Online Appendix for "Land Misallocation and Productivity"

Chaoran Chen Diego Restuccia Raül Santaeulàlia-Llopis[†]

July 2022

A Data

The Malawi Integrated Survey of Agriculture (Malawi-ISA) is part of a new generation of household surveys funded by the Bill & Melinda Gates Foundation (BMGF) and led by the Living Standards Measurement Study (LSMS) team in the Development Research Group (DECRG) of the World Bank to improve the quality and policy relevance of household-level data on agriculture in Sub-Saharan Africa. The Malawi ISA (National Statistical Office, 2012) incorporates an extended and comprehensive agricultural questionnaire on agricultural production and factor inputs, including land quality and rain.

Land size and land quality. We measure household land as the sum of the size (in acres) of each household's plot used for cultivation. We include rented-in land in household land size. For the vast majority of plots, the acres per plot are recorded using GPS with precision of 1% of an acre. For the remaining plots, size is self-reported with an estimate from the household. This leaves virtually no room for error in our measure of land input, see a detailed assessment in Carletto, Savastano and

[†]Chen: York University, Department of Economics, 4700 Keele Street, Toronto, ON M3J 1P3, Canada (chenecon@yorku.ca); Restuccia: University of Toronto and NBER, Department of Economics, 150 St. George Street, Toronto, ON M5S 3G7, Canada (diego.restuccia@utoronto.ca); Santaeulàlia-Llopis: University of Pennsylvania, Universitat Autonoma de Barcelona, BSE and CEPR, Plaça Cívica s/n, Bellaterra, Barcelona 08193, Spain (rauls@movebarcelona.eu).

Zezza (2013). The data also contain detailed information on the quality of land for each plot used in each household. We consider all 11 dimensions of land quality available: elevation, slope, erosion, soil quality, nutrient availability, nutrient retention capacity, rooting conditions, oxygen availability to roots, excess salts, topicality, and workability. The slope (in %) and elevation (in meters) are continuous variables while the rest of land quality variables are categorical such as terrain roughness (plains, lowlands, plateaus, hills, mountains), erosion (1 none, 2 low, 3 moderate, 4 high), nutrient availability, nutrient retention, rooting conditions, oxygen to roots, excess of salts, toxicity and workability (1 constraint, 2 moderate constraint, 3 severe constraint and 4 very severe constraint). These measures are largely from geographical information system such as the Harmonized World Soil Database.

Our benchmark land quality index is defined per household as the predicted value of output (net of rain effects which we discuss next) generated by the joint behavior of all dimensions of land quality controlling for capital and land size, see the first column in Table A-1. We also explore alternative definitions of the land quality index in Table A-1 that depend on the number of land quality dimensions that we incorporate and on the way we control for capital and land. A reassuring aspect of our land quality index, defined from physical measures (e.g., erosion, soil quality, etc.), is that it is positively related to land prices, see Table A-2. Finally, we perform a robustness analysis of our reallocation results with respect to our entire set of land quality indexes without substantial changes in our findings, see Table A-3.

Rain. It is important to control for unanticipated temporary output shocks that can contribute to explain the variation in output and productivity across households in the data. Rain shocks are among the most important shocks in agriculture. We use the annual precipitation which is the total rainfall in millimetres (mm) in the last 12 months. Our benchmark measure of output (value added) is net of the rain effects. Specifically, we group observations into 10 bins sorted by their observed level of rain, and then regress the (log) value added on rain deciles to net the effect of rain

Table A-1: Land Quality Index and Its Dimensions

	Land Quality Index					
	Benchmark	Alternative Definitions				
Dimensions:	q_0	q_1	q_2	q_3	q_4	q_5
Elevation	✓	✓	✓	✓	✓	✓
Slope	✓	✓	✓	✓	✓	✓
Terrain roughness	✓	–	✓	–	✓	–
Erosion	✓	–	✓	–	✓	–
Nutritient availability	✓	–	✓	–	✓	–
Nutritient retention	✓	–	✓	–	✓	–
Rooting conditions	✓	–	✓	–	✓	–
Oxygen to roots	✓	–	✓	–	✓	–
Excess salts	✓	–	✓	–	✓	–
Toxicity	✓	–	✓	–	✓	–
Workability	✓	–	✓	–	✓	–
Additional controls:						
Capital	✓	✓	–	–	✓ (Const.)	✓ (Const.)
Land size	✓	✓	–	–	✓ (Const.)	✓ (Const.)

Notes: Summary definitions of different measures of land quality index. Our benchmark measure utilizes all 11 dimensions of land quality in addition to elevation and slope, controlling for capital and land size. For alternative measures q_4 and q_5, we control for capital and land size but restrict the coefficients to be identical to capital and land shares. Data from the Malawi ISA 2010-11 (National Statistical Office, 2012).

shocks. As rain might be more relevant in some months than others we also tried to control for an alternative measure of rain, the wettest quarter within the last 12 months. We find an output gain of 2.88, similar to 2.82 in our baseline.

Labor. In Malawi, not only the household head but also a large proportion of the households members, which average 4.6 per household, contribute to agricultural work. The household head is identified as the person who makes economic decisions in the household (e.g., use of production or transfers). We define household members as individuals that have lived in the household at least 9 months in the last 12 months. These household members potentially include family (e.g.

Table A-2: Land Quality Index and Land Price

	Land Quality Index					
	Benchmark	Alternative Definitions				
	q_0	q_1	q_2	q_3	q_4	q_5
Correlation with land price	0.189	0.200	0.202	0.204	0.196	0.196

Notes: Rank correlation between the land quality index q and the price of land computed as the self-reported estimated land value (under the hypothetical scenario in which the owner sells the land). We separately calculate this correlation for our benchmark measure of land quality and 5 alternative measures of land quality defined in Table A-1. The correlation is significant at the one percent level for all land quality measures. Data from the Malawi ISA 2010-11 (National Statistical Office, 2012).

Table A-3: Output Gain with Different Land Quality Indexes

	Benchmark	Alternative Definitions				
	q_0	q_1	q_2	q_3	q_4	q_5
Output gain	2.82	2.84	2.83	2.84	2.82	2.84

Notes: Output gain associated with our benchmark measure of land quality and 5 alternative measures of land quality defined in Table A-1. Data from the Malawi ISA 2010-11 (National Statistical Office, 2012).

children, spouses, siblings, and parents) and also non-relatives (e.g. lodgers and servants). Individual information about each household member's (including children) extensive and intensive margins of labor supply is collected: (i) weeks worked, (ii) days per week, and (iii) hours per day, by plot and by agricultural activities covering the entire agricultural production. For the hired labor and free/exchange labor, we also observe number of days worked by men, women and children by plot and activities. The detailed information on individual agricultural labor days through the entire year avoids the seasonal component of labor supply; that is, we do not rely on data on labor supply related to 'last week/month' behavior. Our benchmark measure of household labor supply is aggregate days of all individuals (household members and non-members) supplied in all plots cultivated by the household in the rainy season. To control for human capital, i.e., the fact that not all hours might contribute the same to agricultural production, we construct household efficiency units by weighting individual hours using the wages of hired labor by age and sex groups as weights. We find that our results are robust to this alternative specification.

Capital equipment and structures. Agricultural capital equipment includes implements (i.e., hand hoe, slasher, axe, sprayer, panga knife, sickle, treadle pump and watering can) and machinery (e.g. ox cart, ox plough, tractor, tractor plough, ridger, cultivator, generator, motorized pump, grain mill, and others). Agricultural capital structures includes chicken houses, livestock kraals, poultry kraals, storage houses, granaries, barns, pig sties, etc. To measure the capital stock per item we use the estimated current selling price of capital items after conditioning on its use. We construct the household agricultural capital stock by aggregating across all agricultural items. The use of the selling price (not available in previous LSMS data) avoids the cumbersome perpetual inventory method adjustment for the age of capital to impute current value from the value at the time of purchase which requires recalling and depreciation assumptions by asset's age. We note that we observe a small set of farmers who have zero measured capital but report cultivated land and positive production. This may be because the data do not record a common set of very small tools and structures used by farmers. We hence follow Adamopoulos et al. (2022) and impute an amount of capital to all farms representing the value of this set of small tools and structures, with the value equal to 10% of the median of the calculated capital value.

Trimming strategy. The cost of misallocation summarized by the output gain is known to be sensitive to extreme values of inputs and outputs. We trim our sample to exclude apparent extreme values. Specifically, we trim the top and bottom 0.5% of each of output, land, capital, and estimated farm TFP. This trimming strategy substantially reduces measured dispersion of farm TFP by between 9 to 16 percent: the variance of log farm TFP shrinks from 1.67 to 1.40 in the 2010-11 cross-section and from 0.96 to 0.87 in the panel sample. In the context of the misallocation literature, trimming is also potentially a conservative strategy if high productivity units should in fact be allocated more inputs.

Panel data and farmer ability. We use the panel structure of the data to estimate fixed-effect farm productivity s_i. There are in total 608 farm households who appear in the 2010-11 wave (the benchmark year for the quantitative analysis) and one of the 2013 re-interview, and 2016-17 and 2019-20 waves (National Statistical Office, 2020). Among them, 410 households appear in two waves, 175 households appear in three waves, and 23 households appear in all four waves.

We argue that the fixed-effect measure of farm productivity captures farmer's ability rather than location characteristics or transitory shocks. To illustrate this issue, we regress fixed-effect farm productivity $\log(\bar{s}_i)$ on observed location and farm characteristics and report the resulting coefficients with standard errors in parenthesis in Table A-4.

Table A-4: Fixed-Effect Farm Productivity and Observables

Dependent variable	Fixed-effect farm productivity $\log(\bar{s}_i)$
Location and farm observables:	
Log distance to road	−0.06
	(0.05)
Log distance to population center	−0.02
	(0.11)
Log rainfall	0.46
	(0.44)
Health of household head	0.17
	(0.42)
Log age	−0.55
	(0.23)
Schooling of household head	0.56
	(0.21)
Number of observations	596

Notes: Regression coefficients of fixed-effect farm productivity on location and farm observables such as the log distance to the nearest road and population center, the log rainfall of 2010, a dummy variable indicating household head's health status, the log age of household head, and a dummy variable indicating household head's education status. Standard errors are reported in parenthesis.

We find that farm productivity is not correlated with the log distance to the nearest road or population center, which suggests that farm productivity does not capture location characteristics such as access to markets. We also consider two transitory shocks: the log of rainfall in 2010 and

a dummy indicator which equals to one if the household head was admitted to a hospital in 2010 and zero otherwise. As expected, the coefficients on these two variables are insignificant, which is reassuring given that our measure of farm productivity was constructed devoid of transitory shocks. To illustrate that our measure of farm productivity captures the ability of farmers, we consider two variables measuring non-transitory components of ability such as the log of the household head's age and a dummy indicator which equals to one if the household head ever attended school. We indeed find significant coefficients for these two variables implying that old farm operators are less productive in farming than young farm operators; and that educated farm operators have higher productivity than uneducated farm operators.

To the extent that our measure of farm productivity may still reflect other factors than farmer ability (mismeasurement), we note that the change in the output gain from an efficient reallocation is roughly proportional to the change in the dispersion of farm productivity. This implies that, for example, if dispersion of farm productivity is only one half our measured dispersion, then the output gain from an efficient reallocation of 96 percent in our baseline panel data would only be half, 48 percent.

B Measurement Error and Misallocation

We assess the extent to which our results may be affected by measurement error through two approaches. First, we explore the panel dimension of the data to estimate household-farm fixed effects of productivity and inputs that abstract from time and transitory variation, including potential measurement error. Second, we explore additional counterfactual experiments to provide bounds on the relevance of measurement error for output gains.

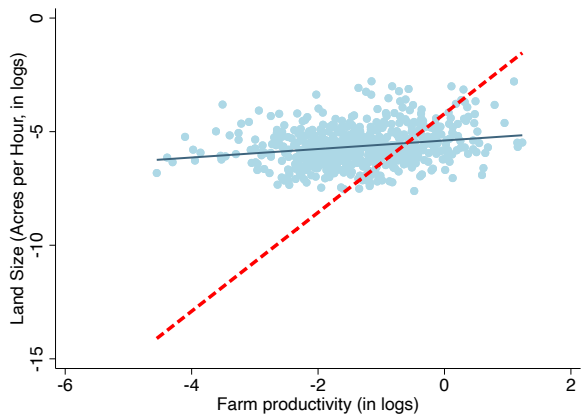
B.1 Patterns of Misallocation with Panel Data

Recall that the patterns of misallocation illustrated in Figure 1 in the article are characterized with the 2010-11 wave of cross-sectional data. We note that these patterns remain remarkably similar if we instead use the panel sample. Figure B-1 illustrates the patterns using the panel data to estimate a household-farm fixed effect of productivity. The correlation of farm productivity and inputs are very similar in the cross-section and panel data, for instance, the log correlation of land input and productivity is 0.17 in the cross-section and 0.21 in the panel, and similarly the log correlation of capital and productivity is 0.02 in the cross-section and also 0.02 in the panel.

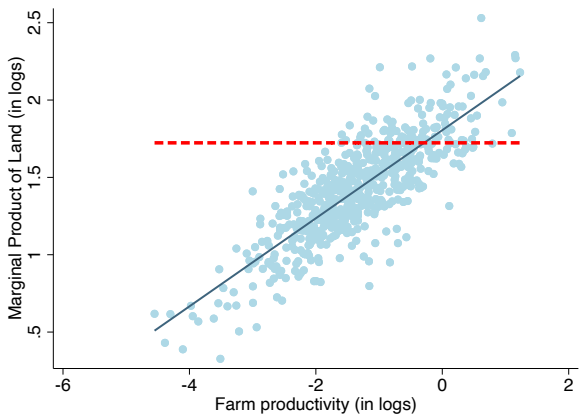
B.2 Recall Bias for Agricultural Production and Labor Input

In rural settings the underreporting of agricultural production is a recurrent issue for survey data (Deaton, 1997; de Magalhães and Santaeulàlia-Llopis, 2018). There are two aspects of the Malawi ISA design that help mitigate and study this issue. First, in many instances the survey provides internal consistency checks (e.g., households are asked total sales, and also sales by crop and by plot; the interviewer must check *in situ* that the two sums coincide or otherwise re-interview). Second, the ISA collects data not only on agricultural production but also on consumption that includes food consumption (in physical units) from own production. This provides a unique opportunity to externally validate agricultural production using consumption data. In this context, a reassuring result is that in rural household-farms that do not sell their agricultural production and have little or no consumption purchases (i.e., about 50% of the entire rural sample), the reported agricultural production and the reported consumption net of transfers imply very similar quantities, which suggests a small scope for measurement error (from recall or elsewhere) in agricultural production, see de Magalhães and Santaeulàlia-Llopis (2018).

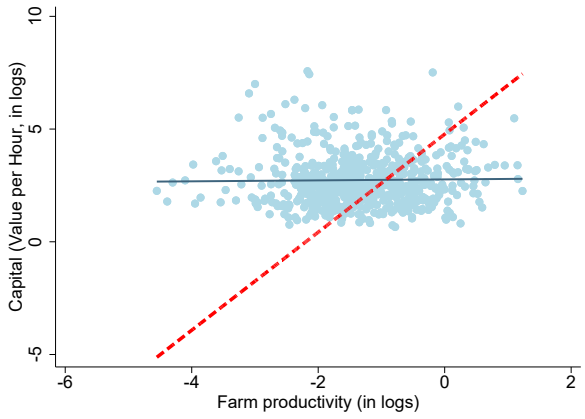
Not only agricultural production is collected retrospectively, but also labor input. To further inves-



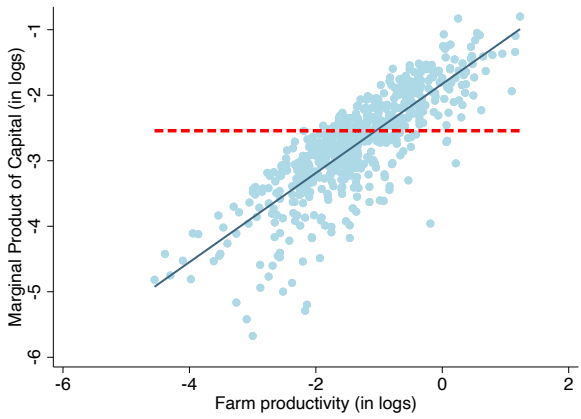
(a) Land Size vs. Farm Productivity



(b) MPL vs. Farm Productivity



(c) Capital vs. Farm Productivity



(d) MPK vs. Farm Productivity

Figure B-1: Patterns of Misallocation—Panel Specification

Notes: Panel (a) reports actual and efficient land operational size in farms ℓ_i with respect to farm productivity s_i. Panel (b) reports actual and efficient marginal product of land (MPL) with respect to farm productivity s_i. Panel (c) reports actual and efficient capital in farms k_i with respect to farm productivity s_i. Panel (d) reports actual and efficient marginal product of capital (MPK) with respect to farm productivity s_i. Each (blue) dot represents a household farm in the data whereas the (red) dashed line represents the efficient allocation. Farm productivity is the household-farm fixed effect while farm inputs are from the 2010/11 wave. All variables have been logged.

tigate the basis for potential recall bias in the collection of production and labor input, we note that in Malawi there is only one main harvest associated with the only rainy season. We then re-conduct our entire analysis for only the households farms that are surveyed shortly (i.e., within four months) after harvest, the output gain is 2.78-fold which is only slightly lower than the 2.82-fold output gain in our benchmark specification. This finding suggests that our results are robust to recall bias.

B.3 Bounds of Output Gains

We design the following experiment to better understand the nature of the output gain. Suppose a planner allocates the observed input sets $\{k_i\}$ and $\{l_i\}$ to farmers in a particular fashion, holding farm-level TFP unchanged, how large are the gains from reallocation? In other words, in this experiment, we hold the marginal distributions of $\{k_i\}$, $\{l_i\}$, and $\{s_i\}$ constant but we allow for, for instance, assigning l_i to an arbitrary farmer j.

The assignment with lowest possible output gain is positive assortative matching between inputs and farm productivity, which yields an output gain of 1.35-fold. The highest possible output gain is then obtained with negative assortative matching, which is 5.38-fold. Random assignment, which means $\{k_i, \ell_i\}$ are uncorrelated with $\{s_i\}$, yields an output gain of around 3.1-fold. Note that with finite sample (7,505 observations) the output gain associated with this random assignment varies with the particular draw of random numbers. This comparison highlights the importance of the correlation. Our baseline output gain, 2.82-fold, is not much lower than that of random assignment. This is exactly because in our data, the correlation between inputs and farm productivity is very low, as documented in Figure 1 in the article.

B.4 A Structural Interpretation of Measurement Error

We estimate a structural model of measurement error to assess their potential importance in our quantitative results. We denote true capital, land, and output as k_i, ℓ_i, and y_i. Capital and land inputs are functions of true productivity z_i:

$$\ln(k_i) = \zeta^k \ln(z_i) + d_i^k, \quad \ln(\ell_i) = \zeta_\ell \ln(z_i) + d_i^\ell.$$

An efficient allocation implies $\zeta^k = \zeta^\ell = 1$ and $d_i^k = d_i^\ell = 0$. Hence $\zeta^k \neq 1$ and $\zeta^\ell \neq 1$ indicate correlated distortions as in Bento and Restuccia (2017), while $d_i^k \neq 0$ and $d_i^\ell \neq 0$ indicate non-systematic distortions.

We assume that variables (inputs and outputs) may be observed with error, i.e.,

$$\begin{aligned}
\ln(\tilde{k}_i) &= \ln(k_i) + \varepsilon_i^k = \zeta^k \ln(z_i) + d_i^k + \varepsilon_i^k, \\
\ln(\tilde{\ell}_i) &= \ln(\ell_i) + \varepsilon_i^\ell = \zeta^\ell \ln(z_i) + d_i^\ell + \varepsilon_i^\ell, \\
\ln(\tilde{y}_i) &= \ln(y_i) + \varepsilon_i^y,
\end{aligned}$$

where ε^k, ε^ℓ, and ε^y represent (log) additive measurement error. The estimated farm productivity $\tilde{z}_i$ is then

$$\tilde{z}_i = \frac{\tilde{y}_i}{\left(\tilde{k}_i^\alpha \tilde{\ell}_i^{1-\alpha_i}\right)^\gamma} = z_i \frac{\exp(\varepsilon_i^y)}{\left(\exp(\varepsilon_i^k)^\alpha \exp(\varepsilon_i^\ell)^{1-\alpha}\right)^\gamma}.$$

We now contrast the output gain calculated from observed variables with that calculated from true variables. Note that true aggregate inputs and output are identical to observed ones since measurement errors are mean zero. The true output gain from reallocating resources is

$$e = \frac{Y^e}{Y^a} = \frac{\left(\sum_i z_i\right)^{1-\gamma} \left(K^\alpha L^{1-\alpha}\right)^\gamma}{\sum_i z_i^{1-\gamma} \left(k_i^\alpha \ell_i^{1-\alpha}\right)^\gamma},$$

while the measured output gain is

$$\tilde{e} = \frac{\tilde{Y}^e}{\tilde{Y}^a} = \frac{\left(\sum_i \tilde{z}_i\right)^{1-\gamma} \left(K^\alpha L^{1-\alpha}\right)^\gamma}{\sum_i \tilde{z}_i^{1-\gamma} \left(\tilde{k}_i^\alpha \tilde{\ell}_i^{1-\alpha}\right)^\gamma}.$$

To structurally estimate this framework, we make the parametric assumption that ε^k, ε^ℓ, and ε^y are all normally distributed with variance σ_m^2. The parameter σ_m^2 governs the precision in measurement and we assume that additive measurement error is of the same magnitude for all inputs and outputs. In addition, we assume d^k and d^ℓ, the idiosyncratic distortions, follow normal distributions with variance σ_k^2 and σ_ℓ^2, and true productivity follows a normal distribution with variance σ_s^2.

We use this framework to answer the following question: If true output gain e is only half of measured output gain $\tilde{e}$, what is the implied magnitude of measurement error? We estimate this framework, which consists of six parameters: $\{\sigma_k^2, \sigma_\ell^2, \sigma_m^2, \zeta^k, \zeta^\ell, \sigma_s^2\}$, to match the following five moments from Malawi micro data: the variances of *observed* capital and labor input, the correlations between farm productivity and capital/labor input, and the measured output gain. We also use the fact that true output gain e is half the measured output gain as our sixth moment to restrict the value of σ_m^2.

Considering our most conservative output gain associated with the fixed effects of inputs and productivity in the panel of 1.67-fold, half of this level renders a "true" output gain of 1.34-fold. We find that the magnitude of the measurement error must be huge, the estimated $\sigma_m^2 = 0.22$ and as a result, the variance of (log) observed land input $\tilde{\ell}_i$ must be almost two times larger than (log) true land input ℓ_i. Given that our measure of land input is cultivated land at the household level measured via GPS, we think this magnitude of measurement error is unlikely, but nevertheless our analysis helps frame the extent to which remaining measurement error may be driving reported reallocation gains.

References

Adamopoulos, Tasso, Loren Brandt, Jessica Leight, and Diego Restuccia. 2022. "Misallocation, selection, and productivity: A quantitative analysis with panel data from china." *Econometrica*, 90(3): 1261–1282.

Bento, Pedro, and Diego Restuccia. 2017. "Misallocation, Establishment Size, and Productivity." *American Economic Journal: Macroeconomics*, 9(3): 267–303.

Carletto, Calogero, Sara Savastano, and Alberto Zezza. 2013. "Fact or Artifact: The Impact of Measurement Errors on the Farm Size - Productivity Relationship." *Journal of Development Economics*, 103(C): 254–261.

Deaton, Angus. 1997. *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy.* The Johns Hopkins University Press.

de Magalhães, Leandro, and Raül Santaeulàlia-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: 350–371.

National Statistical Office. 2012. "Malawi Third Integrated Household Survey 2010-2011 [MWI-2010-IHS-III-v01-M]." World Bank Microdata Library [distributor] https://microdata.worldbank.org/index.php/catalog/1003, accessed July 5, 2022.

National Statistical Office. 2020. "Malawi Integrated Household Panel Survey 2010-2013-2016-2019 (Long-Term Panel, 102 EAs) [MWI-2010-2019-IHPS-v05-M]." World Bank Microdata Library [distributor] https://microdata.worldbank.org/index.php/catalog/3819, accessed July 5, 2022.