added is misleading, unless average differences across locations are taken into account and netted out in the analysis. Doing so reveals that among households within a location, households with higher agricultural productivity are the ones who are more likely to also pursue entrepreneurship.

We investigate these patterns systematically by implementing the following regression specification

$$Y_{igct} = \beta\, P_{ict} + \mathbf{X}'_{icgt}\gamma + \lambda_g + \delta_{ct} + \varepsilon_{igct}, \tag{12}$$

where Y_{igct} is the outcome of interest for household i surveyed in location g, country c, and wave t. P_{ict} is the percentile (divided by 10) the household belongs to in the distribution of absolute advantage in country c and wave t. In our first set of results discussed next, Y_{igct} is a dummy variable that equals 1 if a household is active in non-farm entrepreneurship, and P_{ict} is the percentile (divided by 10) the household belongs to in the distribution of absolute advantage in agriculture. Our coefficient of interest is β, which captures any systematic relationship between absolute advantage in agriculture and likelihood to engage in non-farm entrepreneurship. $\mathbf{X}_{icgt}$ is a vector of household-level characteristics. λ_g and δ_{ct} indicate location and country-wave fixed effects respectively. These capture and net out time-invariant location characteristics and differential trends in the likelihood of engaging in non-farm entrepreneurship between different countries and survey rounds. We allow the residual unobserved determinants of entrepreneurship ε_{igct} to be correlated among household-level observations that belong to the same location by clustering standard errors at the same level.

Table 2 shows the corresponding coefficient estimates. In column 1, we implement a regression specification that includes the household's percentile in the distribution of value added in agriculture (divided by 10), $P(VA_a)$ as a regressor, conditioning on the the full set of location fixed effects. In column 2, we instead use the percentile (divided by 10) in the distribution of value added per hour, $P(VA_a/h_a)$. Point estimates are consistent with the top and bottom right graphs in Figure 3. The point estimate of β is zero when considering agricultural value added, and positive and significant when considering value added per hour. Households that are higher in the distribution of agricultural value added per hour are more likely to engage in non-farm entrepreneurship than households that are lower.[19] In columns 3 and 4, we include as regressors a number of household-level characteristics, and the full set of country-wave fixed effects. Given the differences across countries in the way assets are recorded, we allow the coefficient of the asset index to vary flexibly across countries by including its interaction with the four country dummies. Coefficient estimates show that entrepreneurship rates are systematically higher among households that work more overall and have more female members. Entrepreneurship is also more likely among households that have more land and more assets, which is suggestive

[19] Table A.2 in Appendix A shows the coefficient estimates obtained without conditioning on location fixed effects, which are consistent with the top and bottom left graphs in Figure 3.

of the presence of fixed cost to start a non-farming enterprise combined with credit constraints. According to the results in column 4, households in the top percentile of the distribution of agricultural value added per hour are 7 percentage points more likely to engage in non-farm entrepreneurship than households in the bottom percentile. That is, the correlation of advantages is negative. The corresponding coefficient estimate is significant at the 1% level.

Non-farm entrepreneurship We next analyze households that are active in non-farm entrepreneurship. We restrict the sample to these households, and derive the percentile they belong to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. We also identify with a dummy equal to 1 those households that are also active in farming. Figure 4 mirrors Figure 3 and illustrates the relationship between these variables. The top and bottom left graphs of Figure 4 show that the likelihood of doing farming is significantly lower for those households at the top of the distribution of profits from entrepreneurship, suggesting that comparative and absolute advantage are positively correlated in this sector. But, the top and bottom right graphs show that this relationship disappears when comparing individuals within locations. This is consistent with the results we obtain when implementing the regression specification from equation 12, using the farming dummy as the dependent variable and the household's percentile in the distribution of profits from non-farm entrepreneurship as the main independent variable. Columns 5 to 8 of Table 2 show the corresponding coefficient estimates. Households that are higher in the distribution of profits from non-farm entrepreneurship or profit per hour are no more likely to engage in farming than households that are lower.

Summary The results so far suggest that absolute and comparative advantage are negatively correlated in agriculture, and not correlated in non-farm entrepreneurship. Importantly, differences across locations confound these correlations when estimated by comparing households across locations within countries. Figures A.1 to A.4 in Appendix A show that this pattern holds consistently in three out of the four countries in our sample, with Ethiopia being the exception.

In what follows, we discuss the possible mechanisms underlying this pattern. Yet, no matter what causes it, the correlation shown here is the one that determines the relationship between sectoral size and productivity. These findings cast doubt on self-selection on household unobservables playing a major role for the low average agricultural productivity in developing countries: if anything, the misalignment of advantages in agriculture together with selection should imply that average agricultural productivity in poor countries is relatively *high*.

4.1 Robustness

Alternative Definitions This section investigates the robustness of this first set of results. We start by using alternative definitions of a household's activity along the extensive margin. First, instead of classifying households' activity using information on value added in agriculture and profits from entrepreneurship, we use information on hours worked. In Table A.4 in Appendix

A, we report the coefficients we obtain when regressing a dummy equal to 1 if any household member reports any hours worked in any of the household non-farming enterprise on the household's percentile (divided by 10) in the distribution of value added in agriculture. Similarly, in Table A.5, we report the coefficient estimates from a regression of a dummy equal to 1 if any household member reports any hours worked in the household farm on the household's percentile in the distribution of profits from entrepreneurship. The results are very similar to those we report in Table 2.

Second, we adopt a stricter definition and label a household as active in non-farm entrepreneurship (farming) only if we can retrieve information on business profits (value added in agriculture) and if it devotes at least 15% of the total hours worked by household members to that activity. Tables A.6 and A.7 in Appendix A report the coefficient estimates when using as dependent variable the corresponding dummy. Results are once again similars to the ones obtained at baseline.

Third, we adopt a definition of farming activities that excludes livestock and related activities. We do so because live animals can also be considered assets and activities related to them are very different from the ones related to crops. We redefine value added in agriculture by excluding revenues and costs from growing and selling livestock and livestock products. This means that we also relabel households as active in farming if we can derive value added in agriculture excluding these activities. Tables A.8 and A.9 in Appendix A show the results when doing so, once again very similar to the ones in Table 2.

Specialization Even if the household as a whole is active in both farming and non-farm entrepreneurship, it could still be the case that each household member is fully specialized in only one of these activities. We defer a fuller discussion of this issue to Section 7, but already investigate the extent to which our results could be driven by full specialization within households. In Table A.10 (A.11) in Appendix A, we show the results obtained when adopting an alternative definition for extensive margin activity. We implement the same regression as in equation 12 but use as dependent variable a dummy that equals 1 if any household member reports any hours worked in a household non-farming enterprise (farm) *and* if at least one household member reports hours worked in both activities. The results are quantitatively and qualitatively similar to the ones obtained before.

Hours Worked for Others Next, we take into account the possibility that households supply hours of work outside the household in the form of paid or unpaid work, temporary or not, apprenticeship, etc. The average number of total hours worked by household members outside their household is equal to 15 in our data, which is not negligible. Tables A.12 to A.15 in Appendix A report coefficient estimates separately for the subsample of households in which any member reports positive hours worked outside the household and the subsample in which this is not the case. Once again, no meaningful differences emerge with the results reported in Table 2.

Alternative Measures of Absolute Advantages We have thus far measured absolute advantage using percentiles of value added and value added per hour in agriculture, and percentiles of profits and profits per hour in non-farm entrepreneurship. We can also use these variables directly as measures of absolute advantage. In order to reduce the sensitivity of results to extreme values, we apply a cube root transformation. This reduces the skewness of the distribution but, unlike the logarithm transformation, can also be applied to zero and negative values.[20] Table A.16 and A.17 in Appendix A show the corresponding coefficient estimates. These are consistent with those reported in Table 2.

In addition, we can use information on revenues and hours worked to directly estimate z^a and z^n. Specifically, we regress the log value of agricultural production on the log of hours worked in that sector, together with the full set of location and wave fixed effects. We take the residuals of this regression and derive the percentile the household belongs to in the corresponding country-wave distribution. We do the same in the other sector by using the log of annual sales of household-run non-farm entreprises and the log of hours worked in that sector.[21] We then use percentiles of these newly obtained measures of absolute advantage as the main regressor in equation 12. Table A.18 and A.19 in Appendix A show the corresponding coefficient estimates. These are consistent with those reported in Table 2. The only exception is the coefficient reported in the last column of Table A.19, which suggests that comparative and absolute advantages are positively correlated in the entrepreneurship sector, a result we will return to later on.

Subsistence vs. Market Production To conclude, we check whether systematic differences exist between households doing only farming and those that also engage in entrepreneurship in the split of agricultural output between internal consumption and market production. The fraction of market revenues from agriculture over the total value of agricultural output ranges from 20% in Ethiopia to 37% in Uganda. This shows that the majority of agricultural output is consumed within the household. Importantly for our analysis, these numbers are not meaningfully different between the group of households that only do farming and the one of households that also engage in non-farm entrepreneurship. The fraction of market revenues from agriculture over the total value of agricultural output ranges from 22 to 35% for the first group, and from 13 to 38% for the second group.

[20] Specifically, we apply the cube root transformation $\text{sign}(x) \times |x|^{1/3}$. In order to ease the interpretation of coefficient estimates, we also normalize the transformed variable by its standard deviation.

[21] The following caveats apply to these newly obtained measures of absolute advantage. First, the previous measures were based on value added and profits, thus taking into account the costs associated with each activity. We are here using the value of agricultural production and sales respectively, since this allows us to retain all observations, including those households that have negative values for profits. Second, hours worked are endogenous to absolute advantage, biasing the estimated coefficient and thus the residual we derive from these regressions. In the case of agriculture, we address this bias by adopting a control function approach and including a third degree polynomial of lagged production expenditures in the production function regression. We do not have comparable information for non-farming activities. Notice however that the bias induced by the endogeneity of hours worked does not affect the derived percentile measures insofar as it does not change the ranking of estimated absolute advantage across households.

To summarize, our results are robusts to several alternative definitions of a household's activity and an alternative, estimated measure of absolute advantage. They hold not only in the entire sample but also separately for the subsamples of households without internal specialization, those who supply hours worked outside the household, and those that do not. Finally, they are unlikely to be driven by differences in market production across groups. All of these results indicate misalignment of advantages in agriculture.

5 Mechanisms of Selection Along the Extensive Margin

What determines the correlation between absolute and comparative advantage that we observe in the data? The underlying distribution of abilities in the population plays an important role, which we explore first. We turn to the role of entry or operating costs in Section 5.2.

5.1 Underlying Distribution of Abilities

Denoting the correlation of advantages in agriculture and non-farming entrepreneurship by $\rho\left(z_i^a/z_i^n, z_i^a\right)$ and $\rho\left(z_i^n/z_i^a, z_i^n\right)$ respectively, the following proposition summarizes our results.

Proposition 1. *The signs of the (approximated) correlations between comparative and absolute advantage are given by*

$$\begin{aligned} sign\left[\rho\left(\frac{z_i^a}{z_i^n}, z_i^a\right)\right] &= sign\left[\frac{CV\left(z_i^a\right)}{CV\left(z_i^n\right)} - \rho\left(z_i^a, z_i^n\right)\right] \\ sign\left[\rho\left(\frac{z_i^n}{z_i^a}, z_i^n\right)\right] &= sign\left[\frac{CV\left(z_i^n\right)}{CV\left(z_i^a\right)} - \rho\left(z_i^a, z_i^n\right)\right] \end{aligned} \tag{13}$$

where $CV\left(z_i^j\right) = \sigma_j/\mu_j$ is the coefficient of variation in the population for sector $j = \{a, n\}$ and $\rho\left(z_i^a, z_i^n\right)$ is the correlation coefficient of abilities in the population. See Appendix A for a proof.

Abstracting from trivial cases where the distributions of sectoral abilities coincide or are degenerate in at least one sector, several insights arise from Proposition 1.[22] First, given that $\rho\left(z_i^a, z_i^n\right) \leq 1$, the correlation of advantages is always positive in one sector – the sector with higher dispersion of abilities as measured by the coefficient of variation.[23] For the sake of exposition, let us assume $CV\left(z_i^n\right) > CV\left(z_i^a\right)$, so that advantages are always aligned in

[22] When the coefficients of variation of abilities in both sectors coincide, $CV\left(z_i^a\right) = CV\left(z_i^n\right)$, the correlation of advantages will be positive in both sectors if abilities are not perfectly positively correlated. If $\rho\left(z_i^a, z_i^n\right) = 1$, advantages are uncorrelated in both sectors, $\rho\left(z_i^a/z_i^n, z_i^a\right) = \rho\left(z_i^n/z_i^a, z_i^n\right) = 0$. When the distribution of abilities in one sector is degenerate, for instance $CV\left(z_i^n\right) = 0$, abilities are uncorrelated in this sector, i.e. $\rho\left(z_i^n/z_i^a, z_i^n\right) = 0$. It is likely that these cases are not empirically relevant.

[23] Advantages can never be misaligned in both sectors: assume that advantages in agriculture are misaligned, i.e. those with low agricultural comparative advantage z^a/z^n have high agricultural absolute advantage z^a. If entrepreneurial advantages were also misaligned, those same households have not only high entrepreneurial comparative advantage z^n/z^a, but also low entrepreneurial absolute advantage z^n. But then these households have high z^a and low z^n, which contradicts the assumption on their comparative advantage.

entrepreneurship. Second, when abilities are not positively correlated, $\rho\left(z_i^a, z_i^n\right) \leq 0$, advantages are aligned in both sectors. Third, under positive correlation of abilities, advantages in agriculture will be aligned as long as $\rho\left(z_i^a, z_i^n\right) < CV\left(z_i^a\right)/CV\left(z_i^n\right)$, uncorrelated when $\rho\left(z_i^a, z_i^n\right) = \bar{\rho} \equiv CV\left(z_i^a\right)/CV\left(z_i^n\right)$, and misaligned otherwise. The more different the sectors are in terms of the dispersion of abilities in the population – as reflected by lower $CV\left(z_i^a\right)/CV\left(z_i^n\right)$ – the lower is the correlation of abilities in the population, $\bar{\rho}$, that ensures that advantages remain aligned in agriculture.

The first equation in 13 determines a threshold $\bar{\rho}$ for the correlation of abilities below which advantages in agriculture will be aligned, i.e. an upper bound for the correlation of abilities that ensures that they are, in the words of Young, "at worst weakly correlated" (Young 2014). This threshold, derived using a second-order approximation to the covariance of advantages, turns out to be exact under log-normal abilities (Heckman and Sedlacek 1985) or, in general, when the difference of log abilities is log concave (Heckman and Honoré 1990). Although widely used in the literature on selection of migrants (see for instance Borjas 1987), this result seems to have been overlooked in the recent work on selection and sectoral productivity differences.

This proposition suggests that our empirical finding of misaligned advantages in agriculture is generated by a joint distribution of abilities with a larger dispersion of abilities in non-farm entrepreneurship, combined with a *strong* positive correlation of abilities ($\rho > \bar{\rho}$). In this scenario, the best potential farmers mostly do not pursue farming, but entrepreneurship: with a high correlation of abilities, they are also the best potential entrepreneurs. The high dispersion of entrepreneurial abilities then makes entrepreneurship the more profitable choice. The worst potential farmers would also be the worst entrepreneurs, and thus choose farming due to its lower ability dispersion. Finally, there is an intermediate range of agricultural ability where part-time entrepreneurial activity is optimal.

This scenario also implies aligned advantages in non-farm entrepreneurship. The fact that we find no correlation of advantages in that sector suggests the presence of additional factors, besides the underlying distribution of abilities, that shape the correlation of advantages in the data. Next, we explore whether the introduction of fixed costs of entry may interact with selection in generating the observed correlations of advantages.

5.2 Sector-specific Fixed Costs

It is natural to think that sectoral choices could also be affected by the presence of fixed operating costs or entry costs. In fact, a prominent explanation for sectoral differences in labor productivity relies on the presence of such factors. We therefore now extend the model discussed in Section 2 along these lines and allow for the presence of fixed costs of operating in any of the two sectors, τ^j.[24] These costs, if negative, should be interpreted as amenities and, in principle, may be correlated with abilities.

As before, households take prices as given and allocate labor to maximize income net of oper-

[24] While we model the costs as fixed operating costs, fixed costs of entry would have a similar effect in our setting.

ating costs:

$$y_i = \kappa z_i^a f\left(l_i^a\right) - \tau^a \mathbf{1}(y_i^a > 0) + z_i^n g\left(1 - l_i^a\right) - \tau^n \mathbf{1}(y_i^n > 0), \tag{14}$$

where $\mathbf{1}(\cdot)$ is the indicator function. The *i-th* household compares the payoffs of being active in farming only, in non-farm entrepreneurship only or active in both sectors, and decides accordingly. This household will operate in both sectors as long as

$$\kappa z_i^a f\left(\tilde{l}_i^a\right) - \tau^a + z_i^n g\left(1 - \tilde{l}_i^a\right) - \tau^n \geqslant \max\left[\kappa z_i^a f\left(1\right) - \tau^a, z_i^n g\left(1\right) - \tau^n\right], \tag{15}$$

where $\tilde{l}_i^a$ is the optimal labor allocation. In terms of comparative and absolute advantage, this becomes

$$\kappa \frac{z_i^a}{z_i^n} f\left(\tilde{l}_i^a\right) - \frac{\tau^a}{z_i^n} + g\left(1 - \tilde{l}_i^a\right) - \frac{\tau^n}{z_i^n} \geqslant \max\left[\kappa \frac{z_i^a}{z_i^n} f\left(1\right) - \frac{\tau^a}{z_i^n}, g\left(1\right) - \frac{\tau^n}{z_i^n}\right]. \tag{16}$$

or equivalently

$$\kappa f\left(\tilde{l}_i^a\right) - \frac{\tau^a}{z_i^a} + \frac{z_i^n}{z_i^a} g\left(1 - \tilde{l}_i^a\right) - \frac{\tau^n}{z_i^a} \geqslant \max\left[\kappa f\left(1\right) - \frac{\tau^a}{z_i^a}, \frac{z_i^n}{z_i^a} g\left(1\right) - \frac{\tau^n}{z_i^a}\right]. \tag{17}$$

The first expression provides information about the sign of the correlation of advantages in non-farm entrepreneurship $\rho\left(z_i^n / z_i^a, z_i^n\right)$, while the second is informative about the same correlation in agriculture, $\rho\left(z_i^a / z_i^n, z_i^a\right)$.[25]

The last two expressions make clear that, in the presence of barriers, sectoral choices are not uniquely determined by comparative advantage, but also depend on the level of fixed costs, τ^a and τ^n, and on absolute advantages. It follows that in the presence of fixed operating costs the correlation of advantages captured by our empirical approach is not generated by the underlying distribution of abilities alone, but also depends on its interaction with fixed costs.

We illustrate this last point through an example. Assume the correlation of abilities is positive but weak, in particular, $\rho\left(z_i^a, z_i^n\right) = \bar{\rho} = CV\left(z_i^a\right)/CV\left(z_i^n\right) \in (0, 1)$. According to Proposition 1, in such a scenario advantages will be aligned in entrepreneurship, $\rho\left(z_i^n / z_i^a, z_i^n\right) > 0$, and uncorrelated in farming, $\rho\left(z_i^a / z_i^n, z_i^a\right) = 0$. Panel (a) of Figure 5 reproduces this scenario in the absence of barriers. On the one hand, since abilities are positively correlated and they are more dispersed in entrepreneurship, the most able entrepreneurial households fully specialize in this activity. It follows that comparative and absolute advantage are aligned in entrepreneurship. On the other hand, those with a high comparative advantage in agriculture specialize in farming. For some of them agricultural comparative advantage is high because they are good farmers, but for some others it is high because they are very poor entrepreneurs. Those engaged in both activities have weak comparative advantage in agriculture. Some of them are relatively good at both activities, while others are relatively bad at both. As a result, the average agricultural ability of those specialized in farming turns out to coincide with that of those engaged in both activities: $\bar{z}_A^a = \bar{z}_B^a$, and advantages are correctly measured as uncorrelated in agriculture.

[25] In Appendix A we show that $sign\left[\rho\left(z_i^a / z_i^n, z_i^n\right)\right] = -sign\left[\rho\left(z_i^n / z_i^a, z_i^n\right)\right]$.

Consider now the introduction of a fixed cost to enter non-farm entrepreneurship ($\tau^n > 0$). This situation is illustrated in panel (b) of Figure 5, with the broken line indicating indifference between farming only and both activities – the choice between non-farm entrepreneurship and both activities is unaffected, as the fixed cost is due in both cases. The fixed cost pushes some of the households that in the absence of this cost would choose to engage in both activities to do only farming. These households have relatively low entrepreneurial ability. Given the positive correlation of abilities, these households also have relatively low agricultural ability. It follows that the group of households that remains active in both sectors has, on average, higher ability in both activities. This reduces the correlation of advantages in both sectors. In agriculture it turns from zero to negative (as $\bar{z}^a_F < \bar{z}^a_B$). In entrepreneurship, it is reduced, and for large enough τ^n can change from positive to zero (if $\bar{z}^n_B = \bar{z}^n_E$).

The results in Section 4 show that absolute and comparative advantage are negatively correlated in agriculture and uncorrelated in non-farm entrepreneurship. This section shows that, while not consistent with selection solely based on comparative advantage, the observed data pattern can be rationalized in the presence of fixed entry costs.

This analysis suggests an additional empirical approach to inferring some of the properties of the underlying distribution of abilities. For this, note that fixed entry costs (or amenities) do not affect the optimal allocation of labor to activities for those households who engage in both activities – equation 5. As shown in equation 6, the hours supplied to an activity increase in a household's comparative advantage in that activity. As a consequence, the ratio between total hours worked in the two sectors is a revealed measure of comparative advantage that is unaffected by any fixed costs of operation or entry. Any systematic relationship between value added and relative labor supply in a sector then is informative of the correlation between absolute and comparative advantage net of fixed entry costs.[26] We next turn to estimating this relationship.

6 Selection Along the Intensive Margin

To investigate the relationship between comparative advantage as revealed by relative hours supplied to each sector and absolute advantage in the sector, we now restrict the sample to households that are active in both farming and non-farm entrepreneurship. As in Section 4, we start by investigating sectoral correlations in agriculture. We implement the regression specification given in equation 12, with relative labor supply in agriculture – the ratio of total hours worked in agriculture over those in non-farm entrepreneurship – as the dependent variable. We start by including as the only regressor the household's percentile (divided by 10) in the distribution of agricultural value. Column 1 of Table 3 reports the corresponding coefficient estimate. We condition on the full set of location fixed effects, and cluster standard errors at the same level. The estimated relationship is positive and significant at the 10% level. Households in the top percentile of the agricultural value added distribution on average work slightly more hours in agriculture relative to entrepreneurship than households in the bottom percentile. In column 2,

[26] This is the case even if fixed costs are heterogeneous across households.

we instead use the household's percentile in the distribution of agricultural value added per hour as the main regressor. The coefficient of interest is negative, and highly significant, indicating that, among households engaged in both activities, households with higher agricultural value added per hour work significantly fewer hours in this sector relative to non-farm entrepreneurship. Note that this coefficient is much larger in absolute value than the coefficient in column 1. Table 1 shows that the average household active in both sectors allocates 40.7% of total hours worked to agriculture (36.5 hours compared to 53.1 hours in entrepreneurship). Taking this as a benchmark, the estimate in column 2 of Table 3 implies that moving up one decile in the distribution of agricultural value added per hour is associated with a reduction in the share of time allocated to agriculture of about 5 percentage points, or 4.5 hours.

As discussed earlier, decreasing returns to scale imply that value added per hour is an upward-biased measure of absolute advantage for households that are active in both sectors, and more so the lower the absolute amount of hours worked. This implies that the estimate in column 2 is a lower bound for the true correlation between absolute and comparative advantage in agriculture, while the estimate in column 1 is an upper bound. The ordering of the two estimated coefficients is consistent with this bounding argument.

In column 3 and 4, we include the full set of household-level controls together with country-wave fixed effects. The relative supply of hours worked in agriculture is higher for households with more land, and for those that work more hours in total. We also find some evidence that the amount of hours worked in agriculture relative to non-farm entrepreneurship is lower for households with more assets. Perhaps more importantly, the estimates of our main coefficient of interest in columns 3 and 4 support the same conclusion as those in columns 1 and 2: Households that are more productive in agriculture supply relatively fewer hours in that sector, a sign of weak comparative advantage. We conclude that absolute and comparative advantages are negatively correlated in agriculture. This result suggests that the patterns we found in Section 4 for the agricultural sector cannot be entirely due to the presence of fixed costs, and must at least in part be due to selection on ability.

Section 5 showed that, in the absence of fixed costs, misalignment of advantages in agriculture implies alignment of advantages in entrepreneurship. Columns 5 to 8 of Table 3 provide evidence of the latter.[27] We test whether there is a systematic relationship between relative labor supply to non-farm entrepreneurship and profits in that sector. In column 5 of Table 3, we show coefficients from a regression of the relative labor supply in non-farm entrepreneurship – the ratio of total hours worked in non-farm entrepreneurship over those in agriculture – on the household's percentile (divided by 10) in the distribution of profits from entrepreneurship. The regression models and estimates in columns 6 to 8 are ordered as in columns 2 to 4. Once again, the ranking of the estimated coefficients is consistent with the bounding argument outlined above, with the estimate in columns 5 (and 8) being a lower bound for the true correlation between absolute and comparative advantage in entrepreneurship, and the estimate in column 5 (and 7) being an upper bound. Households with higher profits from non-farm entrepreneurship work significantly more hours in this sector relative to agriculture, while no systematic differ-

[27] Note that this is not directly implied by the findings in columns 1 to 4 of Table 3.

ences emerge in relative labor supply across percentiles of the distribution of hourly profits. Taking again the time allocation of the average household active in both sectors as benchmark, the estimate in column 5 of Table 3 implies that moving by one decile in the distribution of profits from entrepreneurship is associated with an increase in the share of time allocated to entrepreneurship of about 2.3 percentage points, or about 2 hours. We conclude that absolute and comparative advantages are positively correlated in non-farm entrepreneurship.

Abilities, Frictions, and Selection Summarizing, our analysis finds a negative correlation of advantages in agriculture at both the extensive and intensive margins, a positive correlation of advantages in non-farm entrepreneurship at the intensive margin, and no significant correlation of advantages in non-farm entrepreneurship at the extensive margin.

The discussion in Section 5 indicates a unique setting that is consistent with these observations. First, the patterns of alignment at the intensive margin indicate that the coefficient of variation is higher for the distribution of ability in non-farm entrepreneurship relative to agriculture, $CV\left(z_i^n\right) > CV\left(z_i^a\right)$, and that the correlation of abilities is high, i.e. $\rho\left(z_i^a, z_i^n\right) > \bar{\rho} = CV\left(z_i^a\right)/CV\left(z_i^n\right)$. Second, the absence of correlation of advantages at the extensive margin in non-farm entrepreneurship indicates the presence of fixed costs in entrepreneurship.

Robustness As in Section 4.1, we verify whether the results we obtain in this section are robust to alternative definitions of activity along the extensive margin. In Tables A.20 and A.21 in Appendix A, we restrict the sample to those households that report positive hours worked in both sectors, with no meaningful changes to the results. In Tables A.22 and A.23 we restrict the sample to households that devote at least 15% of the total hours worked by household members to each activity. In this case, results are less conclusive because of the reduced variation in the dependent variable. This is not the case when we adopt a stricter definition of farming activities that excludes livestock and related activities. Results in Tables A.24 and A.25 are again very similar to those presented in Table 3.

In Tables A.26 and A.27 in Appendix A, we restrict the sample to households that are not fully specialized, i.e. where at least one household member reports hours worked in both the household non-farming enterprise and the household farm. Coefficient estimates are similar to the ones reported in Table 3. Finally, results in Tables A.28 to A.31 show that, with the exception of Table A.31, estimates are consistent with the baseline results discussed above both in the subsample of households with positive hours worked outside the household and in the subsample without.

7 Households, Individuals, and Selection Over Time

Our findings indicate that comparative and absolute advantage are negatively correlated in the agricultural sector, and positively correlated in entrepreneurship. Given our unit of observation

in the data, these findings apply to households, not individuals. This motivated our assumption that production and economic choices occur at the household level. In our theoretical framework, ability or productivity are household-level attributes or, alternatively, the attributes of a single household member who acts as manager and makes production decisions on behalf of all members. A natural question is whether our findings at the household level can also be informative of the correlation of advantages and abilities at the individual level. In particular, we want to rule out the case in which comparative and absolute advantage in agriculture are negatively correlated at the household level, but positively correlated at the individual level.

7.1 Households or Individuals: Theoretical Insights

To address this concern, we first consider an alternative model that endows each individual in a household with a vector of sector-specific ability or productivity. Households then choose an allocation of individuals and their working hours to activities. Abstracting from productivity interactions across individuals in a household, individuals will sort into activities based on their individual comparative advantage. In this model, a household will specialize in an activity if all its members have strong comparative advantage in that activity. It will engage in both activities either if household members have strong comparative advantage in different activities, or if one or more members have weak comparative advantage and therefore do not specialize.

This is illustrated in Figure 6, where now every dot represents an individual. In the figure, individuals with strong comparative advantage specialize in either agriculture or entrepreneurship, while those with intermediate comparative advantage pursue both activities. The lines of indifference are drawn to yield proportions of individuals engaged in each activity similar to the data.

A negative correlation of advantages in agriculture at the household level can come about in only two ways. The first one is shown in panel (a): if the correlation of advantages at the individual level is negative and entrepreneurial abilities are more dispersed, then the correlation at the household level is also negative if individuals within a household are relatively "similar". (Some illustrative households are labelled A, B, etc.) Only then is it the case that the households with the lowest agricultural productivity are more likely to specialize in farming. This situation is akin to multi-dimensional positive assortative matching as defined in Lindenlaub (2017).[28] In this case, individual-level productivity ranks are similar to household ones.

The second possibility is shown in panel (b): if the correlation of advantages at the individual level is positive, the correlation at the household level is negative if individuals within a household are relatively "different" (multi-dimensional negative assortative matching). Only then is it the case that the households with the highest agricultural productivity are active in both sectors (e.g. household A), while those with low agricultural productivity are active only in agriculture (e.g. household C).[29]

[28] Multi-dimensional sorting problems are very challenging and the literature studying them is in its infancy. Therefore we do not study a full model in this section, but resort to a graphical representation.

[29] This is similar if the correlation of advantages at the individual level is negative and the dispersion of agricultural

In the data, slightly more than half the households specialize in agriculture. In panel (a), these households consist of individuals with strong comparative but weak absolute advantage in agriculture – selection at the individual level mirrors that at the household level. In panel (b), they consist of individuals with weak comparative advantage in agriculture (e.g. household C). Individuals with strong comparative advantage in agriculture meanwhile match with others with strong comparative advantage in entrepreneurship, and thus are in households that are active in both sectors (like household A).

Importantly, the two settings yield different predictions regarding which households will first enter entrepreneurship as the threshold determining selection across sector changes over time. Consider a decline in κ, driven by any factor that makes agricultural work relatively less attractive. This change makes the indifference lines pivot counterclockwise. In panel (a), it is clear that this prompts the most productive farmers to take up entrepreneurship. When individuals match with similar individuals, these farmers will come from households with high agricultural productivity. In panel (b), in contrast, it is clear that the least productive specialized farmer will switch.

In the following, we exploit the panel dimension of the data, which allows us to capture switches in a household's activity. We investigate the behavior of households engaged in different activities at different points in time, with a particular focus on determining which of these two scenarios is a better description of the data.

7.2 Selection Over Time

We begin by reporting in Table 4 the fraction of households in each wave that is active in agriculture only, in non-farm entrepreneurship only, or in both. Evidence shows that the fraction of households active in both sectors has been growing between 2009 and 2016, from 26% to 37%. This is true in all countries in our sample with the exception of Uganda.[30] The fraction of households active in farming only has decreased in Malawi and Nigeria, but remained stable in Ethiopia and Uganda. Table A.33 in Appendix A reports the transition matrices across these different groups between waves 1 and 2, and 2 to 3. The fraction of households transitioning from being active only in one sector into being active only in the other is negligible, while transitions from doing only farming (entrepreneurship) to doing both and vice versa are more common, covering around 10% (2%) of households in the sample. In light of these non-trivial transition probabilities, we can complement the cross-sectional analysis above with a systematic analysis of sectoral transitions.

We implement a panel data regression analysis. We restrict the sample to households that in wave 1 are only active in farming, and investigate their probability of being active in non-farm

abilities is larger.

[30] Table A.32 reports the same numbers separately for each country and wave.

entrepreneurship through wave 3. We implement the following regression specification

$$Entrep_{igct} = \sum_{t=2}^{3} \beta_t \, Wave_t \times Rank_{ig} + \mathbf{X}'_{icgt}\gamma + \lambda_i + \delta_{ct} + \varepsilon_{igct}, \tag{18}$$

where $Entrep_{igct}$ is a dummy variable that equals 1 if household i surveyed in location g, country c, and wave t is active in non-farm entrepreneurship. $Wave_t$ is a wave dummy identifier. $Rank_{ig}$ is defined according to where the household stands in the location-specific ranking of agricultural value added and agricultural value added per hour in the first wave of data. That is, $Rank_{ig}$ is time-invariant and takes a value of 1 if household i is the most productive farming household in its location g in the first wave of the data, 2 if it is the second most productive, etc. $\mathbf{X}_{icgt}$ is a vector of household-level characteristics. λ_i and δ_{ct} capture household and country-wave fixed effects respectively, which allow to control for and net out both time-invariant household-level characteristics and country-specific time trends. As before, we allow the residual unobserved determinants of entrepreneurship ε_{igct} to be correlated among household-level observations that belong to the same location by clustering standard errors at the same level. The coefficient β_t captures whether the likelihood of taking up non-farm entrepreneurship in wave 2 or 3 is correlated with the household's absolute advantage in agriculture.

This analysis is useful for two reasons. First, by revealing the position of switchers – who have the weakest comparative advantage among all initially specialized households – in the distribution of absolute advantage, it yields estimates of the correlation of advantages conditional on household fixed effects. These allow to capture and net out time-invariant unobserved differences in e.g. wealth and access to technologies beyond those captured by observables. Second, as discussed in the previous subsection, it reveals whether the correlation of advantages we find at the household level is driven by a similar correlation of advantages at the individual level.

Table 5 reports the estimated coefficients across different regression specifications, from one that includes only household and wave fixed-effects to the fully saturated one. As in the previous analysis, we define the ranking position of the household in terms of either agricultural value added or agricultural value added per hour. The estimated β_t is negative and significant for all waves and across all specifications. Columns 1 and 2 are consistent with each other in showing that households having a lower rank, i.e. higher agricultural value added or value added per hour in wave 1, are differentially more likely to take up non-farm entrepreneurship in subsequent waves. The magnitude and significance of coefficient estimates is only marginally affected by the inclusion of time-varying household-level controls in columns 3 and 4. This pattern is remarkably consistent across countries, as indicated by the coefficient estimates reported in Table A.34 in Appendix A, and despite the fact that the time interval between waves is different across countries.

We can exploit the panel dimension of the data to also investigate the role played by changes in household composition. In survey waves other than the first, we can identify households members that were previously listed, but moved out in the time between the previous and the current interview. We define for each household a dummy equal to 1 if any household member

moved out since the last interview and, similarly to the analysis in Section 4, we regress it on the household's percentile in the distribution of value added (or value added per hour) in agriculture, and again on the household's percentile in the distribution of profits from entrepreneurship (or profits per hour). Tables A.35 and A.36 in Appendix A report the corresponding coefficient estimates. We find some evidence that the exit of members is systematically more likely to occur among more productive households. If these members were to migrate for work outside of agriculture, this piece of evidence would be once again consistent with misalignment of advantages in that sector. But coefficient estimates are no longer statistically significant when we control for household characteristics and location fixed effects.

The evidence in this section is consistent with the one presented in Section 4 and 6. Farming households at the margin of entrepreneurship have a lower comparative advantage in agriculture than inframarginal ones. It is thus natural that their gains from switching sector are limited (Hicks, Kleemans, Li, and Miguel 2017). Yet, evidence shows that they are among the most productive farming households. Results from this panel data analysis provide further indication that absolute and comparative advantage are negatively correlated in agriculture, and that the household level patterns revealed in the previous sections are driven by similar patterns at the individual level.

8 Alternative Explanations

In this section, we explore several mechanisms other than selection on ability, grouped into a few distinct categories, and discuss to what extent they are consistent with the data.

Distortions Along the Intensive Margin The presence of constraints along the intensive margin may affect the allocation of hours worked across sectors within the household. For example, it could be the case that the effective marginal cost of agricultural inputs or capital is higher for some farming households, and that this induces them to allocate more of their time to non-farm entrepreneurship. But the results in Section 4 show that it is the most productive farming households who are systematically more likely to engage in non-farm entrepreneurship, both at the extensive and intensive margin. Constraints at the intensive margin would reduce input use and weaken the absolute advantage of these households – but not overturn it. Perhaps more importantly, all results are not sensitive to controlling for various household characteristics such as size of land, fraction of land rented, asset index, etc., which we would expect to correlate with constraints – or size-dependent distortions – to both agricultural and non-agricultural activities along both the extensive and intensive margin.

Diversification as Insurance The choice of the household may be driven by considerations other than joint profit maximization across activities. In particular, farming households may turn to non-farm entrepreneurship in response to negative shocks to agricultural output. This is consistent with the notion of *necessity entrepreneurs*, see De Giorgi and Di Falco (2018) among

others. Yet, this appears once again inconsistent with our findings, as households affected by a negative shock and therefore turning to entrepreneurship should have lower agricultural value added. We find instead that entrepreneurship rates are higher among most productive farming households.

Alternatively, households may choose *ex ante* to diversify, to reduce risk. For this to drive our finding that more productive farming households are more likely to engage in entrepreneurship, it would be necessary for these to be more risk averse, which is possible but seems implausible. Moreover, it is not clear how this would explain the intensive margin results showing that more productive farmers allocate more of their time to non-farm entrepreneurship.

Seasonality Returns to agricultural work have a strong seasonal component. During low agricultural season, households may allocate less of their time to agriculture and more of it to non-farm entrepreneurship. Our finding that more productive farmers allocate more of their time to non-farm entrepreneurship is based on information from the post-harvest questionnaire. Seasonality in agriculture can explain this result only insofar as, at the time of the interview, returns from agricultural work are differentially lower for more productive farmers, prompting them to increase the time they allocate to entrepreneurship. This could be the case if, for instance, more productive farmers grow a set of crops that require less post-harvest work. Yet, this does not explain why these farmers are also highly productive entrepreneurs, unless abilities are strongly positively correlated across sectors.

Heterogeneous Fixed Costs One possible reason why entrepreneurship rates are higher among the most productive farming households is that they face lower costs to enter entrepreneurship. For this factor to drive our findings, these costs would need to be orthogonal to all household characteristics we control for. But then again, our analysis in Section 6 is robust to the presence of fixed entry costs.

Missing Land Market If land endowments were fixed and there was no way to sell or rent out land, households with a high comparative advantage in non-farm entrepreneurship would still use this land and thus remain active in farming. This could explain why these households are not systematically different in terms of profits from entrepreneurship. It would also be consistent with the evidence that 90% of the households in our sample who only pursue non-farm entrepreneurship report to have no land. Yet it cannot explain the observed negative correlation of advantages in agriculture, nor the allocation of hours among those pursuing both activities.

Suppose that, in addition, an exogenous production capacity constraint puts a strict upper bound on agricultural output. The most productive farming households hit such a constraint earlier and are pushed into non-farm entrepreneurship. This would be consistent with the results in columns 1 to 4 of Table 2 showing that entrepreneurship rates are not systematically different across households at different percentiles of the distribution of value added in agriculture while they increase systematically with the percentile the household belongs to in the distribution

of agricultural value added per hour. Among households active in both activities, the same limit on output implies that the most productive farming households are left with more hours to allocate to non-farm entrepreneurship, thus have higher profits and – if in addition the production function in non-farm entrepreneurship is close to linear – no different profits per hour. This would be consistent with the results presented in Table 3.

However, this scenario not only features very strong assumptions, but is also incompatible with some of the other empirical results. First, there is substantial variation in agricultural output, also among those active in both activities. This would require heterogeneity in the bound on agricultural output. Second, in Table 3, the main coefficient in column 5 is very similar to the one in column 7. This indicates that, in a scenario where the relationship between value added in entrepreneurship and supply of labor to that sector is uniquely determined by the upper bound to agricultural production, the latter needs to be orthogonal to all other household characteristics that we include as controls, in particular assets. This is unlikely.

Finally, the mechanisms we are considering here still do not explain why, among households active in both sectors, absolute advantages across sectors are significantly positively correlated. This is empirically true using the various measures considered thus far, i.e. agricultural value added and value added per hour, profits from non-farm entrepreneurship and profits per hour, estimated absolute advantages. In other words, it does not explain why those same highly productive farming households that hit the production capacity constraint earlier and do more non-farm entrepreneurship are also highly productive entrepreneurs. This finding is instead fully consistent with our main interpretation of results and the selection mechanism we propose.

9 Conclusions

Agricultural productivity is lower than that of other sectors. The agricultural productivity gap is particularly large in poor countries. A recent influential literature argues that an important source of this difference is worker self-selection. This mechanism relies on a positive correlation of comparative advantage and absolute advantage in the agricultural sector. We test this hypothesis using household-level data from Ethiopia, Malawi, Nigeria, and Uganda. Our empirical analysis delivers four sets of results. First, around one third of households engage in both agriculture and non-farm entrepreneurship. Second, those households active in both sectors have systematically higher agricultural productivity than those doing only farming. Third, among households active in both sectors, those with higher agricultural productivity supply relatively fewer hours in agriculture while those with higher profits from entrepreneurship supply relatively more hours in this sector. Fourth, over time, households starting a non-farming enterprise have higher baseline agricultural productivity than those who remain only farmers.

These results all imply that comparative and absolute advantage are misaligned in agriculture, casting doubt on the importance of worker self-selection as a root cause of the agricultural productivity gap. The literature suggests other possible explanations such as distortions to the land market (Adamopoulos and Restuccia 2014), or to the use of intermediate inputs (Donovan

2018). Yet, some of our results suggest that selection may still play a role, but along a different margin: land quality (Adamopoulos and Restuccia 2018). When comparing households across locations, the evidence in Section 4 shows that non-farm entrepreneurship rates are higher in places where agricultural productivity is lower. The reason for this could be differences in land quality. It might be the case that, as the agricultural sector shrinks, average agricultural productivity could increase not because the worst farmers switch to non-agriculture – as the worker self-selection story would argue – but because the worst agricultural land is converted to other uses or abandoned. Across countries, only the most productive land would be devoted to agriculture in rich countries, while in poor countries, less suitable land would also be used for farming. We are exploring this hypothesis in separate work.

References

Adão, R. (2016). Worker Heterogeneity, Wage Inequality, and International Trade: Theory and Evidence from Brazil. Mimeo.

Adamopoulos, T., L. Brandt, J. Leight, and D. Restuccia (2021). Misallocation, Selection and Productivity: A Quantitative Analysis with Panel Data from China. *Econometrica, accepted*.

Adamopoulos, T. and D. Restuccia (2014). The size distribution of farms and international productivity differences. *American Economic Review 104*(6), 1667–97.

Adamopoulos, T. and D. Restuccia (2018, April). Geography and Agricultural Productivity: Cross-Country Evidence from Micro Plot-Level Data. NBER Working Papers 24532, National Bureau of Economic Research, Inc.

Alvarez, J. A. (2019). The Agricultural Wage Gap: Evidence from Brazilian Micro-Data. *American Economic Journal: Macroeconomics*. forthcoming.

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

Caselli, F. (2005). Accounting for Cross-Country Income Differences. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth*, Volume 1 of *Handbook of Economic Growth*, Chapter 9, pp. 679–741. Elsevier.

De Giorgi, G. and S. Di Falco (2018). Farmers to Entrepreneurs. Mimeo.

de Magalhaes, L. and R. Santaeulalia-Llopis (2018). The consumption, income, and wealth of the poorest: An empirical analysis of economic inequality in rural and urban Sub-Saharan Africa for macroeconomists. *Journal of Development Economics 134*(C), 350–371.

Donovan, K. (2018). Agricultural Risk, Intermediate Inputs, and Cross-Country Productivity Differences. Mimeo.

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

Gollin, D., S. Parente, and R. Rogerson (2002). The role of agriculture in development. *American economic review 92*(2), 160–164.

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

Heckman, J. J. and G. Sedlacek (1985). Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy 93*(6), 1077–1125.

Herrendorf, B. and T. Schoellman (2015). Why is Measured Productivity so Low in Agriculture? *Review of Economic Dynamics 18*(4), 1003–1022.

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

Hicks, J. H., M. Kleemans, N. Y. Li, and E. Miguel (2017). Reevaluating Agricultural Productivity Gaps with Longitudinal Microdata. NBER Working Papers 23253, National Bureau of Economic Research, Inc.

Jones, C. I. and P. M. Romer (2010). The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital. *American Economic Journal: Macroeconomics 2*(1), 224–245.

Lagakos, D., S. Marshall, A. M. Mobarak, C. Vernot, and M. E. Waugh (2020). Migration costs and observational returns to migration in the developing world. *Journal of Monetary Economics*.

Lagakos, D., A. M. Mobarak, and M. E. Waugh (2018). The welfare effects of encouraging rural-urban migration. Working Paper 24193, National Bureau of Economic Research.

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

Lindenlaub, I. (2017). Sorting multidimensional types: Theory and application. *The Review of Economic Studies 84*(2), 718–789.

Pulido, J. and T. Święcki (2018). Barriers to Mobility or Sorting? Sources and Aggregate Implications of Income Gaps Across Sectors and Locations in Indonesia. Mimeo.

Restuccia, D. and R. Santaeulalia-Llopis (2017). Land misallocation and productivity. Working Paper 23128, National Bureau of Economic Research.

Restuccia, D., D. T. Yang, and X. Zhu (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.

Young, A. (2013). Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics 128*(4), 1727–1785.

Young, A. (2014). Structural transformation, the mismeasurement of productivity growth, and the cost disease of services. *American Economic Review 104*(11), 3635–67.

Tables and Figures

Table 1: Summary Statistics

	Only Agriculture	Only Entrep.	Both	Full Sample
Observations	20621	4101	10376	35098
	59%	12%	30%	100%
Household Size	5.274	4.687	5.867	5.381
	(0.019)	(0.042)	(0.027)	(0.015)
	20536	4093	10361	34990
Female HH Members	2.106	1.941	2.108	2.087
	(0.012)	(0.027)	(0.020)	(0.010)
	20536	4093	10361	34990
Hours in Agriculture	47.283	4.141	36.575	39.070
h_a	(0.385)	(0.269)	(0.460)	(0.276)
	19850	3940	10176	33966
Hours in Entrepreneurship	18.541	70.744	53.094	34.948
h_n	(0.270)	(0.856)	(0.510)	(0.264)
	19850	3940	10176	33966
Total Hours	65.664	75.020	90.140	73.993
$h_a + h_n$	(0.501)	(0.904)	(0.730)	(0.384)
	20621	4101	10376	35098
Hours in Agriculture	59.189		52.273	48.406
$h_a > 0$	(0.434)		(0.563)	(0.320)
	15857		7120	27078
Hours in Entrepreneurship		76.407	63.962	25.034
$h_a > 0$		(0.858)	(0.543)	(0.249)
		3648	8447	32716
HH Members with			0.938	0.277
$h_a, h_n > 0$			(0.014)	(0.005)
			10361	35083
Female HH Members with			0.211	0.048
$h_a, h_n > 0$			(0.006)	(0.001)
			7285	32007
Land Size (ha)	1.488	0.516	2.464	1.782
	(0.087)	(0.086)	(0.899)	(0.289)
	19297	410	9076	28783
Fraction Rented	0.068	0.115	0.070	0.070
	(0.002)	(0.016)	(0.002)	(0.001)
	19297	410	9076	28783
Asset Index	9.433	13.536	12.044	10.683
	(0.073)	(0.167)	(0.112)	(0.058)
	20529	4053	10355	34937

Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The table reports the estimated average of each variable across the different subsamples, together with the corresponding standard error and the number of observations. Households doing only agriculture are those for which we can derive information on value added in agriculture, but not on profits from non-farm entrepreneurship. Households doing only entrepreneurship are those for which we can derive information on profits from non-farm entrepreneurship, but not on value added in agriculture. Households doing both are those for which we can derive information on both value added in agriculture and non-farm entrepreneurial profits.

Table 2: Selection Along the Extensive Margin

	Any Entrepreneurship				Any Farming			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	0.001		-0.000					
	(0.001)		(0.001)					
$P(VA_a/h_a)$		0.003**		0.007***				
		(0.001)		(0.001)				
$P(VA_n)$					-0.001		-0.001	
					(0.002)		(0.001)	
$P(VA_n/h_n)$						0.001		0.002
						(0.001)		(0.001)
Hours in Agriculture			-0.004***					
			(0.000)					
Hours in Entrepreneurship							-0.002***	
							(0.000)	
Land Size (ha)			-0.000	-0.000				
			(0.002)	(0.002)				
Fraction Rented			0.010	0.030*				
			(0.013)	(0.016)				
Total Hours			0.004***	0.001***			0.002***	0.000***
			(0.000)	(0.000)			(0.000)	(0.000)
Household Size			-0.002	0.000			0.018***	0.019***
			(0.002)	(0.002)			(0.002)	(0.002)
Females			0.014***	0.007**			0.003	0.005
			(0.003)	(0.003)			(0.003)	(0.004)
Asset Index – Ethiopia			0.002***	0.002***			0.002***	0.002***
			(0.000)	(0.001)			(0.001)	(0.001)
Asset Index – Malawi			0.004***	0.006***			-0.000	-0.001**
			(0.001)	(0.001)			(0.000)	(0.000)
Asset Index – Nigeria			0.002***	0.003***			0.001	0.001
			(0.001)	(0.001)			(0.001)	(0.001)
Asset Index – Uganda			-0.000	0.001			0.002**	0.001
			(0.001)	(0.001)			(0.001)	(0.001)
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	30931	22893	27420	21489	14377	11963	13958	11909
R^2	0.247	0.247	0.337	0.292	0.515	0.539	0.572	0.570

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable in columns 1 to 4 is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. The dependent variable in columns 5 to 8 is a dummy equal to 1 if we can derive information on value added in agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour.

Table 3: Selection Along the Intensive Margin

	h_a/h_n				h_n/h_a			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	0.028*		0.000					
	(0.015)		(0.018)					
$P(VA_a/h_a)$		-0.121***		-0.114***				
		(0.021)		(0.023)				
$P(VA_n)$					0.132***		0.120***	
					(0.035)		(0.034)	
$P(VA_n/h_n)$						-0.037		-0.046
						(0.029)		(0.030)
Land Size (ha)			0.010***	0.018***				
			(0.003)	(0.004)				
Fraction Rented			-0.183	-0.205				
			(0.241)	(0.354)				
Total Hours			0.006***	0.004***			0.002	0.000
			(0.001)	(0.001)			(0.002)	(0.002)
Household Size			0.018	0.040			-0.048	-0.031
			(0.018)	(0.025)			(0.038)	(0.043)
Females			0.006	0.006			0.037	0.012
			(0.040)	(0.054)			(0.061)	(0.074)
Asset Index – Ethiopia			-0.006	-0.013			-0.005	-0.008
			(0.010)	(0.019)			(0.007)	(0.008)
Asset Index – Malawi			-0.027**	-0.044			0.143*	0.219**
			(0.013)	(0.031)			(0.073)	(0.107)
Asset Index – Nigeria			-0.012***	-0.015*			0.023***	0.030***
			(0.004)	(0.008)			(0.009)	(0.009)
Asset Index – Uganda			-0.011*	-0.015**			0.077***	0.094***
			(0.006)	(0.006)			(0.028)	(0.028)
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8268	5703	7119	5237	6914	5703	6902	5692
R^2	0.336	0.351	0.346	0.359	0.274	0.265	0.285	0.280

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households for which we can derive information on both value added in agriculture and profits from non-farm entrepreneurship. The dependent variable in columns 1 to 4 is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. The dependent variable in columns 5 to 8 is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour.

Table 4: Activities Over Time

	Only Agriculture	Only Entrep.	Both	Full Sample
Wave 1	63.44%	10.88%	25.68%	100%
	7606	1304	3079	11989
Wave 2	61.37%	9.56%	29.07%	100%
	7228	1126	3424	11778
Wave 3	50.99%	15.35%	33.66%	100%
	4923	1482	3250	9655
Wave 4	51.64%	11.28%	37.07%	100%
	865	189	621	1675

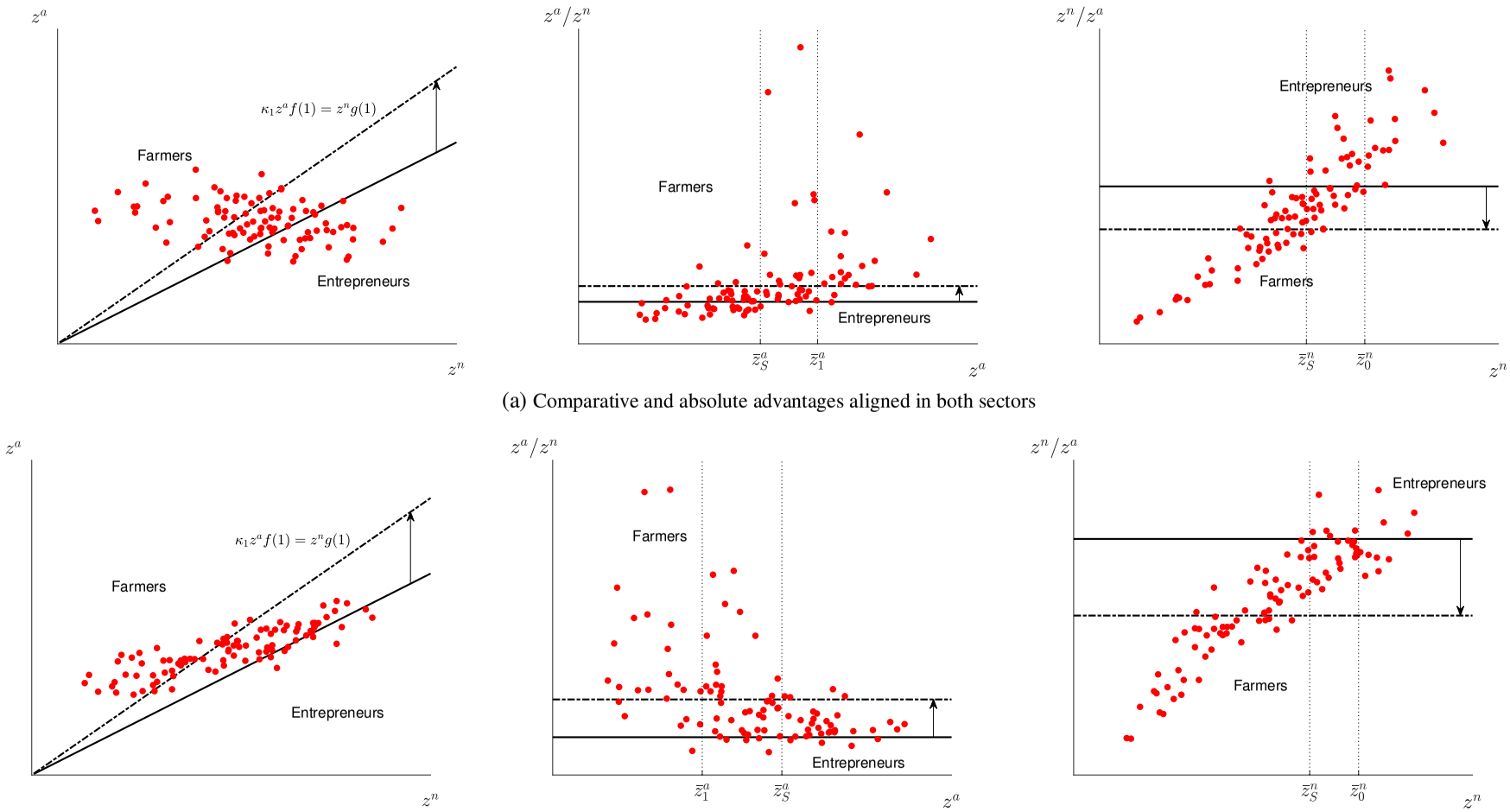
Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The table reports the relative and absolute number of households across the different subsamples over different waves. Households doing only agriculture are those for which we can derive information on value added in agriculture, but not on profits from non-farm entrepreneurship. Households doing only entrepreneurship are those for which we can derive information on profits from non-farm entrepreneurship, but not on value added in agriculture. Households doing both are those for which we can derive information on both value added in agriculture and non-farm entrepreneurial profits.

Table 5: Transitions To Entrepreneurship

	Any Entrepreneurship			
	(1)	(2)	(3)	(4)
Wave 2 $\times$ $Rank(VA_a)$	-0.007***		-0.007***	
	(0.002)		(0.002)	
Wave 3 $\times$ $Rank(VA_a)$	-0.008***		-0.009***	
	(0.003)		(0.003)	
Wave 2 $\times$ $Rank(VA_a/h_a)$		-0.009***		-0.007***
		(0.002)		(0.002)
Wave 3 $\times$ $Rank(VA_a/h_a)$		-0.012***		-0.010***
		(0.003)		(0.003)
Hours in Agriculture			-0.002***	
			(0.000)	
Land Size (ha)			-0.051***	-0.041**
			(0.015)	(0.020)
Fraction Rented			-0.003	-0.011
			(0.023)	(0.027)
Total Hours			0.002***	0.000***
			(0.000)	(0.000)
Household Size			0.006	0.011**
			(0.005)	(0.005)
Females			0.012*	0.002
			(0.006)	(0.007)
Household FE	Yes	Yes	Yes	Yes
Wave FE	Yes	Yes	n.a.	n.a.
Asset Index	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Observations	18721	14746	16509	13678
R^2	0.547	0.544	0.590	0.574

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households for which we cannot derive any information on profits from entrepreneurship in Wave 1, and observed again over time through Wave 3. $Rank(\cdot)$ is the within-village ranking of agricultural value added or agricultural value added per hour in Wave 1 among these households.

Figure 1: Standard Roy model: Choice of a Single Sector

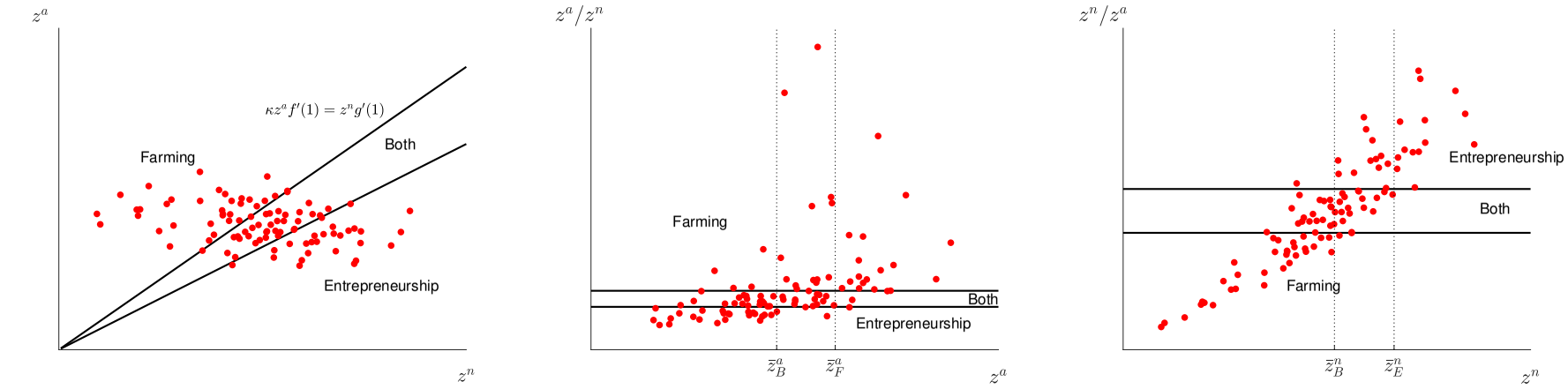


(a) Comparative and absolute advantages aligned in both sectors

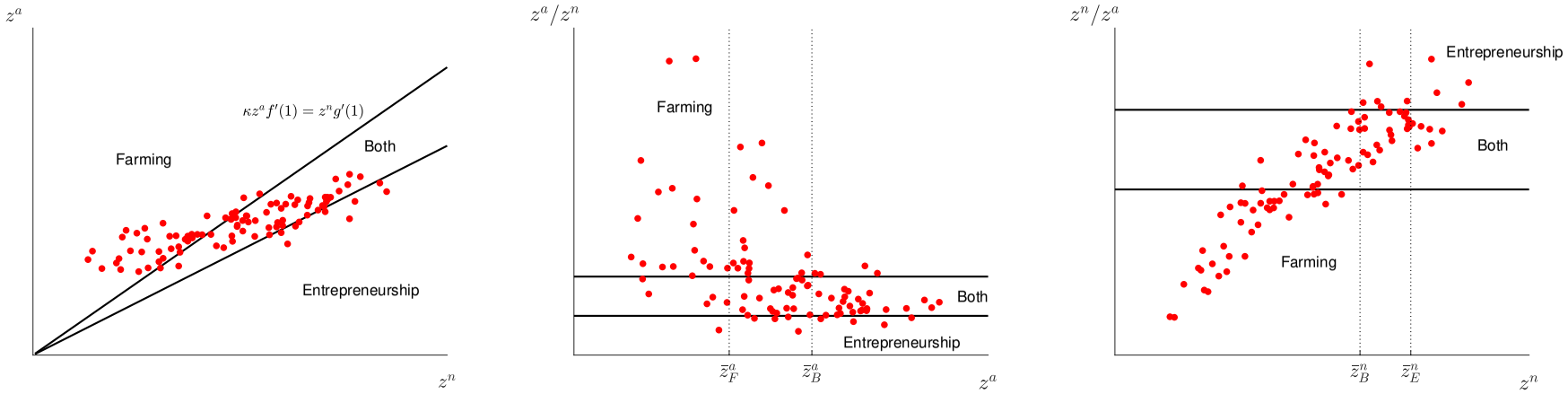
(b) Comparative and absolute advantages misaligned in agriculture

Notes. Simulated data. In both panels, the sample size is 100 and outliers are omitted. In both panels, $\mu_a = \mu_n = 1$, $\sigma_n = 1/3$, $\sigma_a = 1/6$. The correlation between z^a and z^n is -0.2 in panel (a) and 0.85 in panel (b).

Figure 2: Extended Roy model: Continuous Choice of Hours



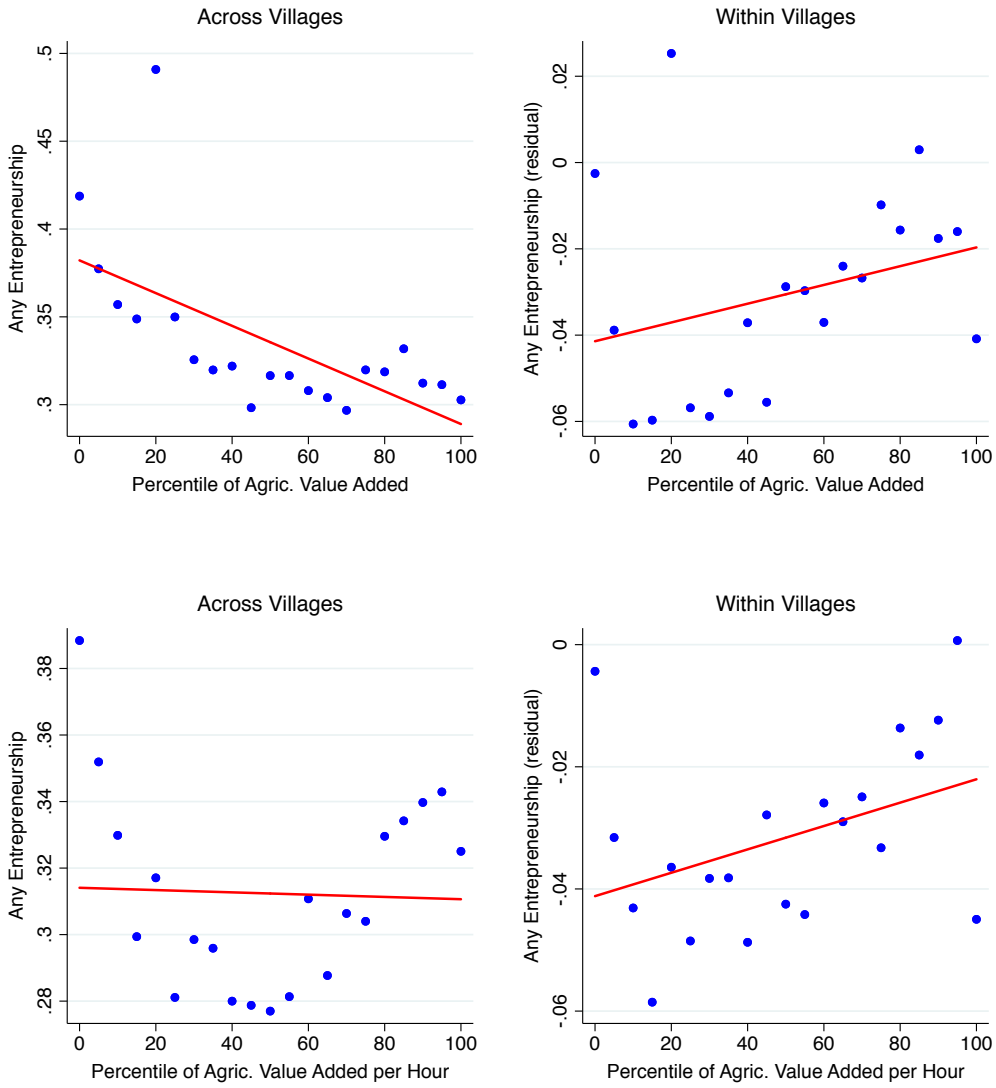
(a) Comparative and absolute advantages aligned in both sectors



(b) Comparative and absolute advantages misaligned in agriculture

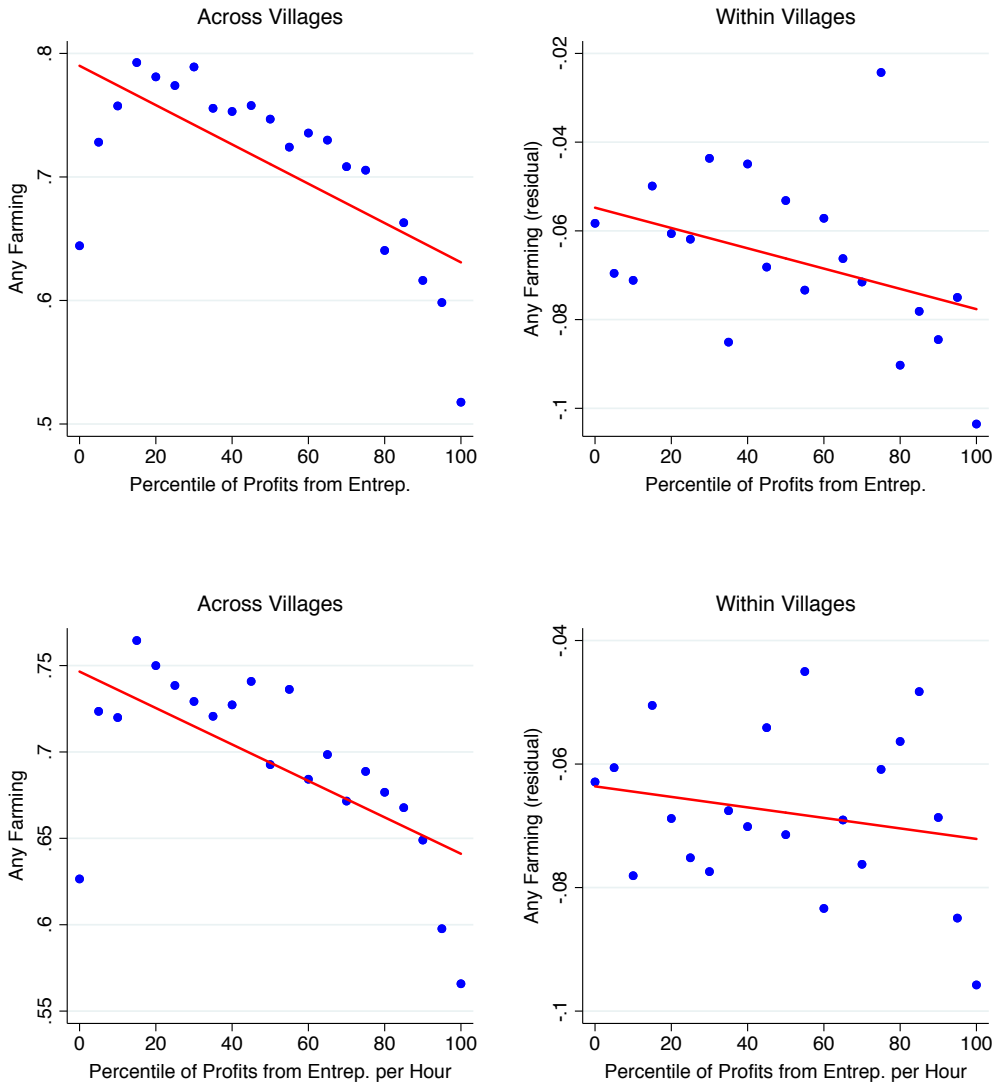
Notes. Same samples as in Figure 1.

Figure 3: Agricultural Value Added and Entrepreneurship



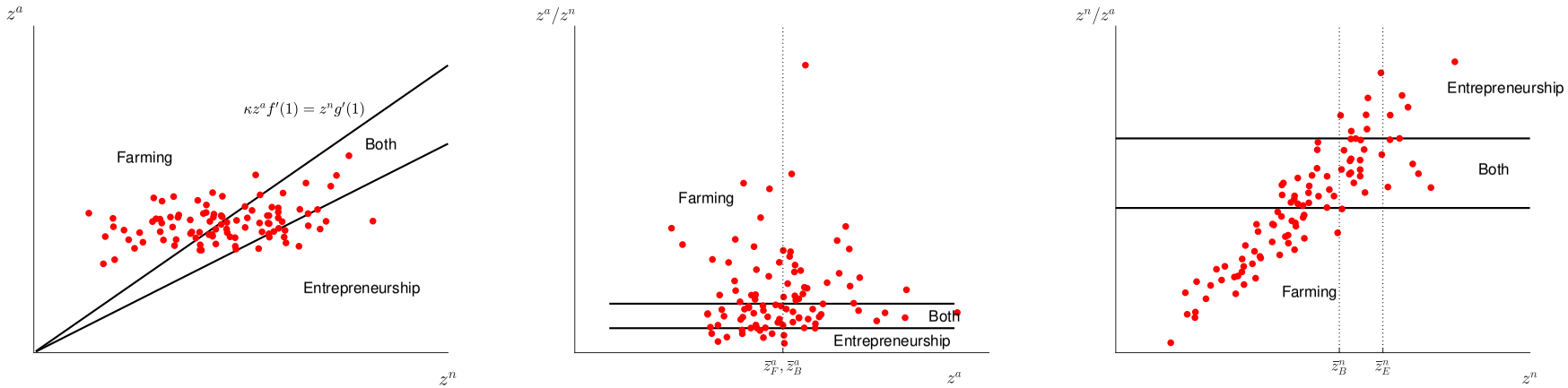
Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The top figures show the fraction of households involved in non-farm entrepreneurship per bin of 5 percentiles of the distribution of value added in agriculture as derived in each country and wave. The bottom figures show the same number per bin of 5 percentiles of the distribution of value added in agriculture per hour as derived in each country and wave. The right figures plot the averaged residuals of the probability of doing entrepreneurship after netting out location (enumeration area) fixed effects.

Figure 4: Profits from Entrepreneurship and Farming

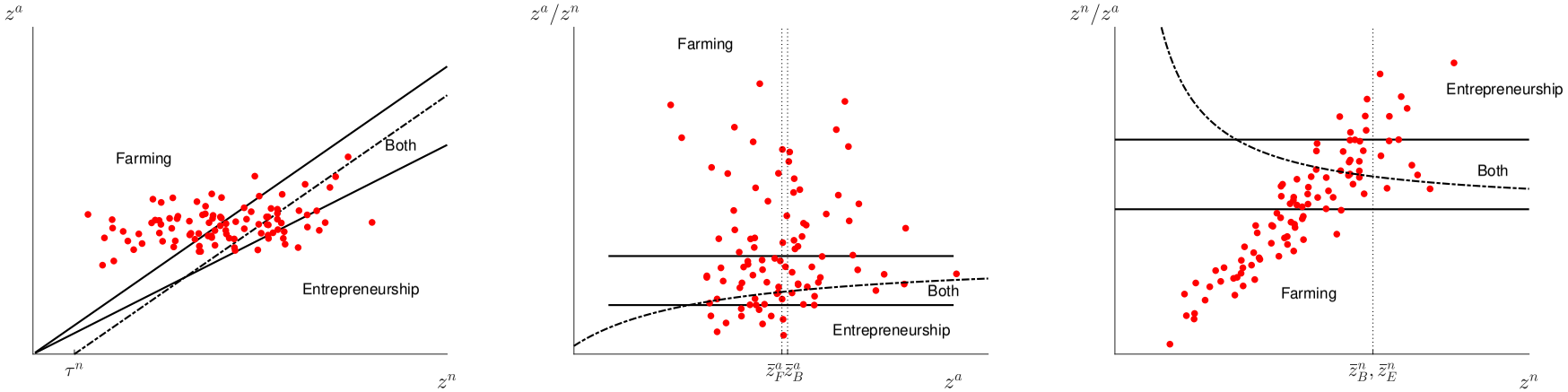


Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The top figures show the fraction of households involved in farming per bin of 5 percentiles of the distribution of profits from non-farm entrepreneurship as derived in each country and wave. The bottom figures show the same number per bin of 5 percentiles of the distribution of profits from non-farm entrepreneurship per hour as derived in each country and wave. The right figures plot the averaged residuals of the probability of doing farming after netting out location (enumeration area) fixed effects.

Figure 5: Extended Roy model: Effect of Sector-specific Fixed Costs



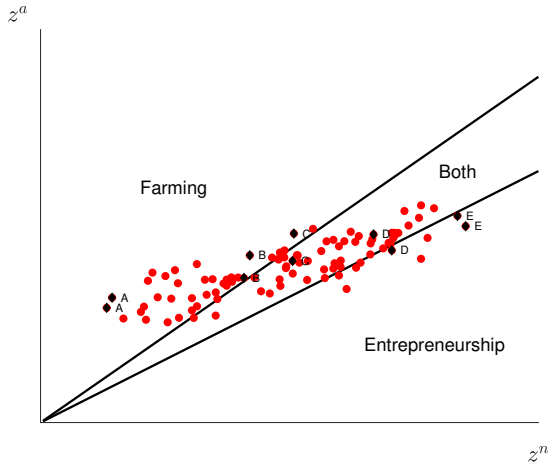
(a) Correlation of advantages with no fixed cost: unaligned in agriculture, aligned in non-agriculture



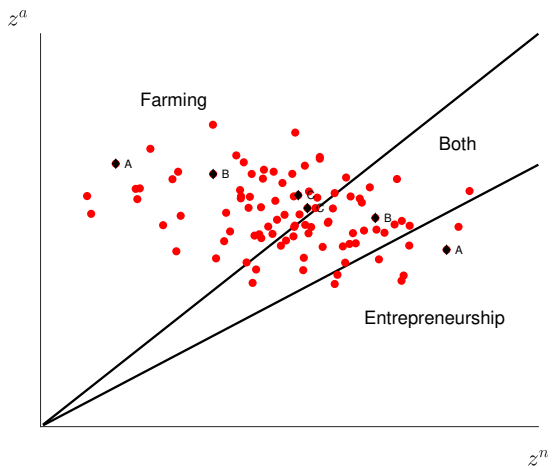
(b) Correlation of advantages with fixed cost in non-agriculture: misaligned in agriculture, unaligned in non-agriculture

Notes. Simulated data. Same sample in both panels. The sample size is 100 and outliers are omitted. In both panels, $\mu_a = \mu_n = 1$, $\sigma_n = 1/3$, $\sigma_a = 1/6$. The correlation between z^a and z^n is $(\sigma_a/\mu_a)/(\sigma_n/\mu_n) = 1/2$.

Figure 6: Choice of sector by households vs individuals



(a) Individual comparative and absolute advantages misaligned in agriculture



(b) Individual comparative and absolute advantages aligned in both sectors

Notes. Same samples as in Figure 1(a) (panel b) and Figure 1(b) (panel a).

Appendix – For Online Publication

A Additional Tables and Figures

Table A.1: Summary Statistics by Country

	Only Agriculture	Only Entrep.	Both	Full Sample
		Panel A. Ethiopia		
Observations	6946	738	2371	10055
	69%	7%	24%	100%
Household Size	5.168	3.954	5.450	5.145
	(0.027)	(0.084)	(0.046)	(0.023)
	6924	738	2371	10033
Hours in Agriculture	53.504	10.797	43.316	48.037
	(0.660)	(0.914)	(1.006)	(0.531)
	6697	688	2328	9713
Hours in Entrepreneurship	4.598	43.686	23.986	12.013
	(0.207)	(1.956)	(0.777)	(0.298)
	6697	688	2328	9713
		Panel B. Malawi		
Observations	3936	27	1420	5383
	73%	1%	26%	100%
Household Size	4.280	4.111	4.477	4.331
	(0.034)	(0.561)	(0.055)	(0.029)
	3934	27	1420	5381
Hours in Agriculture	21.515	2.333	12.337	18.997
	(0.516)	(1.539)	(0.637)	(0.417)
	3934	27	1420	5381
Hours in Entrepreneurship	1.725	49.741	30.761	9.628
	(0.151)	(7.436)	(0.916)	(0.322)
	3934	27	1420	5381
		Panel C. Nigeria		
Observations	5605	2474	3492	11571
	48%	21%	30%	100%
Household Size	5.535	4.894	6.536	5.700
	(0.042)	(0.054)	(0.054)	(0.029)
	5603	2474	3492	11569
Hours in Agriculture	63.571	2.446	45.121	44.794
	(0.925)	(0.269)	(0.954)	(0.579)
	5329	2389	3396	11114
Hours in Entrepreneurship	18.936	71.702	57.130	41.949
	(0.547)	(0.937)	(0.766)	(0.459)
	5329	2389	3396	11114
		Panel D. Uganda		
Observations	4135	862	3091	8088
	51%	11%	38%	100%
Household Size	5.007	4.444	5.596	5.173
	(0.040)	(0.090)	(0.047)	(0.029)
	4076	854	3076	8006
Hours in Agriculture	40.307	3.568	33.155	33.554
	(0.632)	(0.633)	(0.652)	(0.431)
	3891	836	3030	7757
Hours in Entrepreneurship	58.997	90.951	81.372	71.181
	(0.775)	(2.286)	(1.044)	(0.631)
	3891	836	3030	7757

Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The table reports the estimated average of each variable across the different subsamples, together with the corresponding standard error and the number of observations. Households doing only agriculture are those for which we can derive information on value added in agriculture, but not on profits from non-farm entrepreneurship. Households doing only entrepreneurship are those for which we can derive information on profits from non-farm entrepreneurship, but not on value added in agriculture. Households doing both are those for which we can derive information on both value added in agriculture and non-farm entrepreneurial profits.

Table A.2: Agricultural Value Added and Entrepreneurship
Estimates Without and With Village FE

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	-0.009***		-0.006***		0.001		-0.000	
	(0.001)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		0.001		0.003**		0.003**		0.007***
		(0.001)		(0.001)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	30997	22977	27486	21575	30931	22892	27419	21488
R^2	0.003	0.000	0.179	0.080	0.247	0.247	0.337	0.293

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.3: Entrepreneurial Profits and Farming
Estimates Without and With Village FE

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.017***		-0.010***		-0.001		-0.001	
	(0.002)		(0.002)		(0.002)		(0.001)	
$P(VA_n/h_n)$		-0.012***		-0.008***		0.001		0.002
		(0.002)		(0.002)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	14477	12095	14058	12041	14377	11963	13958	11909
R^2	0.012	0.005	0.270	0.154	0.515	0.539	0.572	0.570

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.4: Agricultural Value Added and Entrepreneurship
Alternative Definition Based on Hours Worked

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	-0.006***		-0.003***		-0.000		-0.001	
	(0.002)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		0.001		0.005***		0.001		0.007***
		(0.002)		(0.001)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	30026	22977	27486	21575	29960	22892	27419	21488
R^2	0.001	0.000	0.559	0.452	0.469	0.533	0.622	0.575

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if any member of the household reports any hour worked in household business. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.5: Entrepreneurial Profits and Farming
Alternative Definition Based on Hours Worked

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.022***		-0.010***		-0.005***		-0.004***	
	(0.002)		(0.001)		(0.002)		(0.001)	
$P(VA_n/h_n)$		-0.017***		-0.008***		-0.006***		-0.000
		(0.002)		(0.002)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	14116	12095	14058	12041	14016	11963	13958	11909
R^2	0.017	0.009	0.410	0.152	0.447	0.466	0.597	0.518

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if any member of the household reports any hour worked in household farm. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.6: Agricultural Value Added and Entrepreneurship
Stricter Definition Based on Hours Worked

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	-0.011***		-0.006***		-0.001		-0.000	
	(0.001)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		0.002		0.005***		0.004***		0.008***
		(0.001)		(0.001)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	30997	22977	27486	21575	30931	22892	27419	21488
R^2	0.005	0.000	0.260	0.131	0.262	0.262	0.380	0.304

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship and the household as a whole reports that at least 15% of total hours worked are dedicated to the household business. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.7: Entrepreneurial Profits and Farming
Stricter Definition Based on Hours Worked

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.026***		-0.009***		-0.009***		-0.004***	
	(0.002)		(0.001)		(0.002)		(0.001)	
$P(VA_n/h_n)$		-0.014***		-0.005***		-0.003*		0.002
		(0.002)		(0.002)		(0.002)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	14477	12095	14058	12041	14377	11963	13958	11909
R^2	0.022	0.006	0.407	0.123	0.421	0.442	0.574	0.486

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture and the household as a whole reports that at least 15% of total hours worked are dedicated to the household farm. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.8: Agricultural Value Added and Entrepreneurship
Excluding Livestock-related Activities

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	-0.006***		-0.006***		0.002**		0.000	
	(0.001)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		0.000		0.002		0.002**		0.006***
		(0.002)		(0.002)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	30615	22967	27486	21575	30549	22880	27419	21488
R^2	0.001	0.000	0.179	0.080	0.242	0.247	0.337	0.292

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture (excluding livestock-related activities) as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.9: Entrepreneurial Profits and Farming
Excluding Livestock-related Activities

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.019***		-0.011***		-0.002		-0.002	
	(0.002)		(0.002)		(0.002)		(0.001)	
$P(VA_n/h_n)$		-0.013***		-0.009***		0.000		0.001
		(0.002)		(0.002)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	14477	12095	14058	12041	14377	11963	13958	11909
R^2	0.014	0.006	0.263	0.143	0.540	0.565	0.586	0.586

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture (excluding livestock-related activities). $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.10: Agricultural Value Added and Entrepreneurship
Non-specialized Households

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	0.003***		-0.000		0.006***		0.002*	
	(0.001)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		-0.000		0.003***		0.002		0.006***
		(0.001)		(0.001)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	30997	22977	27486	21575	30931	22892	27419	21488
R^2	0.001	0.000	0.164	0.136	0.187	0.242	0.267	0.282

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship and at least one household member reports hours worked in both the household farm and the household business. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.11: Entrepreneurial Profits and Farming
Non-specialized Households

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.012***		-0.008***		0.001		-0.001	
	(0.002)		(0.001)		(0.001)		(0.001)	
$P(VA_n/h_n)$		-0.013***		-0.005***		-0.002		0.004**
		(0.002)		(0.002)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	14477	12095	14058	12041	14377	11963	13958	11909
R^2	0.005	0.006	0.283	0.163	0.396	0.440	0.483	0.483

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture and at least one household member reports hours worked in both the household farm and the household business. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.12: Agricultural Value Added and Entrepreneurship
Restricted Sample of Households with Any Hours Worked for Others

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	-0.006**		-0.004**		0.002		0.003	
	(0.002)		(0.002)		(0.002)		(0.002)	
$P(VA_a/h_a)$		0.005**		0.001		0.005*		0.006**
		(0.003)		(0.002)		(0.002)		(0.003)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	8179	5374	7000	4938	8010	5136	6824	4694
R^2	0.001	0.001	0.191	0.139	0.342	0.382	0.418	0.421

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report any hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.13: Entrepreneurial Profits and Farming
Restricted Sample of Households with Any Hours Worked for Others

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.016***		-0.011***		-0.001		-0.001	
	(0.003)		(0.002)		(0.002)		(0.002)	
$P(VA_n/h_n)$		-0.014***		-0.012***		-0.001		-0.000
		(0.003)		(0.003)		(0.002)		(0.002)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	5019	4474	4926	4465	4737	4191	4642	4182
R^2	0.010	0.007	0.284	0.161	0.608	0.621	0.645	0.641

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report any hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.14: Agricultural Value Added and Entrepreneurship
Restricted Sample of Households with No Hours Worked for Others

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	-0.009***		-0.007***		0.001		-0.001	
	(0.002)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		-0.000		0.002		0.001		0.005***
		(0.002)		(0.002)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	22818	17603	20486	16637	22740	17492	20406	16530
R^2	0.003	0.000	0.183	0.074	0.274	0.266	0.360	0.312

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report no hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.15: Entrepreneurial Profits and Farming
Restricted Sample of Households with No Hours Worked for Others

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	-0.017***		-0.008***		-0.002		-0.002	
	(0.003)		(0.002)		(0.002)		(0.002)	
$P(VA_n/h_n)$		-0.009***		-0.003		0.001		0.003*
		(0.003)		(0.002)		(0.002)		(0.002)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	9458	7621	9132	7576	9277	7420	8948	7375
R^2	0.012	0.004	0.286	0.196	0.503	0.535	0.569	0.572

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report no hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.16: Agricultural Value Added and Entrepreneurship
No Percentile Transformation

	Any Entrepreneurship							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VA_a	0.012***		-0.014***		0.004		-0.002	
	(0.004)		(0.004)		(0.003)		(0.003)	
VA_a/h_a		0.031***		0.007*		0.011***		0.013***
		(0.004)		(0.004)		(0.004)		(0.004)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	30997	22977	27486	21575	30931	22892	27419	21488
R^2	0.001	0.004	0.178	0.080	0.247	0.247	0.337	0.292

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report no hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. VA_a is value added in agriculture and VA_a/h_a is value added per hour. A cube root transformation $\text{sign}\,(x) \times |x|^{1/3}$ and standard deviation normalization is applied to both variables. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.17: Entrepreneurial Profits and Farming
No Percentile Transformation

	Any Farming							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VA_n	-0.018***		-0.025***		0.002		-0.002	
	(0.007)		(0.006)		(0.004)		(0.004)	
VA_n/h_n		0.002		-0.021***		0.005		0.004
		(0.007)		(0.006)		(0.004)		(0.004)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	14477	12095	14058	12041	14377	11963	13958	11909
R^2	0.002	0.000	0.268	0.154	0.515	0.540	0.572	0.570

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report no hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture. VA_n is profits from non-farm entrepreneurship and VA_n/h_n is profits per hour. A cube root transformation $\text{sign}\,(x) \times |x|^{1/3}$ and standard deviation normalization is applied to both variables. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Table A.18: Estimated Agricultural Advantage and Entrepreneurship

	Any Entrepreneurship			
	(1)	(2)	(3)	(4)
$P(\hat{z}^a)$	0.006***	0.004***	0.005***	0.003**
	(0.002)	(0.002)	(0.001)	(0.002)
Controls	No	Yes	No	Yes
Country-Wave FE	No	Yes	No	Yes
Village FE	No	No	Yes	Yes
Observations	12081	11588	12081	11564
R^2	0.002	0.066	0.295	0.319

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report no hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on profits from non-farm entrepreneurship. $P(\hat{z}^a)$ is the percentile (divided by 10) the household belongs to in the distribution of estimated agricultural productivity as derived in each country and wave. $\hat{z}^a$ is estimated as the residual from a regression of the log of value of agricultural production over the log of hours worked in that sector, together with the full set of location and wave fixed effects. Control variables include: cultivated area, fraction of land that is rented, country-specific asset index.

Table A.19: Estimated Entrepreneurial Advantage and Farming

	Any Farming			
	(1)	(2)	(3)	(4)
$P(\hat{z}^n)$	-0.001	-0.000	-0.002	-0.003**
	(0.001)	(0.001)	(0.001)	(0.001)
Controls	No	Yes	No	Yes
Country-Wave FE	No	Yes	No	Yes
Village FE	No	No	Yes	Yes
Observations	12791	12738	12791	12738
R^2	0.000	0.064	0.518	0.532

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report no hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is a dummy equal to 1 if we can derive information on value added in agriculture. $P(\hat{z}^a)$ is the percentile (divided by 10) the household belongs to in the distribution of estimated productivity in non-farm entrepreneurship as derived in each country and wave. $\hat{z}^n$ is estimated as the residual from a regression of the log of value of sales associated to the household-run enterprise over the log of hours worked in non-farm entrepreneurship, together with the full set of location and wave fixed effects. Control variables include the country-specific asset index.

Table A.20: Agricultural Value Added and Time Allocation
Alternative Definition Based on Hours Worked

	h_a/h_n			
	(1)	(2)	(3)	(4)
$P(VA_a)$	0.010		-0.006	
	(0.022)		(0.025)	
$P(VA_a/h_a)$		-0.122***		-0.115***
		(0.021)		(0.023)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	5702	5702	5236	5236
R^2	0.348	0.354	0.357	0.362

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business. The dependent variable is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.21: Entrepreneurial Profits and Time Allocation
Alternative Definition Based on Hours Worked

	h_n/h_a			
	(1)	(2)	(3)	(4)
$P(VA_n)$	0.149***		0.150***	
	(0.041)		(0.044)	
$P(VA_n/h_n)$		-0.037		-0.047
		(0.029)		(0.032)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	5702	5702	5236	5236
R^2	0.269	0.265	0.261	0.257

Notes. * p-value$<$ 0.1; ** p-value$<$0.05; *** p-value$<$0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business. The dependent variable is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, country-specific asset index.

Table A.22: Agricultural Value Added and Time Allocation
Stricter Definition Based on Hours Worked

	h_a/h_n			
	(1)	(2)	(3)	(4)
$P(VA_a)$	0.011**		0.010*	
	(0.005)		(0.005)	
$P(VA_a/h_a)$		-0.041***		-0.040***
		(0.005)		(0.006)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4998	4998	4583	4583
R^2	0.389	0.398	0.401	0.409

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and devote at least 15% of their total hours worked to each activity. The dependent variable is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.23: Entrepreneurial Profits and Time Allocation
Stricter Definition Based on Hours Worked

	h_n/h_a			
	(1)	(2)	(3)	(4)
$P(VA_n)$	0.030***		0.033***	
	(0.007)		(0.008)	
$P(VA_n/h_n)$		-0.024***		-0.029***
		(0.007)		(0.007)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4998	4998	4583	4583
R^2	0.311	0.309	0.334	0.333

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and devote at least 15% of their total hours worked to each activity. The dependent variable is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, country-specific asset index.

Table A.24: Agricultural Value Added and Time Allocation
Excluding Livestock-related Activities

	h_a/h_n			
	(1)	(2)	(3)	(4)
$P(VA_a)$	0.007		-0.012	
	(0.016)		(0.019)	
$P(VA_a/h_a)$		-0.126***		-0.118***
		(0.022)		(0.023)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4998	4998	4583	4583
R^2	0.389	0.398	0.401	0.409

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households for which we can derive information on both value added in agriculture (excluding livestock-related activities) and profits from non-farm entrepreneurship. The dependent variable is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture (still excluding livestock-related activities) as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.25: Entrepreneurial Profits and Time Allocation
Excluding Livestock-related Activities

	h_n/h_a			
	(1)	(2)	(3)	(4)
$P(VA_n)$	0.147***		0.142***	
	(0.036)		(0.037)	
$P(VA_n/h_n)$		-0.024		-0.036
		(0.029)		(0.032)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4998	4998	4583	4583
R^2	0.311	0.309	0.334	0.333

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households for which we can derive information on both value added in agriculture (excluding livestock-related activities) and profits from non-farm entrepreneurship. The dependent variable is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, country-specific asset index.

Table A.26: Agricultural Value Added and Time Allocation
Non-specialized Households

	h_a/h_n			
	(1)	(2)	(3)	(4)
$P(VA_a)$	0.015		-0.000	
	(0.024)		(0.027)	
$P(VA_a/h_a)$		-0.106***		-0.102***
		(0.021)		(0.025)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4608	4608	4209	4209
R^2	0.362	0.366	0.370	0.373

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and having at least one household member reporting hours worked in both. The dependent variable is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.27: Entrepreneurial Profits and Time Allocation
Non-specialized Households

	h_n/h_a			
	(1)	(2)	(3)	(4)
$P(VA_n)$	0.160***		0.161***	
	(0.051)		(0.056)	
$P(VA_n/h_n)$		-0.031		-0.041
		(0.034)		(0.040)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4608	4608	4209	4209
R^2	0.255	0.251	0.266	0.263

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and having at least one household member reporting hours worked in both. The dependent variable is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, country-specific asset index.

Table A.28: Agricultural Value Added and Time Allocation
Restricted Sample of Households with Any Hours Worked for Others

	h_a/h_n			
	(1)	(2)	(3)	(4)
$P(VA_a)$	0.007		-0.005	
	(0.009)		(0.012)	
$P(VA_a/h_a)$		-0.051***		-0.056**
		(0.019)		(0.022)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	2599	1506	2078	1331
R^2	0.516	0.543	0.534	0.564

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and also report any hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.29: Entrepreneurial Profits and Time Allocation
Restricted Sample of Households with Any Hours Worked for Others

	h_n/h_a			
	(1)	(2)	(3)	(4)
$P(VA_n)$	0.074		0.127*	
	(0.069)		(0.070)	
$P(VA_n/h_n)$		0.007		0.002
		(0.073)		(0.081)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	1728	1506	1539	1331
R^2	0.419	0.401	0.444	0.427

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and also report any hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, country-specific asset index.

Table A.30: Agricultural Value Added and Time Allocation
Restricted Sample of Households with No Hours Worked for Others

	h_a/h_n			
	(1)	(2)	(3)	(4)
$P(VA_a)$	0.047**		0.020	
	(0.020)		(0.022)	
$P(VA_a/h_a)$		-0.121***		-0.100***
		(0.024)		(0.025)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	5303	3846	4653	3562
R^2	0.348	0.373	0.370	0.386

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and do not report any hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is the ratio of total hours worked by the household in agriculture vs. non-farm entrepreneurship. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.31: Entrepreneurial Profits and Time Allocation
Restricted Sample of Households with No Hours Worked for Others

	h_n/h_a			
	(1)	(2)	(3)	(4)
$P(VA_n)$	0.091**		0.094**	
	(0.039)		(0.042)	
$P(VA_n/h_n)$		-0.074**		-0.086***
		(0.030)		(0.031)
Controls	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes
Village FE	Yes	Yes	Yes	Yes
Observations	4806	3846	4501	3562
R^2	0.280	0.301	0.253	0.278

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. Sample is restricted to those households that report positive hours worked in both the household farm and the household business and do not report any hours worked outside the household (paid or unpaid, temporary or not, apprenticeship). The dependent variable is the ratio of total hours worked by the household in non-farm entrepreneurship vs. agriculture. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, country-specific asset index.

Table A.32: Activities Over Time by Country

	Only Agriculture	Only Entrep.	Both	Full Sample
		Panel A. Ethiopia		
Wave 1	66.92%	12.12%	20.96%	100%
	2142	388	671	3201
Wave 2	68.96%	6.5%	24.54%	100%
	2346	221	835	3402
Wave 3	71.21%	3.74%	25.06%	100%
	2458	129	865	3452
		Panel B. Malawi		
Wave 1	76.86%	.39%	22.75%	100%
	2176	11	644	2831
Wave 2	68.97%	.63%	30.41%	100%
	1760	16	776	2552
		Panel C. Nigeria		
Wave 1	59.44%	17.92%	22.64%	100%
	2213	667	843	3723
Wave 2	55.03%	18.27%	26.69%	100%
	2066	686	1002	3754
Wave 3	32.39%	27.38%	40.23%	100%
	1326	1121	1647	4094
		Panel D. Uganda		
Wave 1	48.12%	10.65%	41.23%	100%
	1075	238	921	2234
Wave 2	51.01%	9.81%	39.18%	100%
	1056	203	811	2070
Wave 3	54.01%	11%	34.99%	100%
	1139	232	738	2109
Wave 4	51.64%	11.28%	37.07%	100%
	865	189	621	1675

Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The table reports the relative and absolute number of households across the different subsamples over different waves per country. Households doing only agriculture are those for which we can derive information on value added in agriculture, but not on profits from non-farm entrepreneurship. Households doing only entrepreneurship are those for which we can derive information on profits from non-farm entrepreneurship, but not on value added in agriculture. Households doing both are those for which we can derive information on both value added in agriculture and non-farm entrepreneurial profits.

Table A.33: Transition Matrices

Wave 1 to 2	Only Agriculture	Both	Only Entrep.
Only Agriculture	52.65%	10.96%	0.85%
Both	9.1%	16.66%	1.04%
Only Entrepreneurship	0.94%	2.5%	5.3%

Wave 2 to 3	Only Agriculture	Both	Only Entrep.
Only Agriculture	44.62%	12.29%	1.43%
Both	6.67%	20.15%	2.04%
Only Entrepreneurship	0.82%	2.47%	9.52%

Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The table reports the relative number of households across the different subsamples and their transitions from Wave 1 (row) to 2 (column) and from Wave 2 to 3. Households doing only agriculture are those for which we can derive information on value added in agriculture, but not on profits from non-farm entrepreneurship. Households doing only entrepreneurship are those for which we can derive information on profits from non-farm entrepreneurship, but not on value added in agriculture. Households doing both are those for which we can derive information on both value added in agriculture and non-farm entrepreneurial profits.

Table A.34: Transitions To Entrepreneurship by Country

	Entrepreneurship Dummy					
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A. Ethiopia			
Wave 2 × $Rank(VA_a)$	-0.005**	-0.005**	-0.004*			
	(0.002)	(0.002)	(0.002)			
Wave 3 × $Rank(VA_a)$	-0.006**	-0.006**	-0.005**			
	(0.003)	(0.003)	(0.003)			
Wave 2 × $Rank(VA_a/h_a)$				-0.004*	-0.004*	-0.003
				(0.002)	(0.002)	(0.003)
Wave 3 × $Rank(VA_a/h_a)$				-0.006**	-0.006**	-0.006**
				(0.003)	(0.003)	(0.003)
Observations	6062	6062	5923	5346	5346	5237
R^2	0.524	0.524	0.539	0.517	0.517	0.533
			Panel B. Malawi			
Wave 2 × $Rank(VA_a)$	-0.010***	-0.010***	-0.010***			
	(0.002)	(0.002)	(0.002)			
Wave 2 × $Rank(VA_a/h_a)$				-0.011***	-0.011***	-0.011***
				(0.003)	(0.003)	(0.003)
Observations	3692	3692	3512	2362	2362	2298
R^2	0.561	0.561	0.558	0.556	0.556	0.555
			Panel C. Nigeria			
Wave 2 × $Rank(VA_a)$	-0.008**	-0.008**	-0.004			
	(0.004)	(0.004)	(0.004)			
Wave 3 × $Rank(VA_a)$	-0.014***	-0.014***	-0.008*			
	(0.004)	(0.004)	(0.005)			
Wave 2 × $Rank(VA_a/h_a)$				-0.010**	-0.010**	-0.007
				(0.004)	(0.004)	(0.005)
Wave 3 × $Rank(VA_a/h_a)$				-0.017***	-0.017***	-0.013**
				(0.005)	(0.005)	(0.006)
Observations	6058	6058	5191	4653	4653	4060
R^2	0.582	0.582	0.584	0.582	0.582	0.583
			Panel D. Uganda			
Wave 2 × $Rank(VA_a)$	-0.017**	-0.017**	-0.018*			
	(0.008)	(0.008)	(0.009)			
Wave 3 × $Rank(VA_a)$	-0.017**	-0.017**	-0.018**			
	(0.008)	(0.008)	(0.008)			
Wave 4 × $Rank(VA_a)$	-0.017**	-0.017**	-0.018**			
	(0.009)	(0.009)	(0.009)			
Wave 2 × $Rank(VA_a/h_a)$				-0.010	-0.010	-0.012
				(0.009)	(0.009)	(0.011)
Wave 3 × $Rank(VA_a/h_a)$				-0.019**	-0.019**	-0.021**
				(0.008)	(0.008)	(0.009)
Wave 4 × $Rank(VA_a/h_a)$				-0.013	-0.013	-0.017
				(0.010)	(0.010)	(0.010)
Observations	3547	3547	3036	2907	2907	2571
R^2	0.481	0.481	0.493	0.474	0.474	0.495
Household FE	Yes	Yes	Yes	Yes	Yes	Yes
Wave FE	Yes	n.a.	n.a.	Yes	n.a.	n.a.
Country-Wave FE	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes

Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Sample is restricted to those households for which we cannot derive any information on profits from entrepreneurship in Wave 1, and observed again over time through Wave 3. $Rank(\cdot)$ is the within-village ranking of agricultural value added or agricultural value added per hour in Wave 1 among these households. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (column 3), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.35: Agricultural Value Added and Migration

	Household Member Moved Out							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_a)$	0.006***		0.002*		0.005***		0.001	
	(0.002)		(0.001)		(0.001)		(0.001)	
$P(VA_a/h_a)$		0.004**		0.002		0.003**		0.001
		(0.002)		(0.001)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	20229	15201	18096	14305	20153	15098	18026	14207
R^2	0.002	0.001	0.333	0.331	0.376	0.388	0.412	0.420

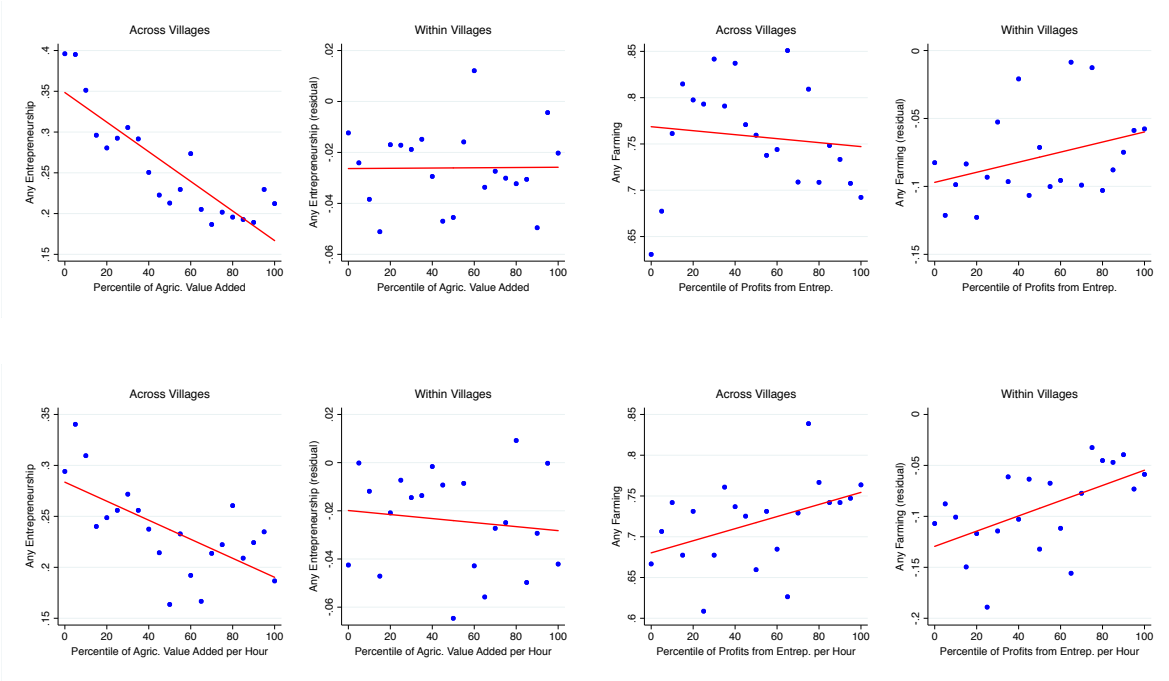
Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each but the first wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if any household member moved out of the household since the last interview. $P(VA_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added in agriculture as derived in each country and wave. $P(VA_a/h_a)$ is the percentile (divided by 10) the household belongs to in the distribution of value added per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in agriculture (columns 3 and 7 only), total cultivated area, fraction of land that is rented, country-specific asset index.

Table A.36: Entrepreneurial Profits and Migration

	Household Member Moved Out							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$P(VA_n)$	0.000		-0.000		0.000		-0.001	
	(0.002)		(0.001)		(0.001)		(0.001)	
$P(VA_n/h_n)$		-0.000		0.000		-0.001		-0.001
		(0.001)		(0.001)		(0.001)		(0.001)
Controls	No	No	Yes	Yes	No	No	Yes	Yes
Country-Wave FE	No	No	Yes	Yes	No	No	Yes	Yes
Village FE	No	No	No	No	Yes	Yes	Yes	Yes
Observations	10071	8430	9776	8378	9929	8253	9631	8201
R^2	0.000	0.000	0.332	0.303	0.432	0.422	0.458	0.445

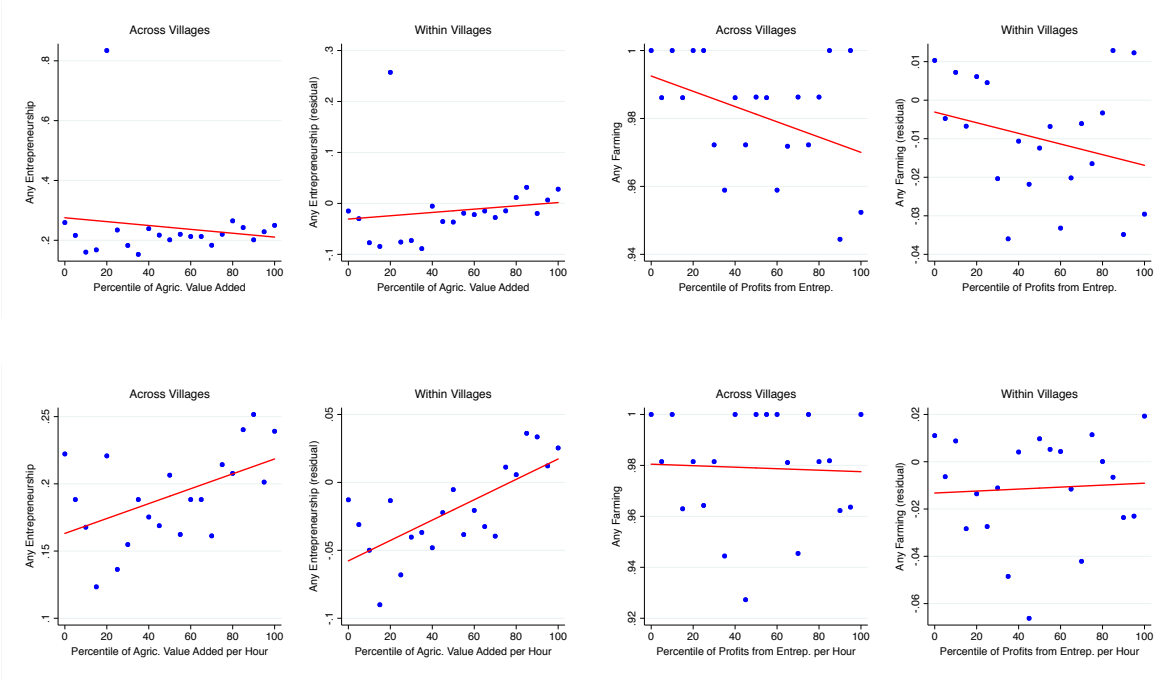
Notes. * p-value< 0.1; ** p-value<0.05; *** p-value<0.01. Standard errors in parenthesis, clustered at the location (enumeration area) level. Unit of observation is the household surveyed in each but the first wave of the LSMS-ISA panel dataset for Ethiopia, Malawi, Nigeria, and Uganda. The dependent variable is a dummy equal to 1 if any household member moved out of the household since the last interview. $P(VA_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship as derived in each country and wave. $P(VA_n/h_n)$ is the percentile (divided by 10) the household belongs to in the distribution of profits from non-farm entrepreneurship per hour. Control variables include: total number of household members, total number of female household members, total number of hours worked by all household members, total number of hours in non-farm entrepreneurship (columns 3 and 7 only), country-specific asset index.

Figure A.1: Figures by Country - Ethiopia



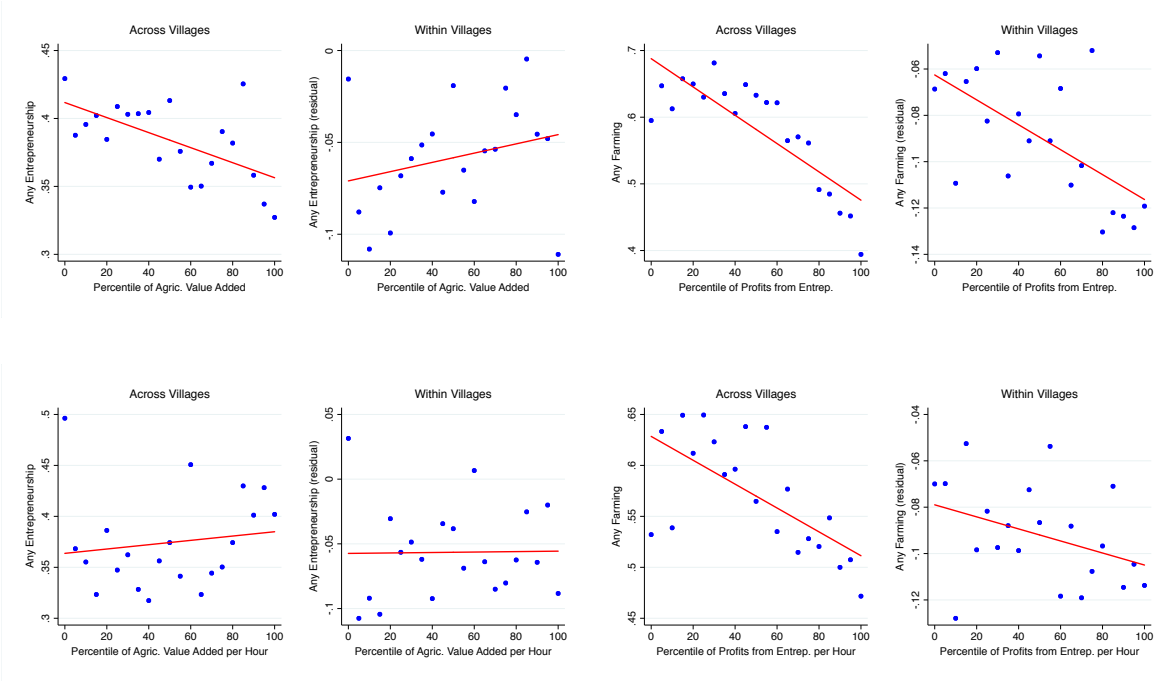
Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Ethiopia. The top figures show the fraction of households involved in non-farm entrepreneurship (left) or farming (right) per bin of 5 percentiles of the distribution of value added in agriculture (left) or profits from non-farm entrepreneurship (right). The bottom figures show the same number per bin of 5 percentiles of the distribution of value added in agriculture per hour (left) or profits from non-farm entrepreneurship per hour (right).

Figure A.2: Figures by Country - Malawi



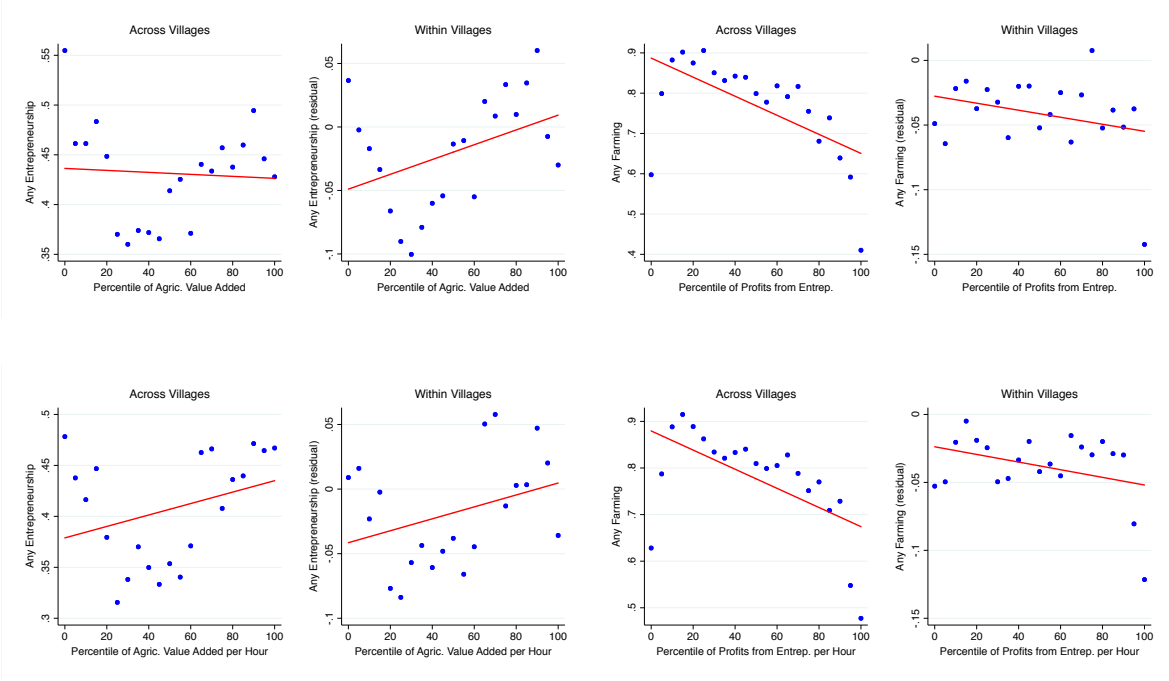
Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Malawi. The top figures show the fraction of households involved in non-farm entrepreneurship (left) or farming (right) per bin of 5 percentiles of the distribution of value added in agriculture (left) or profits from non-farm entrepreneurship (right). The bottom figures show the same number per bin of 5 percentiles of the distribution of value added in agriculture per hour (left) or profits from non-farm entrepreneurship per hour (right).

Figure A.3: Figures by Country - Nigeria



Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Nigeria. The top figures show the fraction of households involved in non-farm entrepreneurship (left) or farming (right) per bin of 5 percentiles of the distribution of value added in agriculture (left) or profits from non-farm entrepreneurship (right). The bottom figures show the same number per bin of 5 percentiles of the distribution of value added in agriculture per hour (left) or profits from non-farm entrepreneurship per hour (right).

Figure A.4: Figures by Country - Uganda



Notes. Unit of observation is the household surveyed in each wave of the LSMS-ISA panel dataset for Uganda. The top figures show the fraction of households involved in non-farm entrepreneurship (left) or farming (right) per bin of 5 percentiles of the distribution of value added in agriculture (left) or profits from non-farm entrepreneurship (right). The bottom figures show the same number per bin of 5 percentiles of the distribution of value added in agriculture per hour (left) or profits from non-farm entrepreneurship per hour (right).

B Proofs

B.1 Proof of Proposition 1

We start from

$$\rho\left(z_i^a/z_i^n, z_i^a\right) = \frac{Cov\left(z_i^a/z_i^n, z_i^a\right)}{\sigma_{z_i^a/z_i^n}\sigma_{z_i^a}}$$

The sign of this correlation is thus given by the sign of $Cov\left(z_i^a/z_i^n, z_i^a\right)$. This is equal to

$$Cov\left(\frac{z_i^a}{z_i^n}, z_i^a\right) = E\left(\frac{z_i^a}{z_i^n}z_i^a\right) - E\left(\frac{z_i^a}{z_i^n}\right)E\left(z_i^a\right) = E\left(\frac{\left(z_i^a\right)^2}{z_i^n}\right) - E\left(\frac{z_i^a}{z_i^n}\right)\mu_a$$

where, in order to evaluate the two expectations in the last expression, we use the following second order Taylor series approximations around the means

$$\frac{\left(z_i^a\right)^2}{z_i^n} \approx \frac{\mu_a^2}{\mu_n} + \frac{2\mu_a}{\mu_n}\left(z_i^a - \mu_a\right) - \frac{\mu_a^2}{\mu_n^2}\left(z_i^n - \mu_n\right) + \frac{1}{2}\left[\frac{2}{\mu_n}\left(z_i^a - \mu_a\right)^2 + 2\frac{\mu_a^2}{\mu_n^3}\left(z_i^n - \mu_n\right)^2 - 4\frac{\mu_a}{\mu_n^2}\left(z_i^a - \mu_a\right)\left(z_i^n - \mu_n\right)\right]$$

and

$$\frac{z_i^a}{z_i^n} \approx \frac{\mu_a}{\mu_n} + \frac{1}{\mu_n}\left(z_i^a - \mu_a\right) - \frac{\mu_a}{\mu_n^2}\left(z_i^n - \mu_n\right) + \frac{1}{2}\left[\frac{2\mu_a}{\mu_n^3}\left(z_i^n - \mu_n\right)^2 - \frac{2}{\mu_n^2}\left(z_i^a - \mu_a\right)\left(z_i^n - \mu_n\right)\right].$$

Taking expectations we get

$$\begin{aligned}
Cov\left(\frac{z_i^a}{z_i^n}, z_i^a\right) &= E\left(\frac{\left(z_i^a\right)^2}{z_i^n}\right) - E\left(\frac{z_i^a}{z_i^n}\right)\mu_a \\
&\approx E\left(\begin{array}{c} \frac{\mu_a^2}{\mu_n} + \frac{2\mu_a}{\mu_n}\left(z_i^a - \mu_a\right) - \frac{\mu_a^2}{\mu_n^2}\left(z_i^n - \mu_n\right) + \\ \frac{1}{2}\left[\frac{2}{\mu_n}\left(z_i^a - \mu_a\right)^2 + 2\frac{\mu_a^2}{\mu_n^3}\left(z_i^n - \mu_n\right)^2 - 4\frac{\mu_a}{\mu_n^2}\left(z_i^a - \mu_a\right)\left(z_i^n - \mu_n\right)\right] \end{array}\right) \\
&\quad -E\left(\begin{array}{c} \frac{\mu_a}{\mu_n} + \frac{1}{\mu_n}\left(z_i^a - \mu_a\right) - \frac{\mu_a}{\mu_n^2}\left(z_i^n - \mu_n\right) + \\ \frac{1}{2}\left[\frac{2\mu_a}{\mu_n^3}\left(z_i^n - \mu_n\right)^2 - \frac{2}{\mu_n^2}\left(z_i^a - \mu_a\right)\left(z_i^n - \mu_n\right)\right] \end{array}\right)\mu_a \\
&= \frac{\mu_a^2}{\mu_n} + \frac{\sigma_a^2}{\mu_n} + \frac{\mu_a^2\sigma_n^2}{\mu_n^3} - 2\frac{\mu_a}{\mu_n^2}Cov\left(z_i^a, z_i^n\right) - \left(\frac{\mu_a}{\mu_n} + \frac{\mu_a\sigma_n^2}{\mu_n^3} - \frac{1}{\mu_n^2}Cov\left(z_i^a, z_i^n\right)\right)\mu_a \\
&= \frac{\sigma_a^2}{\mu_n} - \frac{\mu_a}{\mu_n^2}Cov\left(z_i^a, z_i^n\right),
\end{aligned}$$

Since we are only interested in the sign it follows that

$$\begin{aligned}
sign\left[Cov\left(\frac{z_i^a}{z_i^n}, z_i^a\right)\right] &= sign\left[\frac{\sigma_a^2}{\mu_n} - \frac{\mu_a}{\mu_n^2}Cov\left(z_i^a, z_i^n\right)\right] = \frac{\mu_n^2}{\mu_a}\frac{1}{\sigma_a\sigma_n}sign\left[\left(\frac{\sigma_a^2}{\mu_n} - \frac{\mu_a}{\mu_n^2}Cov\left(z_i^a, z_i^n\right)\right)\right] \\
&= sign\left[\frac{\mu_n^2}{\mu_a}\frac{1}{\sigma_a\sigma_n}\left(\frac{\sigma_a^2}{\mu_n} - \frac{\mu_a}{\mu_n^2}Cov\left(z_i^a, z_i^n\right)\right)\right] = sign\left[\frac{CV\left(z_i^a\right)}{CV\left(z_i^n\right)} - \rho\left(z_i^n, z_i^a\right)\right],
\end{aligned}$$

and therefore

$$sign[\rho\left(\frac{z_i^a}{z_i^n}, z_i^a\right)] = sign\left[Cov\left(\frac{z_i^a}{z_i^n}, z_i^a\right)\right] = sign\left[\frac{CV\left(z_i^a\right)}{CV\left(z_i^n\right)} - \rho\left(z_i^n, z_i^a\right)\right]$$

as stated in Proposition 1.

B.2 On the relation between the signs of $\rho\left(z_i^a/z_i^n, z_i^n\right)$ and $\rho\left(z_i^n/z_i^a, z_i^n\right)$

Notice that

$$sign\left[\rho\left(\frac{z_i^a}{z_i^n}, z_i^n\right)\right] = sign\left[Cov\left(\frac{z_i^a}{z_i^n}, z_i^n\right)\right] = sign\left[E\left(z_i^a\right) - E\left(\frac{z_i^a}{z_i^n}\right)E\left(z_i^n\right)\right],$$

Using a second-order Taylor series approximation around the means this becomes

$$\begin{aligned} sign\left[\rho\left(\frac{z_i^a}{z_i^n}, z_i^n\right)\right] &\approx sign\left[\mu_a - \left(\frac{\mu_a}{\mu_n} - \frac{Cov\left(z_i^a, z_i^n\right)}{\left(\mu_n\right)^2} + \frac{\sigma_n^2\mu_a}{\left(\mu_n\right)^3}\right)\mu_n\right] = sign\left[\frac{Cov\left(z_i^a, z_i^n\right)}{\mu_n} + \frac{\sigma_n^2\mu_a}{\left(\mu_n\right)^2}\right] \\ &= sign\left[\frac{Cov\left(z_i^a, z_i^n\right)}{\sigma_n\sigma_a} + \frac{\sigma_n\mu_a}{\sigma_a\mu_n}\right] = sign\left[\rho\left(z_i^a, z_i^n\right) - \frac{CV\left(z_i^n\right)}{CV\left(z_i^a\right)}\right], \end{aligned}$$

which from Proposition 1 equals $-sign\left[\rho\left(\frac{z_i^n}{z_i^a}, z_i^n\right)\right]$.

C Data Appendix

Our main source of data is the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA).[31] The LSMS-ISA project is a household survey project established with a grant from the Bill and Melinda Gates Foundation. The project is housed within the Survey Unit of the World Bank's Development Data Group. It provides technical assistance to national statistical offices (NSOs) of its eight partner countries in Sub-Saharan Africa in the design and implementation of multi-topic household surveys. Its purpose is to design and implement systems of multi-topic, nationally representative panel household surveys with a strong focus on agriculture. In each partner country, the LSMS-ISA supports multiple rounds of a nationally representative panel survey with a multi-topic approach designed to improve the understanding of the links between agriculture, socioeconomic status, and non-farm income activities. The frequency of data collection is determined on a country-by-country basis, depending on data demand and the availability of complementary funding. For our purpose, we use data from the following countries, waves, and number of observations

- Ethiopia - Socioeconomic Survey (ESS)
  - Wave 1 – Year 2011/12 – $N = 3,969$
  - Wave 2 – 2013/14 – $N = 3,804$
  - Wave 3 – 2015/16 – $N = 3,726$
- Malawi - Third Integrated Household Survey (IHS3), Integrated Household Panel Survey (IHPS)
  - Wave 1 – 2010/11 – $N = 3,247$
  - Wave 2 – 2013 – $N = 2,872$
- Nigeria - General Household Survey (GHS)
  - Wave 1 – 2010/11 – $N = 4,928$
  - Wave 2 – 2012/13 – $N = 4,716$
  - Wave 3 – 2015/16 – $N = 4,575$
- Uganda - National Panel Survey (UNPS)
  - Wave 1 – 2009/10 – $N = 2,975$
  - Wave 2 – 2010/11 – $N = 2,703$
  - Wave 3 – 2011/12 – $N = 2,748$
  - Wave 4 – 2013/14 – $N = 1,832$.

Each country-year sample follows a stratified two-stage sample design to ensure national representativeness. Enumeration areas (EAs) are selected with probability proportional to size within each district of the country. Random systematic sampling is used to select a certain number of primary households and some replacement households from the household listing for each sample EA. A sub-sample is randomly selected to be visited twice during the first survey to reduce recall associated with different aspects of agricultural data collection. The selected households are then tracked and resurveyed and serve as a baseline for the panel follow-up.

LSMS-ISA surveys typically include three main questionnaires: household (H), agriculture (AG), and community (C). As part of the agriculture questionnaire, fishery questionnaires are sometimes listed independently. In the agriculture questionnaire, households also report information separately on the last completed rainy and dry seasons, or post-harvest and post-planting season. These surveys collect detailed information at the household (and individual) level on income, health, education, expenditure and consumption, labor allocation, asset ownership, and details on agricultural production, business operation, and other economic activities. The surveys undertaken in different countries do not always follow identical methodologies; nevertheless, efforts have been made to follow the same method as much as possible in generating variables used in the empirical analysis. These micro-data allow us to compute measures of household-level value added by agricultural and non-agricultural activity for the four countries considered.

[31] See also http://surveys.worldbank.org/lsms [consulted on October 9, 2018].

All of the LSMS surveys are publicly available from the World Bank website. A basic information document is available for each country, as are the survey questionnaires themselves.

Agricultural Value Added The agricultural activities of each household are generally reported separately for non-permanent crop harvested and sold, permanent crop harvested and sold, livestock sales, livestock products sales, and fishery sales. We follow Gollin, Lagakos, and Waugh (2014) and calculate the agricultural value added $VA_{a,i}$ of household i as the sum of value added from non-permanent crops ($VA^{NPC}_{a,i}$), permanent crops ($VA^{PC}_{a,i}$), livestocks ($VA^{LS}_{a,i}$), livestock products ($VA^{LSP}_{a,i}$), and fishery ($VA^{FS}_{a,i}$), i.e.

$$VA_{a,i} = \left(VA^{NPC}_{a,i} + VA^{PC}_{a,i} + VA^{LS}_{a,i} + VA^{LSP}_{a,i} + VA^{FS}_{a,i}\right)$$

Agricultural activities are questioned and reported in the survey in different seasons. Let

$$z \in \{NPC, PC, LS, LSP, FS\}$$

identify the different agricultural activities and let s identify the rainy and dry seasons respectively in the agricultural activities, or representing high or low landing season in the fishery survey. Similarly to de Magalhaes and Santaeulalia-Llopis (2018) and Restuccia and Santaeulalia-Llopis (2017), we calculate the value added from product c of agricultural sector z as the sum across seasons of each household i's revenues from selling product c in season s ($Rev^z_{c,s,i}$), plus the market value of the product c of that was not sold (recorded as stored, lost, saved for seeds, etc) measured as $P^z_{c,s,i}(Output^z_{c,s,i} - Sold^z_{c,s,i})$, while subtracting the associated costs ($Cost^z_{c,s,i}$), i.e.

$$VA^z_{a,i} = \sum_s Rev^z_{c,s,i} + \sum_s P^z_{c,s,i,r}(Output^z_{c,s,i} - Sold^z_{c,s,i}) - \sum_s Cost^z_{c,s,i}$$

$P^z_{c,s,i,r}$ is the inferred price of the product c in agricultural sector z in season s produced by household i in region r. Prices are imputed as follows:

- If household i sold crop c in season s and reported total sales $Rev^z_{c,s,i}$ and quantity sold $Q^z_{c,s,i}$, we let $P^z_{c,s,i,r}$=$Rev^z_{c,s,i}/Q^z_{c,s,i}$;
- Otherwise, we attribute the average price of the crop sold by other households in the same region if available, meaning $P^z_{c,s,i} = \overline{P}^z_{c,s,j}$ with j being in the same region as i;
- Otherwise, we attribute the regional community price reported in community section, meaning $P^z_{c,s,i} = P^z_{c,s,com}$.

In agricultural production, each household i incurs cost $Cost^z_{c,s,i}$ per season s associated with cost type v. That is

$$Cost^z_{c,s,i} = \sum_v Cost^z_{c,s,i,v}$$

where v ={intermediate goods purchased (fertilizer, seeds, pesticides/herbicides), hired labor, rented capital (and land), transportation} if $z \in \{NPC, PC\}$; v ={intermediate goods purchased (animal feed, vaccinations, other inputs), hired labor, housing equipment, feeding utensils, transportation, veterinary services} if $z \in \{LS, LSP\}$; v ={energy cost (fuel, oil, maintenance), hired labor, rented capital (gears, boats/engines), other cost} if $z = FS$.

Entrepreneurial Profits We define household i's annual non-agricultural value added $VA_{n,i}$ as the sum of profits of all enterprises owned by the household. We identify households engaged in any kind of non-agricultural income-generating activity (owned a non-agricultural business or provided a non-agricultural service, owned a trading business, owned a professional office or offered professional services, etc.) in the last 12 month before the interview. For each household i we compute entrepreneurial profits as the total annual sales minus costs across all enterprises in the household. The value of annual

total sales is annualized from the average monthly sales reported by each enterprise. And the value of annual total costs per enterprise is also annualized from the average monthly costs which consist of variable costs including raw materials, inventory, freight/transport, fuel/oil, electricity, water, insurance, etc. and total wages/salaries paid to hired labor. We thus compute

$$VA_{n,i} = \sum_i (Rev_{n,i} - Cost_{n,i})$$

where $Rev_{n,i}$ is imputed annual revenues in non-agricultural businesses n owned by household i, and $Cost_{n,i}$ is the annual aggregation of any intermediate or factor cost incurred in the same non-agricultural business.

Labor Hours In order to study individual labor supply and the intra-household allocation of time, we use information on hours worked by each individual for the household farm or the household non-farming business. This information belongs to the time use module of the household questionnaire which asks for the amount of hours spent in each activity over the last 7 days. We compute working hours in agriculture as the sum of hours spent in agricultural activities (including livestock and fishing-related activities) whether for sale or for own consumption. We compute working hours in entrepreneurship as the sum of hours spent in any kind of non-farming household business. We then aggregate this information at the household level within and across the two activities.

The household questionnaire was always administered together with the post-harvest questionnaire in all survey waves in Ethiopia and Nigeria. In Malawi, half of households in the sample received the household questionnaire together with the post-planting one during the first visit, while the remaining half received it together with the post-planting one. In Uganda, the documentation provides no information on when the household questionnaire was administered.

Land Use The land available to each household is identified as the cumulative area of plots that any member of the household owns or cultivates. The area of the land is measured by farmer estimation and GPS measurement. We identify the ownership status of the plot as acquired by decision of the local leader, inheritance, or rented. We use this information to calculate the total cultivated area, and fraction of land that is rented, which we also consider a proxy for land market development.

Household characteristics The data provide individual demographic characteristics of household members including sex and birth year. We derive the total number of household members and the total number of female household members. In all waves following the first, the questionnaire asks if any household member left the household since the previous interview, which we use to capture migration. We also derive for each household an index of asset ownership by counting the number of assets the household reports to have. The list of assets is country-specific, therefore so is the index we derive.