Is wealth found in the soil or brain? Investing in farm people in Malawi

M. Mkondiwa;

University of Minnesota, Applied Economics, United States of America

Corresponding author email: mkond001@umn.edu

Abstract:

Should a typical developing country invest more in agriculture or education? At what stage of development is it optimal to invest more in each of these sectors? Every developing country government grapples with these questions annually when designing a national budget. In this paper, I provide estimates of agricultural returns to schooling in Malawi- evidence of such returns implies a more complex non-separable decision process to answer the first question. While a large development economics literature has documented the effects of schooling on agricultural incomes, such estimates are potentially biased because of unobserved heterogeneity and selection bias. In this paper, I use 2010-2013 two period nationally representative panel survey data in Malawi and rely on the exogenous education policy changes and spatial variation in access to schooling to identify effects of schooling on agricultural incomes. In addition, I use recent econometric methods to correct for selection into income activities within a panel data and instrumental variables estimation framework. I find annual agricultural returns to one additional year of schooling in Malawi that range from 3% to 7%.

Acknowledegment:

JEL Codes: O15, O12

#2483

Is wealth found in the soil or brain? Investing in farm people in Malawi

"In sum and substance, the man who is bound by traditional agriculture cannot produce much food no matter ***how rich the land****.* ***Thrift and work*** *are not enough to overcome the niggardliness of this type of agriculture. To produce an abundance of farm products requires that the farmer has access to and has* ***the skill and knowledge to use what*** *science knows about soils, plants, animals, and machines."* Schultz (1964, p.205)

"Chuma chili mthaka- (*Wealth is found in the soil)*"- Malawian proverb

1. Introduction

Is wealth[1] found in the soil or brain? This question is elusive to households in developing countries and -arguably in equal or worse measure- to economists. Until the 1960s, the leading economists- Adam Smith, Ricardo, Malthus- had postulated that only physical capital investments on natural resources like land or soil account for the differences between the haves and have nots. This was a natural earth view. A later position starting with Theodore Schultz and Gary Becker was that land is overrated and argued that income differences are mostly due to human capital differences. For instance, Schultz argues in his 1979 Nobel Laureate Lecture (Schultz 1980) that, "differences in the productivity of soils are not a useful variable to explain why people are poor in long-settled parts of the world". His argument and that of Becker (1994) was that education is of little use in traditional agriculture because farming methods and knowledge are then readily passed from parents to children while in an agricultural sector where technical change is happening, educated farmers are better able to deal with disequilibria by seeking more information and reallocating inputs efficiently.

A mix of the soil versus human capital positions appear to date. The effects of holding one side or the other are not only of academic interest; huge amounts of financial resources across the world are allocated based on these positions. For example, the World Bank Group pushed for structural adjustment programs across the developing world in the 1980s-90s in large part by shifting the

[1] I am using the term "wealth" loosely here to refer to income flows not stocks as is the case in most of economics literature.

focus from agricultural subsidies-mostly fertilizer support programs that reflect investments in the soil-in the new colonial governments to education. In the recent past, investments in the soil have again come back on the agenda across sub-Saharan Africa while rural education programs have remained stagnant. This culminated in the 2008 World Development Report- World Bank's flagship report – focusing on agriculture as the driver to development. There is however a renewed interest in the international community towards evidenced by the detailed 2018 World Development Report which focuses on learning as a means to realize education's promise. The evidence for such policy moves on welfare in developing countries is sparse and difficult to study. For a start, one must assign treatment over many decades to determine which is beneficial in the short, medium and long run. Such a randomized experiment is unlikely and the best approach is exploiting a natural experiment as I attempt in this paper.

The main objective of the paper is to estimate the agricultural returns to schooling in Malawi. Controlling for unobserved human and soil quality, self-selection and endogeneity, such estimate would imply that there is additional wealth in the brain beyond what can be earned in the non-farm sector. The evolution of education and agricultural policies in Malawi offers a unique a natural experiment due to the differential effects of such policies on different households based on age cohorts and access to schooling measured by distance to schools. The identifying assumption is that distance to schooling and cohorts have no effect on agricultural incomes (except through schooling) after accounting for a battery of controls like distance to road, district fixed effects, gender, age and place of birth.

The paper follows the modified Mincer-Becker approach- two-sector extensions- as used for developing country models of schooling effects on choice of income activities and level of investment in rural economies by Taylor & Yunez-Naude (2000), Yang & An (2002), Yang (2004), Duflo (2004), Jollife (2004), and Reimers & Klasen (2013. The evidence on effects of schooling on agricultural incomes and productivity has been rather mixed from this literature. Using data on educational attainment, Reimers & Klasen (2013) find a sizable and significant impact of schooling (average increase of approximately 3.2% per year of schooling) on agricultural productivity that is robust to estimation methods and model specification. Taylor and Yunez use cross sectional data for rural Mexico and a system of selection corrected Mincerian equations to investigate the returns to schooling in crop and non-crop activities. They found that returns to schooling are high- 10% per household head's year of schooling and 5.5% per year of average schooling of other members. There are no returns to staple crop farming. Yang & An use cross sectional data in rural China to analyze the contribution of schooling to sectoral allocation and found that while schooling enhances agricultural and nonagricultural profits, about 14% of total returns arise from optimally choosing activities through the allocation of household-supplied inputs. Fafchamps & Quisumbing (1999) report that one-third of gains to schooling are due to labor reallocation towards non-farm sector in Indonesia. Yang (2004) uses panel data to investigate the sources and determinants of sustained income growth in rural China between 1986 and 1995 and found that schooling plays a critical role in raising the efficiency of farmers to respond to changing market conditions.

Methodologically, this literature can be subdivided into two strands: one based on the assumption that labor allocation across the sectors is the only quasi-fixed factor through which education affects agricultural incomes and another rooted on the assumption that there are a myriad of other factors affecting the choices in activities through which education affects agricultural incomes. Yang (2004) and Jolliffe (2004) argue that labor allocation is the only endogenous factor in production, but this may not be true in most rural areas as other salient resources like bicycles, ox-carts, livestock and hoes may all be quasi-fixed factors through which education can affect agricultural incomes. Taylor & Yunez-Naude (2000) on the other hand assumes there is a host of factors affecting the decision to participate in agriculture which has to be corrected for in the estimation of agricultural returns to schooling. The gap in this literature is the failure to account for the endogeneity of education and self-selection not just due to labor allocation but on the lines of other several salient household, production and marketing features. An example is soil quality which has an intertemporal relationship with education such that such unobserved effects may drive the schooling effects on agricultural incomes. Wantchekon & Stanig (2016) using a combination of unique datasets in sub-Saharan Africa dubbed the relationship between soil quality and poverty in the region as: the curse of good soil. One of the suggested mechanisms through which this can occur according to the authors is through human capital. Loosely, isolated areas with good soil quality may produce enough food leading to higher fertility rates and lower human capital accumulation.

A modeling framework that corrects for unobserved heterogeneity, endogeneity of schooling due to motivation and ability and selection bias is therefore a first step in understanding schooling effects on agricultural incomes. Among the competing methods for doing this, the best approach is by Semykina & Wooldridge (2010) because it does not rely on any known distribution of the errors in the equation of interest, and allows them to be time heteroskedastic and serially correlated in an unspecified way (Jäckle & Himmler 2010). Using this modeling approach and two rounds of nationally representative household survey data (2010, 2013) in Malawi, this paper makes one main contribution. I provide estimates of agricultural returns to schooling that are free from bias from selectivity of economic activities and unobserved soil or human quality within a panel data framework. Previous literature either used cross sectional data with selectivity corrected or used panel data while neglecting the selection issues. I find that agricultural returns to schooling are between 3-7% across all specifications (though not significant in the preferred specification) and that failure to correct for endogeneity biases the estimates upwards.

The plan of the paper is as follows. The next section presents the theoretical and empirical model. Section 3 presents the data and descriptive statistics followed by a section on results. The research and policy implications are presented in section 5. Finally, the conclusion is presented in section 6.

1.1 Background and motivation

The question of investing in the soil or brain is not just a rhetoric among academic economists- it has also been evident in the political rhetoric in Malawi. For instance, the first president of Malawi, the late Hastings Kamuzu Banda was identified with soil centric phrases like *Chuma chili mthaka* meaning wealth is found in the soil and *mchikumbe number 1*, a phrase for exalting the most successful farmer. The third president, late Bingu wa Muntharika is recognized for emphasizing that Malawi is not poor but rather it is the people who are poor implying the huge endowments including soil resources that the country has.

Malawi's educational reform history can be described in tandem with its political story. There are three important political turning points: colonial, post-colonial dictatorship and post-colonial democratic. During the colonial years (up to 1964), the education system was controlled and governed by the missionaries whose emphasis was on vocational skills of the citizens than on literacy. The colonial government had no interest in educating the local population- it was estimated that missionaries invested ten times more than did the government (Heyneman 1980). Access to schooling was therefore extremely low; there were only four secondary schools by 1964 and no university education. During the post-colonial dictatorship period (1964-1994) this ideology did not change much and only few schools were added with noticeable changes in primary school enrolment (30% increase from 1962 to 1972 and 57% between 1972 and 1978). During this period, there was introduction of university of education and increase in literacy. During the democratic dispensation (from 1964), free primary school education was introduced with substantial transformation the curriculum. According to Benson et al. (2018), stocks of human capital have improved since the introduction of free primary education in 1994. It is estimated that this policy alone increased enrolment by half, favoring girls and poor people (World Bank 2018).

Besides education reforms, the year 1994 is associated with several major reforms in the agricultural sector and across the economy. Malawi dropped agricultural subsidy programs and instituted deep liberal reforms in the marketing and production of food and cash crops. It is therefore naïve to attribute outcomes to any of these effects. In this paper, I use these transitions as a potential natural experiment for identifying effects of schooling on agricultural incomes. The figure 1 shows the trend in public expenditure to agriculture and education from 1980-2010. The trends of sectoral expenditures were similar up to 1994 but then diverged due to the reforms. In 2004/5, the government reintroduced heavy agricultural subsidies which led to the increase in public expenditure in the agricultural sector.

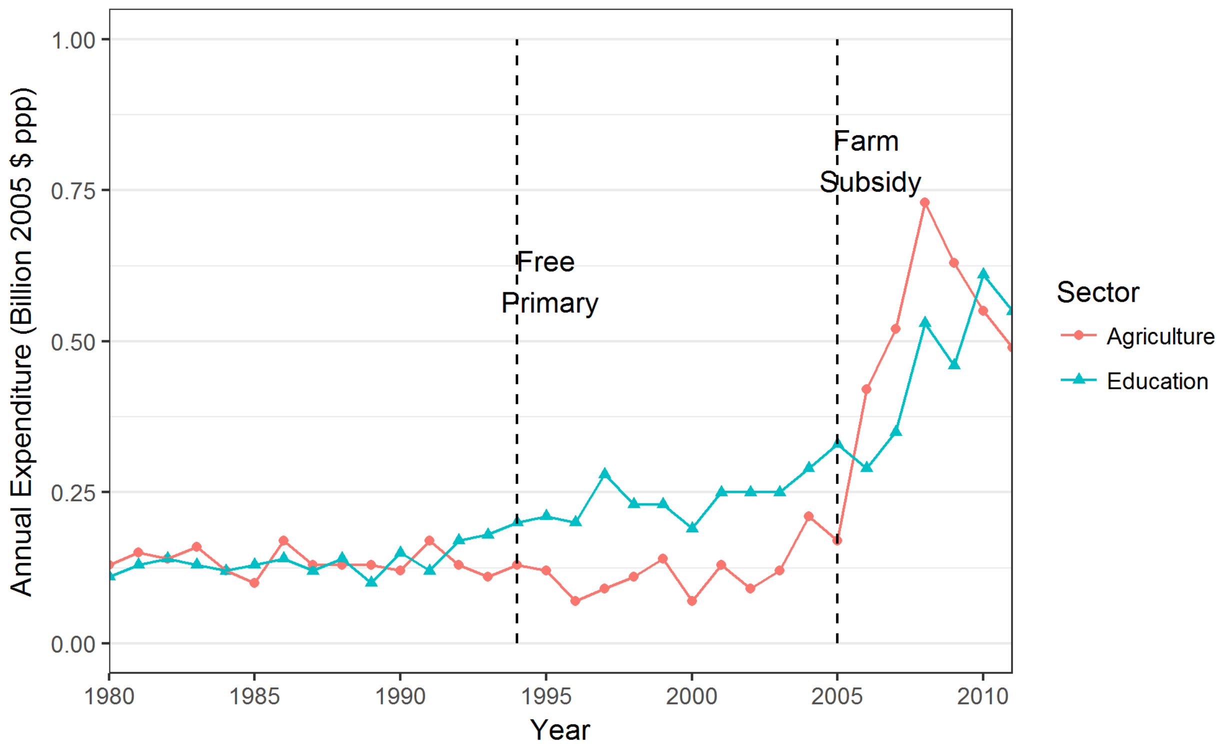

Figure 1 : Annual public expenditures in the agriculture and education sectors

Data source: Statistics on Public Expenditures for Economic Development (SPEED)[2]

Several studies have studied effects of schooling on activity choices and earnings in Malawi. These include; Gondwe (2016), Benson et al. (2018), Chirwa & Matita (2009) and Matita & Chirwa (2009). Gondwe (2016) and Benson (2018) use the same panel data as in this paper to analyze the determinants of employment choices and returns to schooling. Gondwe (2016) find that returns to schooling are high in the formal sector (8.7%) than in the informal sector (4.7%). With this categorization, it is difficult to distinguish the effects on agricultural incomes. Studies by Chirwa and Matita found a wide ranging estimates of returns schooling, 5% to 65% across primary and university education respectively. In this paper, I use econometric tools that in addition to the methods used in this literature are a small step towards causal identification of agricultural returns to schooling.

[2] The data are provided by International Food Policy Research Institute (https://www.ifpri.org/program/statistics-public-expenditures-economic-development-speed)

2. Model

This section presents the theoretical and empirical models that help determine household's investment in the soil or in the brain. The theoretical model provides a justification for the difficulty that the households face in determining as to whether to invest in the soil or in the brain of the children. This problem leads to multiple optima that define the heterogeneity in the investments and returns to schooling. I argue that this heterogeneity in returns to schooling can be explained by heterogeneity in soil quality, level of existing human capital within the household, access to schooling, and proximity and ability to purchase fertilizer from non-farm income sources.

2.1 Theory on human capital and soil quality

In this section, I propose a theory that I argue drives the mechanisms through which schooling affect agricultural incomes. The theory is based on the transitions in human capital and soil fertility overtime based on the evidence across the world and the agronomic theory. It is assumed that education influences both farm and non-farm productivity, that is, using the analogy on sources of wealth, education makes it possible for an individual to rely on wealth both from the soil and in the head. The theoretical hypothesis in this paper is that at low levels of human capital and high levels of soil quality, household invest most of their labor in working on the farm. The investments in material inputs to improve soil fertility during this period are small both because of the high soil quality and the lack of skills to search for new factors of production. As the human capital grows and soil loses its vigor, households invest more in material inputs to compensate for the loss in soil quality.

At high levels of human capital, household sharply reduce their investments in the farm sector, and invest more in enhancing attainment of high knowledge skills. The switch away from the soil may not simply mean leaving agriculture all together, it may imply investing in agro-business, agro-processing and other forms of investments that make the soil investments more productive. This phenomenon therefore suggests three generational changes in human capital in which households find wealth in the soil in first two generations, and immediately find wealth in higher knowledge skills. The number of years for each generation depends on the investment choices the society makes in both the agricultural and education sector. A country that dramatically increases its human capital will eventually find itself at high levels of intensive agriculture, thus both high investments in the soil and in the skills of its farm people.

This narrative is consistent with the transition from Malthus to Solow theory by Hansen & Prescott (2000) or agriculture to industry by Tamura (2002). Tamura for example argues that "for low human capital, the agricultural (Malthus) method dominates the industrial method (Solow), but for high human capital, the industrial method dominates the agricultural method." The literature in line with this argument attributes this transition to the nature of production technology, use of land in traditional method and falling fertility. In this paper, I argue that such transitions are due to falling soil fertility that the switch from agriculture to non-agriculture occurs as soils get degraded and technologies for soil fertility replenishment are not effective.

This theory implies that empirical models need to capture the selection into economic activities that occurs between and within generations and the potentially unobserved heterogeneity from time invariant and variant soil quality and other unobservable factors like motivation. These factors affect both schooling investments and incomes of the households.

2.2 Empirical model

In this section, I present the estimation and identification strategy. The estimation strategy is based on two salient features of the research problem: selection of individuals into agricultural and non-agricultural sectors, and the unobserved heterogeneity across individuals.

2.2.1 Estimation and identification strategy

The structural (outcome) model consists of activity specific (k) equations

$$\ln \Pi_{ikt} = \pi_{ikt} = \beta_0 + X_{ikt}\beta_1 + E_{itk}\beta_2 + \epsilon_{ikt}; \quad \text{for k} = 1, 2 \tag{1}$$

Where $\ln \Pi_{ik}$ is the vector of logarithm of incomes for each household i in activity k. X_i is a vector of exogenous regressors (district dummies, age, household size gender, place of birth, marital status, soil quality measures, elevation, distance to road, distance to nearest population center). E is the endogenous schooling variable measured in years of schooling. The challenge in estimating equation is that π_{ik} is not observed for some activities for some households therefore this self-selection has to be addressed. Secondly, education is endogenous literature due to omitted variable bias from lack of information on the ability or motivation of each household head. In the case of farming, the omitted variable bias can be upward (if higher ability individuals are also good at farming) or downwards (if high ability individuals are lazy and impatient to do strenuous physical work that may pay after 5 months). These justifications are based on anecdotes. A research strategy that addresses these challenges is discussed next.

In addressing endogeneity and selection, I follow the approach suggested by Yang & An (2002) for cross sectional data and Semykina & Wooldridge (2010) for panel data in which two equations are added to the structural equation

Instrument equation:

$$E_{itk} = \delta_0 + W_{ikt}\delta_1 + \eta_{ikt} \quad \textbf{(2)}$$

Selection equation:

$$S_{ikt} = 1[\alpha_0 + M_{ikt}\alpha_1 + W_{ikt}\alpha_2 + \omega_{ikt} > 0] \quad \textbf{(3)}$$

Where $1[.]$ is the indicator function so that S_{ikt} is a selection indicator that equals 1 for $\Pi_{ikt} > 0$ and 0 otherwise, W_{ikt} is a vector of exogenous regressors in the instrument equation and selection equation (includes variables in X_{ikt} and instruments i.e. distance to nearest secondary school and cohort), M_{ikt} is a vector of exogenous regressors in the selection equation only and different from the instruments. η_i and ω_i are independent identically distributed disturbance terms.

These three equations imply six combinations of econometric models:

i) Pooled OLS

This model estimates the naïve outcomes equation (1).

ii) Pooled 2 SLS

Most identification strategies in this literature have concentrated on accounting for selection bias introduced by the choice of economic activities and the omitted variables bias due to unmeasured ability or motivation. In this study, I control for further sources of omitted variables, i.e. soil quality and existing human capital within the community/village. The endogenous variables that are usually considered are labor allocation and education. Two- stage least squares approach is usually used to control for omitted variables by using instruments. In the case of the endogeneity of labor allocation variable, several strategies have been used. For example, (Yang 1997a) uses the number of household members in the labor force as the instrument to correct for the potential endogeneity of family labor supply. Yang (1997b) use household population as an instrument for agricultural labor input and the value of durable consumption goods is chosen as an instrument for productive capital input. Yang & An (2002) use aggregate land/labor and capital\labor ratios and average age of family members as instruments for quasi-fixed factors like labor. For endogenous human capital variables (education), Yang and An use the same instrumental variables. These instruments are suspect on the exclusion restriction. According to (Card 2001) an IV procedure that implicitly compares many subgroups of individuals-say, younger versus older cohorts in several different regions-may be more reliable than one that relies on a single affected subgroup. The LSMS data I use provides the geo-coordinates of each of the household in the sample. I use this unique feature to calculate distances between the homestead and the nearest secondary school point using a secondary schools shapefile. The identifying assumption is that the effect of proximity to secondary school among the young cohorts (those who may be affected by removal of fees) on incomes is only through educational attainment after controlling for a battery of controls including soil fertility, weather, district fixed effects, isolation variables les. This is plausible considering that for poor households, distance to the nearest school is a predictor of school participation, especially where social norms and safety concerns make it difficult for children to travel far from home (World Bank 2018). These reasons strongly apply to the case of Malawi.

iii) Fixed effects panel data model

The two models so far presented are based on selection on observables design in that the available variables in the data are deemed enough to account for all variables that can be correlated with the schooling variable. While the pooled 2SLS is an improvement in correcting for endogeneity, it is still based on assumption of having the observable variables as instruments. The sheer nature of lack of knowledge on the part of the researcher that each household has, there may be unobservable variables that may be correlated with schooling that are yet to be discovered as important. Such unobserved household heterogeneity which may be in the form of land or soil quality, weather, farmer skill, input/output market conditions and motivation can be controlled for through the use of panel data (Mundlak 1961; Xu et al. 2009; Satriawan & Swinton 2007) particularly by adding household fixed effects to the outcomes equation

$$\ln \Pi_{ikt} = \pi_{ikt} = \beta_0 + c_{ik} + X_{ikt}\beta_1 + E_{itk}\beta_2 + \epsilon_{ikt}; \quad \text{for k} = 1, 2 \tag{4}$$

Where the introduced term c_{ik} is the unobserved effect which may be correlated with other variables in the equation. The identification in this framework is therefore based on the time variation in schooling effects within each household-thus causally identifying the effect of schooling on variation in incomes over the two panels controlling for farm's quality and household's management quality.

iv) Fixed effects 2SLS panel data model

The fixed effects model offers a straightforward way to control for time-invariant unobserved heterogeneity like specific production skills, biophysical conditions related to location and social relationships that may be related to schooling and affect agricultural incomes. The model does not however address the time variant sources of endogeneity which can be addressed using instrumental variables as justified by pooled two-stage least squares. This model therefore estimates the outcome equation (4) together with the instrument equation. I use the same IVs as in the pooled 2SLS model. This therefore allows a comparison of estimates to the fixed effects models and to the pooled 2SLS model.

v) Test for selectivity with Fixed effects 2SLS panel data model

Perhaps the common source of unobserved heterogeneity in returns to schooling in rural economies is that households derives their earnings from multiple sources which are selected by households on the basis on several reasons unknown to the econometrician. (Taylor & Yunez-Naude 2000) underscored the importance of incorporating endogenous activity choice into returns-from-schooling models. (Wouterse & Taylor 2008) uses a selection model similar to Taylor and Yunez in analyzing the effect of migration on income diversification in Burkina Faso. Each of selection models considers only a case where other variables are exogenous. Selection into agricultural or non-agricultural activities can be tested by estimating the fixed effects two-stage least squares together with a selection term from a modified selection equation.

$$S_{ikt} = 1[\alpha_0 + c_{ik} + M_{ikt}\alpha_1 + W_{ikt}\alpha_2 + \omega_{ikt} > 0] \quad \textbf{(5)}$$

The problem of doing this is that for non-linear models like the probit that are used in first stage in selection models do not behave well under a panel fixed effects estimation. Semykina & Wooldridge (2010) suggests a procedure that relies on the equivalence between the fixed estimation and the Mundlak device (Mundlak 1978) which models the unobserved fixed effect as a function of the time means of the time-varying explanatory variables

$$c_{ik} = \xi_{0k} + \bar{M}_{ik}\xi_1 + \bar{W}_{ik}\xi_2 + \mu_{ik} \quad \textbf{(6)}$$

Where $\bar{M}_{ik}$ and $\bar{W}_{ik}$ are vectors of means of M_{ikt} and W_{ikt} across the two panels, ξ_{0k} is a constant term and μ_{ik} independently and identically distributed and uncorrelated with ϵ_{ikt}. Many variants of the Mundlak device are used in production economics to control for unobserved heterogeneity such as time-constant farmer ability and soil variation (Xu et al. 2009).

Following Semykina & Wooldridge (2010a) procedure 3.1.1, I use the following steps:

a) Estimate probit of C on all observations for each year t. Obtain the estimated inverse Mils ratio, $\hat{\lambda}_{it} = \lambda(\widehat{\xi_{0k}} + \bar{M}_{ik}\widehat{\xi_1} + \bar{W}_{ik}\widehat{\xi_2})$.
b) Using the selected subsample, I estimate the outcomes equation

$$\ln \Pi_{it1} = \beta_0 + \bar{M}_{ik}\xi_1 + \bar{W}_{ik}\xi_2 + X_{it1}\beta_1 + E_{it1}\beta_2 + \gamma_1\hat{\lambda}_{it} + e_{it1}$$

By fixed effects two stage least squares, using instruments $(W_{i2}, \hat{\lambda}_i)$.

vi) Correcting for selectivity with pooled 2SLS (Mundlak type) panel data model

This model is estimated following the parametric procedure 4.1.1 from Semykina & Wooldridge (2010a);

a) Estimate probit of C on all observations for each year t. Obtain the estimated inverse Mils ratio, $\hat{\lambda}_{it} = \lambda(\widehat{\xi_{0k}} + \bar{M}_{ik}\widehat{\xi_1} + \bar{W}_{ik}\widehat{\xi_2})$.

b) Using the selected subsample, estimate the equation

$$\ln \Pi_{it1} = \beta_0 + \bar{M}_{ik}\xi_1 + \bar{W}_{ik}\xi_2 + X_{it1}\beta_1 + E_{it1}\beta_2 + \gamma_1\hat{\lambda}_{it} + e_{it1}$$

By pooled 2SLS, using instruments $(W_{i2}, \hat{\lambda}_i)$.

For all the models, I use cluster bootstrapped standard errors for inference for two reasons. The LSMS data used in the study has a sampling design with two stages: first enumeration areas then households within enumeration areas. According to Abadie et al. (2017), in a two stage sampling design, clustering adjustment is justified by the fact that there are clusters in the population that are not seen in the sample. The second reason is that for computed values from first stage regressions are being used in the next stages therefore to address a generated regressor's problem using bootstrap standard errors is a way to resolve the challenge. For selection procedures in (v) and (v), I use panel clustered bootstrap standard errors as suggested by Semykina & Wooldridge (2010) because the outcome models have estimated right hand variables.

3. Data and descriptive statistics

3.1 Data sources

The primary source of the data for the empirical analysis is the Malawi Living Standards Measurement Survey-Integrated Agriculture Survey (LSMS-ISA) data collected by the National Statistical Office in collaboration with the World Bank's Living Standards Measurement Unit. This paper uses the 2010-2013 panel sample (Integrated Household Panel Survey or IHPS) consisting of about 4000 households out of the 12721 interviewed in the 2010 survey (also called Integrated Household Survey 3 or IHS 3). The detailed survey information is presented in the 2010 survey handbook (NSO 2012) and 2013 panel survey report (NSO 2014). I briefly discuss the sampling design and the nature of the panel component of the survey.

The IHS 3 is a third wave of the nationally representative household surveys conducted every five years in Malawi. The sample for the survey is selected following a two-stage sample design. First stage involves sampling 768 primary sampling units or enumeration areas (EAs) of the 12,575 EAs defined based on the 2008 Population and Housing Census. The panel subcomponent of the sample was based on sampling 204 EAs in 2010. The second stage involved randomly sampling households within each EA based on a listing of all households within the selected EAs. The sample size for the panel subsample was 3246 households in 2010. For the 2013 survey, once a split-off individual was located, the new household that he/she formed/joined since 2010 was also brought into the IHPS sample. Due to this the IHPS includes a total of 4,000 households that could be traced back to 3,104 baseline households with overall attrition rate of only 3.78 percent at the household level (NSO 2014).

In addition to the IHPS data, I also obtained the secondary schools geo-coordinates from the Malawi Spatial Data Platform (http://www.masdap.mw/). The school geolocation data are used to calculate the distance between each household geo-coordinates and the nearest secondary school. This variable is used as one of the instruments in a two-stage least squares estimation.

3.1.1 Variable construction

The LSMS data is publicly available and being used in several studies that include the key variables in the study. It is therefore important to be explicit on how each of the variables is constructed to ensure the reproducibility of the results.

(a). Income variable

The income variable is constructed by summing up all revenue from all economic activities at annual level. The activities include all agricultural activities like crop sales, livestock sales, trees sales- and non-agricultural activities like formal employment and household enterprises. I use sales like few studies (e.g. Pudasaini 1983) instead of profits to avoid problems of measurement errors in the accounting of costs especially in agricultural and informal income generating activities.

(b). Schooling variable

The schooling variable was constructed based on grades the household reported to have complemented. Therefore, grades 1-8 correspond to the respective schooling years. While this may introduce measurement errors due to repetition and jumping of classes, it presents well the type of knowledge the person has upon leaving school.

(c). Cohorts

The two cohorts-young and old- are computed from the age variable following the assumptions of the age group that would be affected by the education reforms (like free primary education) in 1994. During those years and to-date, children start schooling at age 5 or 6. However some would even start at the age of 10 as such the young cohort is considered those less than 30 so that the oldest within the young cohort would be only 14 years in 1994. The older cohort includes all those above 30.

(d). Distance to secondary schools

Most students in Malawi walk to school as such school distance is the key factor to access to schooling. In addition, for those close to school, parents generally have psychological incentives to send children to school because they are not concerned with security concerns as compared to when children travel long distances away.

3.2 Full sample descriptive statistics

Table 1 provides descriptive statistics by panel year and age cohorts of household heads. The panel was balanced by matching all 2010 households to 2013 households including the newly formed households based on splits. Most of the household heads (78%) are in the old cohort i.e. older than 30 years. About 80% of the households reported positive household incomes or revenues. About 62% report positive non-agricultural earnings while only 48% report positive agricultural incomes. These statistics are very important because though over 90% is engaged in agriculture, only a sizable portion use it as a source of income. Agricultural incomes are generally lower than non-agricultural incomes across all cohorts and years. Average levels of schooling are very low (6 years corresponding to grade 6).

Table 1: Sample descriptive statistics (means)

	Panel Year		**Cohort**		
	2010	2013	Old	Young	All
N	3888	3888	6052	1722	7776
Dependent variables					
Total household income reported	0.89	0.72	0.80	0.79	0.80
Non-agricultural income >0	0.74	0.51	0.61	0.66	0.62
Agricultural income > 0	0.47	0.50	0.50	0.42	0.48
Total household income (Malawi kwacha)	200000	140000	180000	120000	170000
Non-agricultural income (Malawi kwacha)	180000	98000	150000	100000	140000
Agricultural income (Malawi kwacha)	20000	40000	33000	17000	30000
Per capita expenditure (Malawi kwacha)	180000	180000	170000	200000	180000
Explanatory variables					
Years of schooling	6.10	6.40	5.90	7.50	6.20
Age (years)	43.00	43.00	48.00	25.00	43.00
Household size	5.20	5.10	5.50	3.70	5.10
Distance to road (km)	7.60	7.70	7.50	8.10	7.70
Distance to population center (km)	29.00	30.00	30.00	30.00	30.00
Distance to parastatal market (km)	7.80	7.80	7.80	7.80	7.80
Distance to auction market (km)	65.00	66.00	65.00	69.00	66.00
Distance to district center (km)	58.00	25.00	42.00	41.00	41.00
Distance to an international border post (km)	23.00	58.00	41.00	40.00	40.00
Distance to nearest secondary school (km)	3.50	3.50	3.50	3.70	3.50
Household is poor	0.33	0.29	0.33	0.23	0.31
Male household head	0.78	0.77	0.76	0.83	0.78
Born in other town or urban center in this district	0.01	0.01	0.01	0.01	0.01
Born in other village in this district	0.20	0.24	0.22	0.22	0.22
Born outside Malawi	0.04	0.03	0.04	0.01	0.03
Born in this town or urban center	0.01	0.01	0.01	0.01	0.01
Born in this village	0.43	0.41	0.41	0.45	0.42
Born in town or urban center in other district	0.06	0.04	0.05	0.05	0.05
Born in village in other district	0.25	0.27	0.26	0.24	0.26
Household is in rural area	0.74	0.75	0.75	0.74	0.74
Literate in Chichewa language	0.70	0.75	0.70	0.82	0.72
Literate in English language	0.43	0.46	0.42	0.54	0.44

Notes: Cohort is based on age. Old cohort is greater than 30 years while young cohort is less than 30 years.

About 72% of the household head can read and write in the local language (Chichewa) while only 44% are literate in English. Most of the households in the sample (74%) reside in rural areas and live in the same village (43%) they born in. Most of the household live in rural areas (74%) and a longer distances from roads, district center, population center, and agricultural markets.

4. Results and discussion

4.1 Estimation results

In this section, I present first the results of the first stage regressions to explain the identification strategy that I have followed. Then the pooled OLS and pooled 2SLS estimates are compared with a discussion on the implications of the differences in the estimates. This is followed by a presentation of results for fixed effects and fixed effects 2SLS estimates. The final subsection presents the test for selection in a fixed effects 2SLS framework and a consistent correction for selection using pooled 2SLS.

(a). Effect of distance and 1994 policy education policy changes on schooling attainment

Figure 2 summarizes the identification strategy followed in the paper. The validity of the instruments depends on relevance and exclusion restriction. The identification of the schooling effects comes from the impact of the 1994 fees removal which affected younger cohorts but not older cohorts and that those closer to a secondary school have higher levels of schooling than those far away.

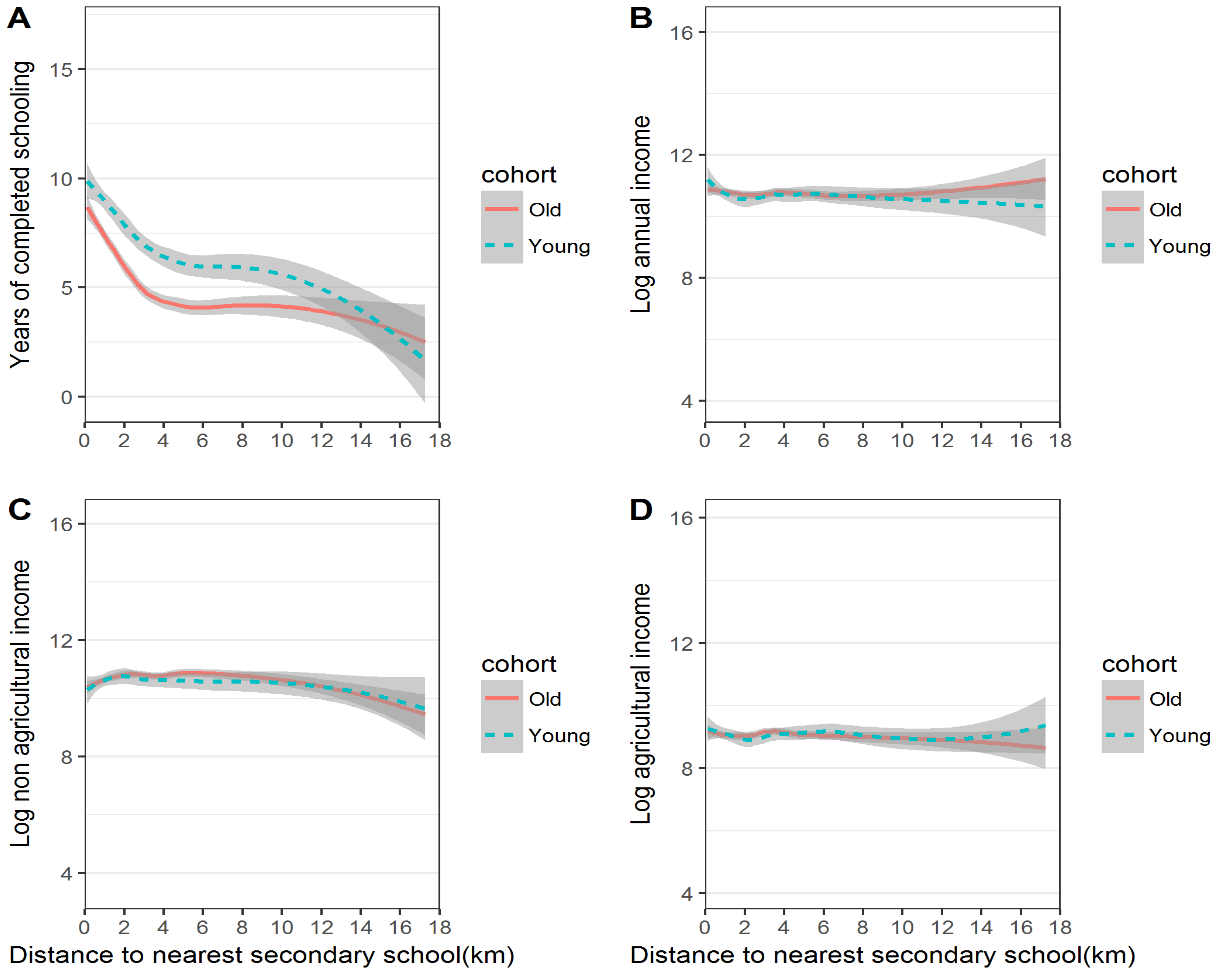

Figure 2: Years of education on y axis, distance to school x, legend: age cohort (2010)

Notes: Panel A shows the relationship between years of schooling and distance to nearest secondary school by cohorts. Younger cohorts have more educational exposure than older cohorts but both fall with increasing distance to secondary school. Panel B shows non-differential effects of distance to nearest secondary school on total income across all cohorts . Panel C and D shows similar trends for non-agricultural and agricultural incomes . Similar pattern is observed for 2013 data.

This is evidenced by panel A. Younger cohorts have higher years of schooling if residing closer to secondary schools. This variation allows identification of schooling effects on income since cohorts and distance to secondary schools do not affect incomes thus any effects of schooling is likely causal (Panel B). Panel C and D demonstrate the selection issues that may arise because non-agricultural incomes are higher than agricultural incomes. The figure 3 and table 6 in the appendices shows the relevance of distance to schooling and cohort as instruments for schooling.

The test for weak instruments has F-statistic of 12.523 with a p-value of 0.000408 implying the instruments are not weak.

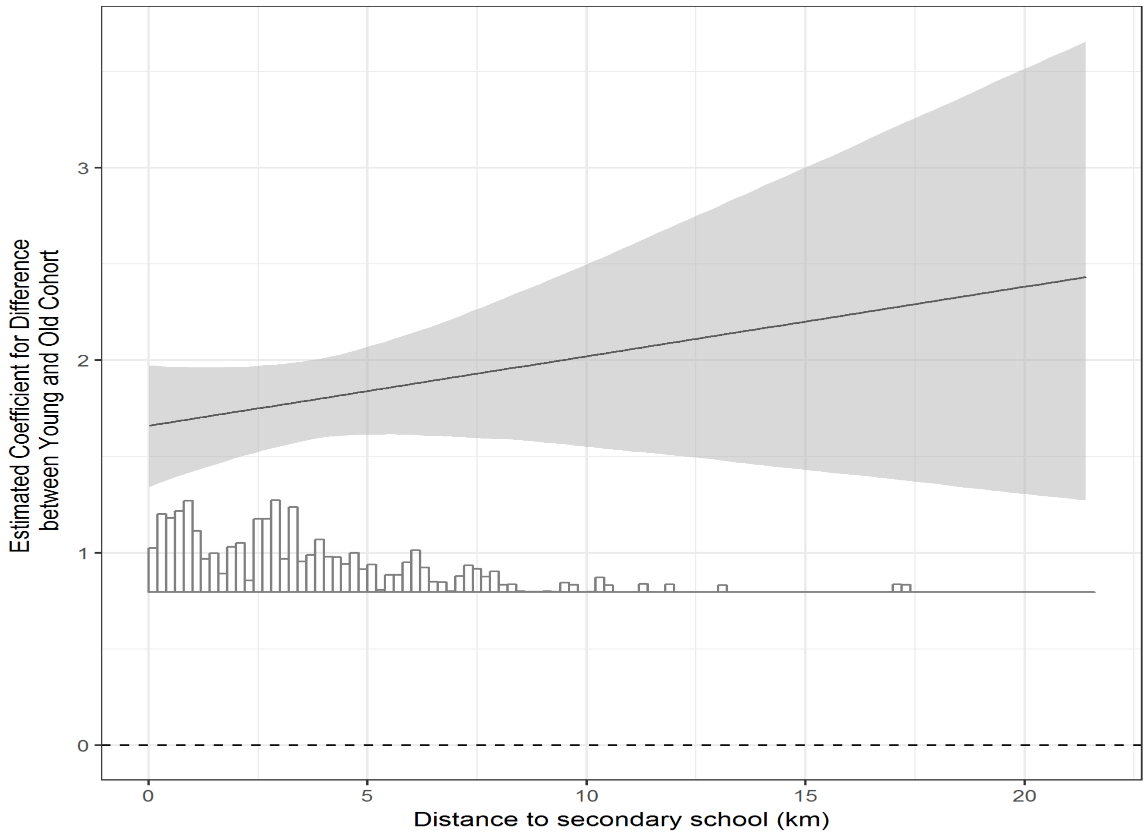

Figure 3: Estimated difference in schooling by cohort and distance to secondary school

Notes: The bars inside show the histogram of the distance to schooling variable. The shaded bands show the 95% confidence interval while the solid line is the marginal effect of cohort on schooling controlling for a battery of controls (e.g. distance to road).

(b). Pooled ordinary least squares and pooled two-stage least squares results

To benchmark the results in the next sections, I present the pooled OLS and pooled 2SLS estimate in table 2. The pooled OLS results in columns (1-3) show that schooling positively affects both agricultural and non-agricultural incomes. An additional one year in schooling increases annual incomes by 7%.

Table 2: Pooled OLS and 2SLS estimates

	Log Household Income *OLS* (1)	Log Agricultural Income *OLS* (2)	Log Non Agricultural Income *OLS* (3)	Log Household Income *2SLS* (4)	Log Agricultural Income *2SLS* (5)	Log Non Agricultural Income 2SLS (6)
Schooling(Years)	0.06***	0.07***	0.07***	0.27	0.05	0.28**
	(0.01)	(0.01)	(0.01)	**(0.18)**	**(0.12)**	**(0.13)**
Age	0.02**	0.05***	0.02**	-0.01	0.05***	0.03**
	(0.01)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)
Age squared	-0.00***	-0.00***	-0.00**	0.00	-0.00***	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	9.31***	6.93***	9.37***	7.67***	7.01***	7.27***
	(0.35)	(0.64)	(0.38)	(1.59)	(1.03)	(1.00)
Observations	6,235	3,763	4,859	6,235	3,763	4,859
R^2	0.17	0.19	0.20	-0.03	0.19	0.01

Notes: Standard errors in parentheses are clustered at enumeration area level and based on 1000 bootstraps. Control variables include: district dummies, place of birth dummies, distance to road, household size, survey year, and gender of the household head.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

The estimates for the 2SLS estimates are different from OLS estimates and not precise based on the cluster bootstrapped standard errors. The estimates are positive and different from zero for non-agricultural incomes. This implies that pooled OLS estimates are biased downward in the non-agricultural sector and upward in the agricultural sector.

(c). Fixed effects and fixed effects-two stage least squares estimates

Table 3 presents household fixed effects and fixed effects two stage least squares regressions for the three dependent variables- log household income, log agricultural income and log non-agricultural income. Column (1) presents FE estimates for log household income. Each year of schooling increases household income by 5%. This is lower than pooled OLS estimates because it expected that household level characteristics farm size and soil quality are positively correlated with incomes such that pooled OLS results are biased upwards. In column (2), the coefficient on schooling is negative and no evidence it is different from zero when FE-2SLS approach is used. Overall, fixed effects estimates are significant for all dependent variables. This is against evidence by Satriawan & Swinton (2007) who find that controlling for unobserved heterogeneity renders most education variables insignificant particularly in the farm sector. However, with instrumental variables included the estimates are unstable. The reason for this may be that the instruments do not vary as much across the two panel years (since the different between first panel and second panel is only 3 years, not many household migrate or move across age cohorts). According to Wooldridge (2010, p.354), FE-2SLS works only when instruments vary overtime. Using Mundlak terms in a correlated random effect model would help resolve this problem as will be shown in the next subsection.

Table 3: Household fixed effects and fixed effects-two stage least squares regressions

	Log Household Income (1)	Log Agricultural Income (2)	Log Non Agricultural Income (3)	Log Household Income (4)	Log Agricultural Income (5)	Log Non Agricultural Income (6)
Schooling(Years)	0.05***	0.03*	0.05***	-0.05	-0.29***	0.10
	(0.01)	(0.02)	(0.02)	(0.07)	(0.10)	(0.11)
Household fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Instruments	No	No	No	Yes	Yes	Yes
N	6,237	3,764	4,860	6,235	3,763	4,859
R^2	0.04	0.20	0.08	0.03	0.08	0.08

Notes: Standard errors in parentheses are based on pair wise panel clustered bootstraps. Control variables include: district dummies, place of birth dummies, distance to road, household size, survey year, gender of the household head

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

(d). Selection into non-agricultural activities and schooling

Table 7 in the appendix presents probit regressions for the incidence of agricultural incomes and non-agricultural incomes. Table 4 shows the estimation results from testing and correcting for selectivity. Table 4 shows that returns to schooling estimates are positive under various model specifications that correct for selection bias and endogeneity. With selection included, pooled

OLS results on schooling are significant but insignificant both with and without Mundlak terms (time avarages on time variant independent variables). In summary, the results illustrate that when endogeneity due to unobserved heterogeneity and selection bias are corrected, a year of schooling increases agricultural incomes by 3-4% slightly lower than pooled OLS estimates (7%)

Table 4: Test and correction for selectivity in agricultural income regressions

	Log (Agricultural income)			
	(1)	(2)	(3)	(4)
Years of schooling	0.06***	0.04**	0.04	0.03
	(0.01)	(0.02)	(0.09)	(0.09)
IMR	-0.80***	-0.64***	0.60**	0.54*
	(0.09)	(0.19)	(0.27)	(0.28)
Constant	7.77***		5.81***	5.94***
	(0.47)		(0.75)	(0.72)
Household fixed effects	No	Yes	No	No
Other controls	Yes	No	Yes	Yes
Instruments	No	No	Yes	Yes
Mundlak means	No	No	No	Yes
N	3,761	3,761	3,761	3,761
R^2	0.15	0.21	0.21	0.21

Notes: Standard errors in parentheses are clustered at enumeration area level and from 1000 bootstraps. Control variables include: district dummies, place of birth dummies, distance to road, household size, survey year, and gender of the household head.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

5. Research and policy implications

Schultz urged relentlessly in a series of articles that human capital is usually underrated for poverty reduction. It is usually difficult to collect income information in agricultural, self and informal employment as such the results so far presented may just be affected by measurement error on the income variable which may lead to attenuation bias. As a robustness check I estimate a set of models with log per capital expenditure as the dependent variable (expenditure is a proxy for household income). This allows analysis of the effects of schooling on poverty.

Table 5: Schooling and per capita expenditure

	Dependent variable: Log (Per capita expenditure)					
	(1)	(2)	(3)	(4)	(5)	(6)
Years of schooling	0.07***	0.14***	0.05***	0.19***	0.20***	0.18***
	(0.002)	(0.02)	(0.003)	(0.02)	(0.06)	(0.02)
Household fixed effects	No	No	Yes	Yes	Yes	Yes
Instruments	No	Yes	No	Yes	Yes	Yes
Other Controls	Yes	Yes	No	No	No	No
N	7,774	7,772	7,776	7,774	1,993	5,781
R^2	0.43	0.31	0.05	0.05	0.04	0.05

Notes: Column 5 is for urban subset the data while column 6 is for rural subset of the data.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Here the story is consistent that schooling positive affects per capita expenditure. One additional year of schooling results in 14-20% increase in per capita across all specifications, rural and urban areas. The estimates are also precisely estimated. Though the theory on agricultural household models is rooted on considering profits as the key dependent, researchers need to be cautious in the case of development countries. There is also a need for better ways of collecting income data and behavioral theoretical models of the agricultural household that link expenditure to incomes.

The limitation of the paper is that it does not address the mechanisms through which schooling affects agricultural incomes. Is it due to worker, entrepreneurial or allocative effect? Each of these mechanisms may point to different policy implications which based on the analysis presented in the paper would be far reaching. For example, if the agricultural returns to schooling are due to worker effect, it would be prudent to target schooling initiatives to children of farm people who are working on the farms. Future research should consider deciphering these mechanisms. On the methodological side, while using recently available shapefiles for secondary school locations is novel; it is not as precise because the spatial data does not show when each of the schools was opened.

6.0 Conclusion

The paper has investigated the effects of schooling on agricultural incomes using panel data from Malawi. The empirical contribution is the use of recent panel data methods developed by Semykina & Wooldridge (2010) for correcting for selection bias, endogeneity and unobserved heterogeneity. This study is important for Malawi and other developing countries because there are competing advocacies of whether government should invest more in education or in the agricultural sector. I argue in this paper that these are non-separable decisions. Previous literature in other developing countries demonstrates the importance of other policies especially in agriculture that allow an increase in the returns to schooling. For instance, Foster & Rosenzweig (1996) based on findings in India suggest that policies resulting in greater technical change are complementary with those increasing investment in schooling. In the case of Malawi, it can be argued that the farm input subsidy program may have increased the agricultural returns to schooling.

The increasing number of papers suggesting the importance of the rural non-farm sector to poverty reduction suggests a shift of directing resources in the future from agriculture to services and industry. The findings in this paper provide an alternative investment strategy- education in basic and productive skills. Human capital through schooling is key to reducing both rural and urban poverty just as the conventional sectors-agriculture, services and industry. This claim is consistent with World Bank recent recommendations on the future of Africa's economic future (World Bank 2017). It is common for policy makers to push for agricultural or employment based interventions to address rural poverty- the neglected sector in this quest is education which can massively pull rural people out of poverty and allow future growth of rural economies.

References

Abadie, A. et al., 2017. When Should You Adjust Standard Errors for Clustering? Available at: http://arxiv.org/abs/1710.02926.

Becker, G.S., 1994. *Human capital: a theoretical and empirical analysis with special reference to education*,

Benson, T., Erman, A. & Baulch, B., 2018. Change and Rigidity in Youth Employment Patterns in Malawi. , (Nso 2014), pp.1–24.

Card, D., 2001. Estimating the Return to Schooling: Progress on Some Persistent Econometric

Problems. *Econometrica*, 69(5), pp.1127–1160.

Chirwa, E.W. & Matita, M.M., 2009. *The Rate of Return on Education in Malawi*, Available at: http://www.eldis.org/vfile/upload/1/document/1105/The Rate of Return on Education in Malawi.pdf.

Fafchamps, M. & Quisumbing, A.R., 1999. Human Capital , Productivity , and Labor Allocation in Rural Pakistan. *Journal of Human Resources*, 34(2), pp.369–406.

Foster, B.A.D. & Rosenzweig, M.R., 1996. Technical Change and Human-Capital Returns and Investments : Evidence from the Green Revolution. *American Economic Journal: Microeconomics*, 86(4), pp.931–953.

Gondwe, A.S., 2016. *An empirical analysis of poverty, inequality and the labor market in Malawi*,

Hansen, G.D. & Prescott, E.C., 2000. Malthus to Solow. *American Economic Review*, (1205–1997).

Heyneman, S.P., 1980. *The Evaluation of Human Capital in Malawi*, Washington D.C.

Jäckle, R. & Himmler, O., 2010. Health and Wages: Panel Data Estimates Considering Selection and Endogeneity. *Journal of Human Resources*, 45(2), pp.364–406. Available at: http://www.socialpolitik.de/tagungshps/2007/paper/Jaeckle.pdf.

Matita, M. & Chirwa, E., 2009. *The impact of education on self-employment, farm activities and household incomes in Malawi*, Available at: http://41.87.6.35:8080/xmlui/handle/123456789/175.

Mundlak, Y., 1961. Empirical Production Function Free of Management Bias. *Journal of Farm Economics*, 43(1), pp.44–56.

Mundlak, Y., 1978. On the pooling of time series and cross section data. *Econometrica*, 46(1), pp.69–85.

NSO, 2014. *Integrated Household Panel Survey 2010-2013: Household Socio-Economic Characteristics Report*,

NSO, 2012. *Malawi Third Integrated Household Survey (IHS3) Basic Information Document*,

Pudasaini, S.P., 1983. The Effects of Education in Agriculture : Evidence from Nepal. *American Journal of Agricultural Economics*, 65(3), pp.509–515.

Reimers, M. & Klasen, S., 2013. Revisiting the role of education for agricultural productivity. *American Journal of Agricultural Economics*, 95(1), pp.131–152.

Satriawan, E. & Swinton, S.M., 2007. Does human capital raise farm or nonfarm earning more ? New insight from a rural Pakistan household panel. , 36, pp.421–428.

Schultz, T.W., 1964. *Transforming traditional agriculture*. University of Chicago Press. Chicago.

Schultz, T.W., 1980. Nobel Lecture : The Economics of Being Poor. *Journal of Political Economy*, 88(4), pp.639–651.

Semykina, A. & Wooldridge, J.M., 2010a. Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, 157(2), pp.375–380. Available at: http://dx.doi.org/10.1016/j.jeconom.2010.03.039.

Semykina, A. & Wooldridge, J.M., 2010b. Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, 157(2), pp.375–380. Available at: http://dx.doi.org/10.1016/j.jeconom.2010.03.039.

Tamura, R., 2002. Human capital and the switch from agriculture to industry. , 27, pp.207–242.

Taylor, J.E. & Yunez-Naude, A., 2000. The Returns from Schooling in a Diversified Rural Economy. *American Journal of Agricultural Economics*, 82(2), pp.287–297. Available at: http://www.jstor.org/stable/1244651%5Cnhttp://www.jstor.org.proxy2.cl.msu.edu/stable/pdfplus/1244651.pdf?acceptTC=true.

Wantchekon, L. & Stanig, P., 2016. *The curse of good soil? Land fertility, roads, and rural poverty in Africa*,

World Bank, 2018. *World Development Report 2018: LEARNING to Realize Education's Promise*,

WorldBank, 2017. *Africa's Pulse: An analysis of issues shaping africa's economic future*, Washington D.C.

Wouterse, F. & Taylor, J.E., 2008. Migration and Income Diversification : Evidence from Burkina Faso. *World Development*, 36(4), pp.625–640.

Xu, Z. et al., 2009. Factors influencing the profitability of fertilizer use on maize in Zambia. *Agricultural Economics*, 40(4), pp.437–446.

Yang, D.T., 2004. Education and allocative efficiency : household income growth during rural reforms in China. *Journal of Development Economics*, 74, pp.137–162.

Yang, D.T., 1997a. Education and off-farm work. *Economic Development and Cultural Change*, 45(3), pp.613–632.

Yang, D.T., 1997b. Education in Production: Measuring Labor Quality and Management.

American Journal of Agricultural Economics, 79(3), pp.764–772.

Yang, D.T. & An, Y.M., 2002. Human capital , entrepreneurship , and farm household earnings. *Journal of Development Economics*, 68, pp.65–88.

Appendices

Table 6: First stage regression results

	Dependent variable: Years of schooling			
	(1)	(2)	(3)	(4)
Distance to school	-0.42***	-0.14***	-0.13***	-0.12***
	(0.02)	(0.02)	(0.02)	(0.02)
Cohort (Young)	1.75***	0.40**	0.44***	0.47**
	(0.11)	(0.16)	(0.16)	(0.19)
Distance to school x Cohort				-0.01
				(0.03)
Birthplace dummies	No	No	Yes	Yes
Controls	No	Yes	Yes	Yes
N	7,774	7,774	7,774	7,774
R^2	0.10	0.33	0.35	0.35
F Statistic	452.99***	92.64***	86.61***	84.79***

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 7: Probit selection equations on agricultural revenue participation

	Dependent variable: Agricultural revenue participation	
	2010	2013
	(1)	(2)
Schooling	0.004	-0.01*
	(0.01)	(0.01)
Age	0.02**	0.04***
	(0.01)	(0.01)
Age squared	-0.0002***	-0.0003***
	(0.0001)	(0.0001)
Household size	0.11***	0.07***
	(0.02)	(0.02)
Distance to road	0.02*	0.01
	(0.01)	(0.01)
Time mean of household size	-0.04*	-0.02
	(0.02)	(0.02)
Time mean of age	0.004	-0.0004
	(0.004)	(0.004)
Time mean of distance to road	-0.02	-0.004
	(0.01)	(0.01)
Male household head	0.24***	0.20***
	(0.06)	(0.06)
Constant	-1.06***	-1.05***
	(0.35)	(0.35)
N	3,886	3,886
Log Likelihood	-2,055.76	-2,201.79
Akaike Inf. Crit.	4,201.52	4,495.59

Notes: All regressions include district and birthplace dummies

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.