Assessing misallocation in agriculture: plots versus farms *

Fernando M. Aragón Diego Restuccia Juan Pablo Rud

June 2024

Abstract

We examine empirically whether the level of data aggregation affects the assessment of misallocation in agriculture. Using data from Ugandan farmers, we document a substantial discrepancy between misallocation measures calculated at the plot and at the farm levels. Estimates of misallocation at the plot level are much higher than those obtained with the same data but aggregated at the farm level. Even after accounting for measurement error and unobserved heterogeneity, estimates of misallocation at the plot level are extremely high, with potential nationwide agricultural productivity gains of 562%. Furthermore, we find suggestive evidence that granular data may be more susceptible to measurement error in survey data and that data aggregation can attenuate the relative magnitude of measurement error in misallocation measures. Our findings suggest caution in generalizing insights on measurement error and misallocation from plot-level analysis to those at the farm level.

JEL classification: O4, O13, Q12.
Keywords: misallocation, agriculture, measurement error, distortions.

*We thank the editor, three anonymous referees, Loren Brandt, and Richard Rogerson for useful comments. All errors and interpretations are our own. Restuccia gratefully acknowledges the support from the Canada Research Chairs program and the Bank of Canada Fellowship. Aragon gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada, grant 435-2018-0227. Rud gratefully acknowledges the support from the Economic and Social Research Council (UK Research and Innovation), Project ES/S014438/1. The views expressed herein are not necessarily those of the Bank of Canada and are the authors alone. *Contact information:* Aragón: Department of Economics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (e-mail: faragons@sfu.ca); Restuccia: Department of Economics, University of Toronto, 150 St. George Street, Toronto, ON M5S 3G7, Canada (e-mail: diego.restuccia@utoronto.ca); Rud: Department of Economics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK (e-mail: juan.rud@rhul.ac.uk).

1 Introduction

A growing literature documents a large dispersion in measures of marginal products of inputs across farms. This finding has been interpreted as evidence of factor misallocation. Under this interpretation, reallocation of agricultural inputs could lead to substantial efficiency gains and aggregate productivity growth (Restuccia and Rogerson, 2017). A concern with this interpretation is that some of the observed dispersion could reflect other factors, such as unobserved heterogeneity or measurement error. In this case, dispersion measures would overestimate the importance of misallocation in agriculture.

To address this potential concern, a promising approach is to use information from disaggregated data, at the plot level, to remove measurement error and unobserved heterogeneity from misallocation measures (Gollin and Udry, 2021). This approach exploits the assumption of efficient allocation of resources across plots within a farm, which implies that the marginal productivity of inputs should be equalized across plots, and, thus, any observed dispersion across plots within farms could be attributed to other sources than misallocation.

However, we do not know if the estimates of misallocation are affected by the level of data aggregation. This limitation is important because most of the existing evidence on misallocation in agriculture uses data at the farm (household) level (e.g., Chen et al., 2023). If aggregation matters, then the insights obtained using plot-level data might not be comparable to existing evidence and could lead to conflicting assessments of the magnitude and sources of misallocation.

We use data from Ugandan farmers to empirically show that the level of data aggregation matters and can lead to quantitatively different conclusions regarding factor misallocation in agriculture. To do so, we calculate and compare misallocation measures using data at the plot level and the same data aggregated at the farm level. Similarly to the recent macroeconomic literature, our main measure of misallocation is potential efficiency gains (henceforth, efficiency gains), i.e., the change in aggregate output that could be obtained if inputs were reallocated across production units according to a hypothetical efficient benchmark. We calculate efficiency gains by assuming reallocation of resources within different geographical units (villages, regions, and nationwide), and by using different estimates of the production function from previous studies.

We find that the efficiency gains at the plot level are extremely large, even after adjusting for measurement error. Using production function estimates from Gollin and Udry (2021), the calculated efficiency gains in Ugandan agriculture are 2,268% allowing for reallocation nationwide, and 311% assuming only reallocation within villages. After purging for measurement error and unobserved heterogeneity, nationwide efficiency gains remain extremely large at 562%. These estimates imply an extent of misallocation much greater than previously documented in the literature. As a comparison, previous studies using farm-level data from China, Ethiopia, and Malawi document nationwide efficiency gains ranging from 53% to 259% (Adamopoulos et al., 2022; Chen et al., 2022, 2023).

Furthermore, estimated efficiency gains for Uganda decrease by more than half when aggregating the same unadjusted data at the farm level. We emphasize that this result is not driven by changes in the calculation procedure, underlying data, or production function parameters which we keep unchanged; but just by the level of data aggregation. We obtain similar results using data from another country (Tanzania), and alternative measures of misallocation such as productivity dispersion and the elasticity of inputs with respect to total factor productivity across production units (i.e., input-productivity elasticity).

We interpret the discrepancy between measures of misallocation at the plot and farm levels as evidence that, when assessing misallocation in agriculture, the level of data aggregation matters. This finding casts doubts on the validity of extrapolating the insights obtained using plot-level analysis to the results obtained at the farm level.

This issue becomes apparent when assessing the role of measurement error in factor misallocation. For example, in a plot-level analysis of our same data, Gollin and Udry (2021) report that “Late-season production shocks, measurement error, and heterogeneity in inputs together account for as much as 70% of the variance in measured productivity. Since these are not susceptible to reallocation, our estimates for the aggregate productivity gains that could be attained from a reallocation exercise are correspondingly smaller" (p. 5). Furthermore, Gollin and Udry (2021) conclude that "[...] commonly used approaches in the literature overstate the dispersion of log TFP by about 100%. The gains from a hypothetical reallocation are thus correspondingly overstated by a factor of two or three" (p. 48).

Our results indicate that this conclusion is unfounded. A substantial proportion of the productivity dispersion at the plot level reflects the higher granularity of the data. We find that aggregating the data at the farm level alone (without any correction for mismeasurement) reduces productivity dispersion by 40%. Moreover, given the large estimates of plot-level productivity dispersion, even a substantial reduction in dispersion still leaves sizeable potential reallocation gains. In this case, even with a lower productivity dispersion at the plot level after correcting for mismeasurement, the efficiency gains range from 562% nationwide to 143% within villages. These magnitudes are quantitatively important and much higher than previous estimates in the literature.

Is there an appropriate level of analysis to understand misallocation in agriculture? Plot-level data can be extremely useful in many applications. For instance, plot-level data has been used to examine the efficiency of intra-household allocation (Udry, 1996; Shandal et al., 2022), gender differences in the adoption of technology and agricultural productivity (Slavchevska, 2015; Ndiritu et al., 2014), and the importance of measurement error in the relationship between size and land productivity (Desiere and Jolliffe, 2018; Abay et al., 2019). However, when evaluating the extent of factor misallocation in agriculture, we recommend using aggregated data at the farm level.

Our recommendation is based on three main reasons. First, data aggregation can reduce the relative magnitude of measurement error in survey data. Comparing self-reported and GPS measures of landholding area reveals systematic measurement error, which is larger at granular levels and reduced when aggregated to the farm level. Second, the contribution of measurement error to misallocation estimates is much smaller when the data are aggregated to the farm level. Using an alternative method to assess measurement error with panel data proposed by Bils et al. (2021), we find that additive measurement error explains less than 20% of the variation in a standard measure of misallocation across farms. This proportion is much lower than the 70% contribution of mismeasurement at the plot level suggested by Gollin and Udry (2021). Third, the analysis at the farm level would be in line with economic theory and policy practice that treats the household farm, not the plot, as the relevant unit of production and decision making (De Janvry et al., 1991; Restuccia, 2020).

The outline of the paper is as follows. Section 2 discusses the data and methods we use to calculate productivity at plot and farm levels, efficiency gains, and other measures of misallocation. Section 3 presents the main comparative results at the plot and farm levels. Section 4 provides suggestive evidence of a larger measurement error in more granular data that is attenuated when the data are aggregated at the farm level, and the results of an alternative method to assess measurement error in misallocation measures at the farm level when panel data are available. Section 5 concludes.

2 Methods

2.1 Data

Our analysis focuses on a dataset that combines four rounds of the Uganda Panel Survey (2009-2010, 2010-2011, 2011-12 and 2013-14). This is a household-level survey collected with support of the World Bank, as part of the LSMS-ISA project, and contains a rich set of socioeconomic and agricultural information. Agricultural information includes the quantity of crop output harvested and the inputs used (land and labor) in each growing season.[1]

We construct a dataset with agricultural output and inputs at the plot level. This dataset is the same as in Gollin and Udry (2021) and uses the same variable definitions.[2] In particular, a plot is defined as a contiguous area in which a farmer grows a specific crop or crop mixture. The output is measured as the value of the harvest. This value is calculated using as prices the median crop values in a community. Labor is the total number of person days (both family members and hired workers) employed in the plot. Land is the self-reported area of the plot.

We aggregate the plot-level data by adding the value of output and inputs of all the plots in a given farm. We define a farm as a set of plots operated by members of the same household. Note that we do not change the definition of variables, but only the level of data aggregation.

2.2 Measuring misallocation

Our approach is in line with current research on misallocation in agriculture (Adamopoulos and Restuccia, 2014, 2020; Chen et al., 2023; Aragón et al., 2022). We quantify misallocation by assessing the potential aggregate efficiency gains. Efficiency gains refer to the change in aggregate output that could be achieved if resources, such as land and labor, were efficiently allocated between production units. While it may be difficult in practice for an economy to achieve the hypothetical efficient allocation, we use it as a benchmark for comparison of the potential reallocation gains of inputs across plots and across farms. We define efficiency gains as:

$$\text{Efficiency gains} = \frac{Y^e - Y^a}{Y^a}, \tag{1}$$

[1] As a robustness check, we replicate our baseline analysis using data from Tanzania. See Online Appendix A for additional details.

[2] We construct the dataset using the raw data and code in Gollin and Udry (2021)'s supplementary material available at `https://www.journals.uchicago.edu/doi/suppl/10.1086/711369`.

where Y^e is the aggregate output assuming an efficient allocation of inputs, and Y^a is the aggregate output with the actual input allocation. We observe the actual input allocation in the data. However, the efficient allocation is a counterfactual. To calculate it, we require more structure.

Finding the efficient allocation. We assume an economy comprising n production units indexed by i.[3] Depending on the level of data aggregation, the production unit would be the plot or the farm. Each production unit produces the same homogeneous good according to the following Cobb-Douglas technology,

$$y_i = s_i(\ell_i^{\alpha} x_i^{1-\alpha})^{\gamma}, \qquad \alpha, \gamma \in (0, 1), \tag{2}$$

where s_i is the total factor productivity of the production unit, and ℓ_i and x_i are the amounts of land and labor allocated to (and used by) production unit i.[4] Note that production units are heterogeneous as they differ in their level of productivity s_i and potentially in the allocation of inputs. We also emphasize that the assumption that γ is between 0 and 1 implies that the production function features decreasing returns to scale and a non-degenerate distribution of production units even in the efficient allocation (Lucas Jr, 1978; Hopenhayn, 1992).

We obtain the efficient allocation by solving the following social planner problem:

$$\begin{aligned} \underset{\{\ell_i, x_i \geq 0\}_{i=1}^n}{\text{maximize}} \quad & Y = \sum_{i=1}^{n} y_i(s_i, \ell_i, x_i) \\ \text{subject to} \quad & L = \sum_{i=1}^{n} \ell_i \text{ and } X = \sum_{i=1}^{n} x_i, \end{aligned}$$

where L and X are the total endowments of land and labor available in the economy.

[3]We treat the panel data as repeated cross-sections, and omit the time subscript for simplicity of exposition.

[4]In what follows, we refer to s_i as total factor productivity or productivity interchangeably.

In the efficient allocation, the marginal product of each factor is equalized across production units. Using this result, we can write down the efficient allocation as

$$\ell_i^e = \frac{s_i^{1/(1-\gamma)}}{\sum_i s_i^{1/(1-\gamma)}} L, \qquad x_i^e = \frac{s_i^{1/(1-\gamma)}}{\sum_i s_i^{1/(1-\gamma)}} X. \tag{3}$$

Note that the efficient allocations of land and labor in production unit i are proportional to $s_i^{1/(1-\gamma)}$, which implies that more productive units are allocated more inputs. In particular, the efficient allocation implies a positive elasticity of inputs (land and labor) with respect to productivity s_i, which is approximately equal to $1/(1-\gamma)$. This value serves as a benchmark for comparing the empirical elasticities implied by actual allocations, which may be different and could even be negative.

Calculating efficiency gains. We first calculate the efficient allocation (3) using estimates of the production function parameters, and data on the actual output and inputs used (ℓ_i^a, x_i^a). L and X are obtained by adding up the actual allocations of all production units. We then calculate the output that each production unit would have obtained in the efficient and actual allocation (y_i^e, y_i^a). These values are obtained by evaluating (2) using the estimated parameters of the production function and the respective allocation of inputs. Finally, we aggregate the output of each production unit to obtain aggregate efficient and actual output $Y^e = \sum_i y_i^e$ and $Y^a = \sum_i y_i^a$, and calculate the efficiency gains using the definition in equation (1).

We also calculate efficiency gains assuming reallocation of inputs within narrower geographical areas, such as regions or villages. To do so, we use the same procedure but instead calculate area-specific efficient allocations. For example, if there are J areas, then the efficient allocation for production units in area j is defined as:

$$\ell_{ij}^{e} = \frac{s_i^{1/(1-\gamma)}}{\sum_{i\in j} s_i^{1/(1-\gamma)}} L_j, \qquad x_{ij}^{e} = \frac{s_i^{1/(1-\gamma)}}{\sum_{i\in j} s_i^{1/(1-\gamma)}} X_j, \tag{4}$$

where $L_j = \sum_{i\in j} \ell_i^a$ and $X_j = \sum_{i\in j} x_i^a$.

In practice, we observe outputs and inputs for each growing season and thus obtain several measures of efficiency gains (one for each period). We report simple averages.

2.3 Production function estimates

we use two sets of estimates of production function parameters from previous studies: (a) estimates from Gollin and Udry (2021) and (b) parameter values commonly used in the macroeconomic literature on misallocation. We acknowledge that these estimates carry considerable uncertainty, and we do not assert that one set is superior to the other. However, we prefer to use them instead of estimating our own production function to focus on the comparative assessment of misallocation in agriculture between the plot and the farm, rather than on providing the most accurate estimate of misallocation.

Gollin and Udry (2021) estimates. First, we rely on the two-stage least squares (2SLS) estimates from Gollin and Udry (2021). They estimate the following Cobb-Douglas production function:

$$\ln y_i = \alpha_L \ln \ell_i + \alpha_X \ln x_i + W_Y \beta + \epsilon_i, \tag{5}$$

where W_Y is a set of other drivers of plot-level output including year-region-season-crop group fixed effects, plot characteristics (soil quality, water source, slope, etc), interaction of soil quality with weather shocks (rain, drought, and floods), and an indicator of the household having received advice on agricultural production. As instruments for $\ln \ell_i$ and $\ln x_i$, Gollin and Udry (2021) use a rich set of household shocks (such as illness events and weather shocks), plot, farmer and household characteristics. Their estimates are $\alpha_L = 0.69$ and $\alpha_X = 0.22$. Note that in terms of our production function (2), these estimates imply values of $\alpha = 0.76$ and $\gamma = 0.91$.

We calculate plot-level productivity ($\ln s_i$) as a residual of the regression (5). We use the microdata and estimates of the production function parameters already provided in Gollin and Udry (2021)'s replication package. We also calculate an adjusted measure of productivity using the correction in Gollin and Udry (2021). This correction is built on the assumptions of efficient allocation within a farm and classical measurement error. Under these assumptions, Gollin and Udry (2021) show that the covariances of the output and inputs of plots within a farm contain information to estimate the dispersion attributed to measurement error and late-season shocks. Their correction subtracts this dispersion from the observed productivity, so that the variance of the adjusted productivity is smaller.[5]

To construct measures of farm-level productivity, we aggregate the output and input measures of all plots operated by the farm household and then calculate the residual assuming the same production function $\ln s = \ln y - \alpha_L \ln \ell - \alpha_X \ln x$. The main distinction from the plot-level residual obtained from equation (5) is that the farm-level residual does not control for other covariates. However, following an alternative approach that incorporates the covariates' information in the farm-level estimates, we obtain quantitatively similar results.[6]

Macro estimates. As an alternative and for comparability, we use parameters commonly used in the macroeconomic literature on misallocation (Adamopoulos et al., 2022; Chen et al., 2022, 2023). We emphasize that these parameters are not estimated using econometric

[5]See Appendix A.2 and equation (18) in Gollin and Udry (2021) for further details.

[6]An alternative approach to measure farm-level productivity is to calculate a weighted average of plot-level productivity of all the plots operated by the household. Denoting the unadjusted productivity of plot p in farm i as s_{ip}, the farm-level productivity would be $\sum_p s_{ip}(\phi_{ip}^L)^{\alpha_L}(\phi_{ip}^X)^{\alpha_X}$, where ϕ_{ip}^L and ϕ_{ip}^X are the shares of farm i's land and labor used in plot p. This alternative measure contains the same information as the productivity at the plot level. We opt for using a simpler measure to preserve comparability with the results using macro-estimates.

methods, but selected to match observed factor income shares. We select $\gamma = 0.70$ to be within the range of values (0.54,0.85) used in previous studies (Restuccia and Rogerson, 2008; Adamopoulos and Restuccia, 2014). This value is also similar to the returns to scale estimated by recent studies using farm-level data such as Shenoy (2017), Aragón et al. (2022), and Manysheva (2021).

We then select $\alpha = 0.57$ to imply a share of land income of 40%. This land share is higher than that observed in U.S. agriculture but comparable to estimates from less developed countries (Chen et al., 2022; Adamopoulos et al., 2022). In the same spirit as before, we calculate productivity, both at the plot and the farm levels, as a residual from equation (2), i.e., $\ln s_i = \ln y_i - \gamma\alpha \ln \ell_i - \gamma(1-\alpha) \ln x_i$.

3 Main results

Table 1 presents our estimates of the efficiency gains for Uganda. We calculate efficiency gains using two sets of production function parameters and two different levels of data aggregation (plot and farm levels). We highlight the following observations:

Obs. 1: Efficiency gains using plot-level data are extremely large. The estimates in column (1) indicate that if the actual allocations of land and labor shift to their efficient levels at the national level, agricultural output would increase by 2,268%. Even when reallocation is limited to smaller geographic areas, the estimated efficiency gains are substantial: 1,552% within regions and 311% within villages.

These estimates are remarkably large and suggest a level of misallocation that far exceeds what has been documented in the macroeconomics literature. To provide a comparative context, the estimated efficiency gains in agriculture at the national level in China, Ethiopia, and Malawi are 53%, 97%, and 259%, respectively (Adamopoulos et al., 2022; Chen et al.,

Table 1: Analysis of misallocation in Uganda

	Gollin and Udry (2021) estimates			Macro estimates	
	Plot level	Plot level (adjusted)	Farm level	Plot level	Farm level
	(1)	(2)	(3)	(4)	(5)
A. Efficiency gains (%)					
Nationwide	2,268	562	1,112	223	107
Region	1,552	438	647	210	98
Parish (Village)	311	143	105	105	40
B. Productivity dispersion					
Var($\ln s_i$)	1.26	0.52	0.77	1.22	0.81
C. Input-productivity elasticities					
Land	−0.16	−0.25	−0.06	0.00	0.21
Labor	−0.03	−0.04	0.08	−0.01	0.18
Efficient	11.11	11.11	11.11	3.33	3.33
D. Parameters					
α	0.76	0.76	0.76	0.57	0.57
γ	0.91	0.91	0.91	0.70	0.70
Number of production units	41,731	41,731	15,377	41,731	15,377

Notes: Efficiency gains refer to the increase in aggregate output associated with changing actual allocation to efficient allocations, expressed in percentage change, and averaged over growing seasons. All columns use the same plot-level data, but columns (3) and (5) use the data aggregated to the farm-level. Columns (1)-(3) use estimates of the parameters of the production function from Gollin and Udry (2021), while columns (4) and (5) use estimates from the macroeconomics literature. Column (1) uses as production unit productivity (s_i) the unadjusted plot-level productivity estimated by Gollin and Udry (2021), while column (2) uses plot-level productivity adjusted for measurement error and unobserved heterogeneity. Columns (3) to (5) calculate productivity as the residual $\ln s_i = \ln y_i - \gamma\alpha \ln \ell_i - \gamma(1-\alpha) \ln x_i$. Input-productivity elasticities are obtained by regressing ln input on $\ln s_i$ and season-year fixed effects.

2022, 2023).

Obs. 2: Efficiency gains remain large even after adjusting for measurement error and unobserved heterogeneity. In column (2), we adjust the measure of productivity to remove other sources of dispersion following Gollin and Udry (2021). We obtain a substantial decrease in productivity dispersion of 59%, from 1.26 to 0.52 (see Panel B), along with a corresponding drop in the estimated efficiency gains between 54 and 75% depending on the geographical scope of reallocation. Despite this large relative reduction in productivity dispersion and efficiency gains, the level of efficiency gains remains quite large: 562% for reallocation at the national level and 143% for reallocation within villages.

A possible interpretation of these two observations is that misallocation in Uganda is actually quite large, higher than previously documented in other countries. Moreover, it would suggest that measurement error and unobserved heterogeneity explain a large proportion of previous estimates of misallocation, as argued by Gollin and Udry (2021). However, this interpretation implicitly assumes that estimates of misallocation using plot and farm-level data (as in previous studies) are comparable. The next observation suggests that this assumption is not warranted, but rather that the level of data aggregation matters for the results.

Obs. 3: Efficiency gains are much smaller when aggregating the same data to the farm level. Data aggregation alone in column (3) reduces the estimated efficiency gains by 50-66% and productivity dispersion by almost 40% relative to the estimates in column (1). This reduction in productivity dispersion and efficiency gains is achieved without any correction for measurement error or unobserved heterogeneity. Moreover, some of the level estimates of efficiency gains at the farm level (column 3) are smaller than those obtained using the adjusted plot-level productivity (column 2). For example, the efficiency gain within villages at the farm level is 105%, while with adjusted plot-level productivity is 143%.

We observe similar qualitative patterns when using a different set of parameters of the production function from the macroeconomics literature (columns 4 and 5). In this case, the efficiency gains at the farm level are also relatively smaller than at the plot level. However, the efficiency gains calculated with these alternative parameters are substantially smaller than those of columns (1) to (3), and closer in magnitude to estimates from previous macroeconomic studies of misallocation. The efficiency gains in column (4), at the plot level, range from 105% within villages to 223% nationwide. These values are well below the efficiency gains calculated using Gollin and Udry (2021)'s production function estimates, with and without adjustments for measurement error in columns (1) and (2). This substantial reduction in the level of misallocation occurs even though the productivity dispersion of the macro estimates is larger (1.22 vs. 0.52).

This result illustrates the importance of production function estimates and the limitation of using productivity dispersion as a measure of misallocation. In general, the magnitude of efficiency gains (and the implied reallocation gains) is a function of productivity dispersion, economies of scale, and the relationship between input allocation and productivity. Thus, production dispersion alone is not enough to compare the extent of misallocation across different contexts.

Our baseline results focus on efficiency gains as our main measure of misallocation. However, we obtain similar results when using other measures, such as the productivity dispersion and the elasticity of each input with respect to total factor productivity (input-productivity elasticity) as documented in panels B and C in Table 1.

In the efficient allocation, the input-productivity elasticities should be positive and larger than one (see equation (3)). In contrast, in the data, we observe that the estimated elasticities at the plot level are quite small and even negative. This result implies that the actual allocations of inputs across plots differ substantially from the efficient allocation. Instead of allocating more resources to more productive plots, less productive plots receive larger amounts of land and labor.

Moreover, estimated input-productivity elasticities at the plot level are not affected by the correction for unobserved heterogeneity and measurement error. We also observe that the estimated elasticities are greater at the farm level than at the plot. This observation is consistent with a lower degree of misallocation when the granular data are aggregated at the farm level.

Discussion. We observe large discrepancies when calculating efficiency gains and other measures of misallocation at the plot and farm levels. These findings are not unique to Uganda. We find similar differences using data from Tanzania (see Appendix A). We interpret these differences as evidence that measures of misallocation using different levels of data aggregation are not comparable. In short, when assessing misallocation in agriculture, data aggregation matters.

This finding has two important implications for researchers and policy makers. First, it underscores the need to take into account the level of data aggregation when assessing misallocation in a country, performing cross-country comparisons, or evaluating the impact of policies. Analysis at the plot level may produce larger estimates that are not comparable to measures using more aggregated data.

Second, it illustrates the limitations of extrapolating insights from plot-level analysis to the farm level. This limitation becomes evident when plot-level data are used to assess the importance of measurement error in misallocation. For example, a researcher using plot-level data would observe that measurement error explains a large fraction of the productivity dispersion and wrongly conclude that estimates using farm-level or other data overstate the magnitude of misallocation. Our analysis indicates a high degree of input misallocation in Ugandan agriculture, highlighting the need for research identifying the sources of this misallocation and potential policy remediation.

4 Measurement error and misallocation

We now examine measurement errors and misallocation using data at different levels of aggregation. First, we show that more granular data feature a relatively larger measurement error than the same data aggregated at the farm level. Second, we apply a method recently proposed by Bils et al. (2021) to assess misallocation in the presence of additive measurement error. Using panel data at the farm level, we find that measurement error accounts for less than 20% of the observed dispersion in measures of misallocation. This share is much smaller than the importance of mismeasurement at the plot level emphasized in Gollin and Udry (2021).

4.1 Measurement error in disaggregated data

Due to data availability, we focus on measurement error in land input. To assess measurement error, we compare two measures of the size of landholdings: self-reported area provided by the farmer and land areas calculated using Global Positioning System (GPS) information collected by the survey enumerators. Although not exempt from potential error, the GPS measure is arguably more precise and less prone to farmer misreporting (Carletto et al., 2017).

We use the same data sources as in the previous section, i.e., four rounds of the Uganda Panel Survey, and construct a data set with information on land areas at the farm and parcel levels. A parcel is a set of plots within a farm; hence, the parcel-level data are less disaggregated than the plot-level data. However, we choose this level of aggregation due to data limitations since the GPS measure is not available at the plot level, only available at the parcel level. This implies that our evidence below of larger measurement error at the parcel level compared to the farm level is conservative of the likely larger measurement error in plot-level self-reported data.

The data comprise around 20,400 parcel-time observations, of which only 9,900 (48%) have GPS data. We note, however, that the analysis in this Section is performed using only the parcel observations for which we have both GPS and self-reported data. To obtain farm-level data, we first drop parcels without GPS data and then aggregate the self-reported (or GPS) areas of the remaining parcels.

Our sample comprises 6,069 farms with at least one parcel with GPS information (3,036 of those have only one parcel with GPS information) compared to 9,148 farms with self-reported data. On average, there are 2.2 parcels per farm with self-reported data, whereas our sample with GPS information comprises 1.6 parcels per farm. Moreover, the distribution of farms and parcels self-reported area is similar regardless of having or not having a GPS measure. This observation reduces concerns about systematic bias in the collection of GPS data.

Figure 1 shows the distribution of the landholding area in hectares using different levels of data aggregation, but the same scale on the axes. The solid lines represent self-reported values, while the dashed lines correspond to GPS values. There are three relevant observations.

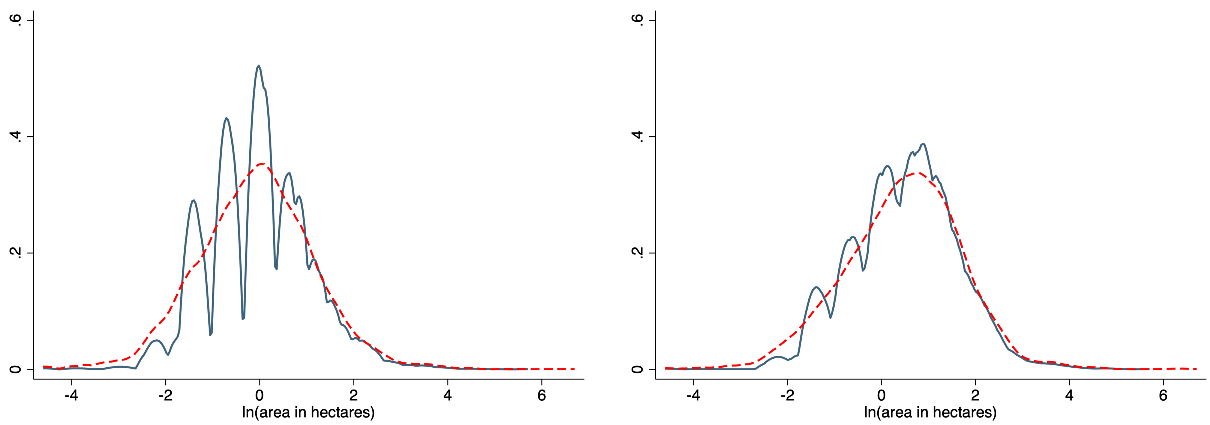
1. There are obvious discrepancies between self-reported and GPS measures of landholding area. The discrepancy has been documented in other studies and interpreted as evidence of measurement error in self-reported values (Judge and Schechter, 2009; Carletto et al., 2015; Gourlay et al., 2019; Abay et al., 2021). Interestingly, the GPS measure follows a smooth bell-shaped distribution, whereas self-reported measures are heaped around certain values. The observed 'heaping' is indicative of respondents (or enumerators) rounding the reported size (Abay et al., 2019, 2021; Carletto et al., 2013).

2. The discrepancy between self-reported and GPS measures is more pronounced among smaller units on the left side of the distribution. This evidence suggests that the

measurement error is not classical, but is correlated with the unit size. This pattern has been documented in other studies. For example, Abay et al. (2021) reports a negative correlation between plot size and land measurement error in four sub-Saharan African countries.

3. The discrepancy between self-reported and GPS measures appears to be greater in more granular data at the parcel level compared with the farm level.

Figure 1: Distribution of landholding size, self-reported and GPS measures



(a) Parcel level (b) Farm level

Notes: Distribution of the log area of land holdings at the parcel level (panel a) and aggregated to the farm level (panel b). Solid lines represent self-reported values, while red-dashed lines represent GPS measures.

The last observation suggests that aggregating data at the farm level might attenuate some of the measurement error at a more granular level. To quantitatively assess this observation, we use two measures of the relative importance of measurement error. First, we examine the average relative error:

$$\text{E(relative error)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{area_i^{SELF} - area_i^{GPS}}{area_i^{GPS}}\right), \quad (6)$$

where $area_i^{SELF}$ and $area_i^{GPS}$ refer to the area of the i-th parcel (or farm) self-reported by the farmer or calculated using GPS. This indicator measures the average overreporting of the landholding size.

Second, we define $\ln error_i = \ln area_i^{SELF} - \ln area_i^{GPS}$, and calculate the relative log-variance of the error:

$$\frac{\text{Var}(\ln error_i)}{\text{Var}(\ln area_i^{GPS})}. \tag{7}$$

This measure captures the dispersion of measurement error relative to the dispersion of GPS landholding sizes. We use this measure as an indicator of how much measurement error could contribute to the dispersion of productivity estimates. To obtain our estimates, we first trim the 1% tails of the relative error and the log error. Then, we estimate (6), (7) and their difference using bootstrapping with 100 replications.

Table 2 reports our findings. We find evidence of measurement error (column 1). The average parcel is reported to be 26.1% larger than its GPS measurement. We find a similar pattern in the relative dispersion of the error. In the parcel data, the log-variance of the error is almost a quarter of the log-variance of the GPS landholding area. However, the relative importance of measurement error decreases when the data are aggregated at the farm level (column 2). For example, at the farm level, average overreporting drops by almost a third, from 26.2% to 17.7%, while the relative log-variance of the error decreases by almost 20%, from 23.2% to 18.0%.

Our central point is that disaggregated data (at the parcel or more granular level) is much worse in terms of measurement error, than data aggregated at the farm level. This conclusion has two caveats. First, due to data limitations, we cannot directly assess the extent of measurement error at the plot level. However, because plot-level data are even more granular than the data at the parcel level, it is likely that measurement error problems in plot-level data are actually worse. Second, our findings are illustrative of the Ugandan case and might not extrapolate to other contexts.

Table 2: Relative importance of measurement error in landholding area

	Parcel-level landholding area (1)	Farm-level landholding area (2)	Difference between (1) − (2) (3)
E(relative error)	0.262	0.177	0.085
	(0.009)	(0.010)	(0.008)
$\frac{\text{Var}(\ln error_i)}{\text{Var}(\ln area_i^{GPS})}$	0.232	0.180	0.051
	(0.005)	(0.005)	(0.004)
No. obs.	9,907	6,069	

Notes: E(relative error) is the average relative error defined as the ratio of self-reported and GPS area minus one. ln error is the difference between ln self-reported area and ln GPS area. Bootstrapped standard errors are in parentheses. Column (3) reports the estimated difference between the measures in columns (1) and (2).

4.2 Measurement error and misallocation at the farm level

If measurement error is larger in more granular data at the plot level, how can we assess the contribution of measurement error to measures of misallocation at the farm level? To answer this question, we implement an alternative method proposed by Bils et al. (2021) to correct for additive measurement error when panel data are available.

Their method is based on the observation that the growth of output (or revenue) of a production unit would be proportional to the growth of inputs. However, this elasticity would be affected by additive measurement error. For example, if there is over-reporting of inputs, then the output (or revenue) would increase proportionally less, and the observed elasticity would be smaller. The method exploits how this elasticity varies across production units to correct measures of misallocation from additive measurement error.

Following Bils et al. (2021) and assuming a Cobb-Douglas production function as in equation (2), we define output per unit of composite input (TFPR or average revenue productivity in their context) as:

$$\mathrm{TFPR}_{it} \equiv \frac{\hat{R_{it}}}{\hat{I_{it}}} \equiv \frac{y_{it}}{\ell_{it}^{\alpha} x_{it}^{1-\alpha}} \tag{8}$$

where $\hat{R_{it}}$ is the observed output in farm i in period t and $\hat{I_{it}}$ is a composite input calculated using the observed inputs of land and labor. Note that TFPR_{it} is not a measure of total factor productivity s_{it}, which would be $\hat{R_{it}}/\hat{I_{it}}^{\gamma}$, but instead a measure of the average and marginal product of inputs.[7]

TFPR_{it} serves as an indicator of distortions because, in an efficient allocation, it is equalized across production units. Therefore, dispersion in TFPR_{it} would signal misallocation. However, if the observed TFPR_{it} contains measurement error, its dispersion would be larger, leading to an overestimation of the extent of misallocation

Bils et al. (2021) show that, with additive measurement error in either output or inputs, the dispersion of the observed TFPR_{it} can be corrected using the following expression:

$$\mathrm{Var}(\ln \tau_{it}) = \mathrm{Var}(\ln \mathrm{TFPR_{it}}) + \mathrm{Cov}(\ln \mathrm{TFPR_{it}}, \ln \beta_k), \tag{9}$$

where τ_{it} is the true value of TFPR_{it} without measurement error, and β_k is the elasticity of observed output with respect to observed inputs, conditional on TFPR_{it} taking a particular value $\mathrm{TFPR_k}$. In the empirical application, TFPR_k corresponds to the k-th decile of the TFPR_{it} distribution.[8]

We implement Bils et al. (2021)'s method using the same data as in Section 3 and the following procedure:[9]

[7] We add the subscript t to acknowledge the time dimension of the panel data. In our previous results, we treated the data as a series of repeated cross-sections and omitted the time subscript to simplify the exposition.

[8] Formally $\beta_k \equiv \frac{\mathrm{Cov_k}(\Delta \hat{R_{it}}, \Delta \hat{I_{it}})}{\mathrm{Var_k}(\Delta \hat{I_{it}})}$, where Δ represents percentage changes, and Cov_k and Var_k are the conditional covariance and variance.

[9] Our procedure follows the steps described in Section 5 of Bils et al. (2021)'s online appendix. The appendix is available at `https://ars.els-cdn.com/content/image/1-s2.0-S0304393221000970-mmc1.pdf`.

1. We aggregate the plot-level data to the farm-level to construct a panel of farms with year-season as the time period.

2. We construct TFPR_{it} using the real value of agricultural output y_{it} as the measure of $\hat{R}_{it}$ and assuming similar values of α as in Table 1. We trim the 1% tails of the TFPR_{it} distribution, but we also present results without trimming.

3. To obtain deciles of the TFPR_{it} distribution, we first calculate the deviation of $\ln(\text{TFPR}_{it})$ from the period average. Then, we average these deviations at the farm level to obtain Tornqvist $\ln(\text{TFPR}_i)$ deviations, and place them into deciles. We trim observations with extreme values of TFPR growth (i.e. within-farm increase or decrease by a factor of 5 or more relative to the period average)

4. We regress output growth on input growth and period fixed effects separately for each decile of the TFPR_{it} distribution. The coefficients on input growth are the β_k estimates. The results of these estimates are reported in Table B.1 in the Appendix.

5. We merge the β_k estimates to the farm data using the decile's cutoff values and the expected value of the Tornqvist $\ln(\text{TFPR}_i)$ conditional on $\ln(\text{TFPR}_{it})$. Bils et al. (2021) implement this correction to adjust for the compression of the distribution of Tornqvist $\ln(\text{TFPR}_i)$.

6. We calculate $\text{Var}(\ln \text{TFPR}_{it})$ and $\text{Cov}(\ln \text{TFPR}_{it}, \ln \beta_k)$ for each decile, and use expression (9) to obtain $\text{Var}(\ln \tau_{it})$ and the ratio $\frac{\text{Var}(\ln \tau_{it})}{\text{Var}(\ln \text{TFPR}_{it})}$. This ratio measures the fraction of the observed dispersion of TFPR_i that reflects distortions, not measurement error.

The results are presented in Table 3. The ratio $\frac{\text{Var}(\ln \tau_{it})}{\text{Var}(\ln \text{TFPR}_{it})}$ ranges from 0.80 to 0.87.[10] The estimated ratio is slightly smaller than the findings of Adamopoulos et al. (2022). They

[10]We obtain values for this ratio closer to one using data from Tanzania. See Table A.2 in the online appendix.

use a similar correction with a panel of Chinese farms and estimate a ratio of around 0.90. However, our estimates are larger than the ratio documented by Bils et al. (2021) for manufacturing sectors in India (0.70-0.76) and the United States (0.27-0.43).

Our findings imply that approximately 13-20% of the variation in observed TFPR_{it} can be attributed to measurement error. This result contrasts with Gollin and Udry (2021), who emphasize that measurement error in plot-level data is quite important and can explain 59-70% of productivity dispersion. We interpret these findings as suggestive evidence that measurement error could play a more important role in assessing misallocation with disaggregated data.

Table 3: Dispersion in marginal products and TFPR in Uganda

	Trimming 1%		No trimming	
	(1)	(2)	(3)	(4)
Value of α	0.76	0.57	0.76	0.57
$\frac{\text{Var}(\ln \tau)}{\text{Var}(\ln \text{TFPR})}$	0.87	0.84	0.81	0.80
$\text{Var}(\ln \tau)$	0.56	0.52	0.65	0.61
$\text{Var}(\ln \text{TFPR})$	0.64	0.62	0.80	0.77

Notes: The table shows the log-variance of the true dispersion (τ), revenue productivity (TFPR), and their ratio estimated using the correction proposed by Bils et al. (2021). TFPR is calculated using the same values of α as in Table 1. Columns 1 and 2 trim the 1% tails of the TFPR distribution.

5 Conclusion

We examine whether the level of data aggregation, plot level or farm level, affects measures of the extent of misallocation in agriculture. Using detailed micro-data from Uganda and Tanzania, we show that the plot-level analysis produces much larger estimates of misallocation than the analysis at the farm level. Even after controlling for unobserved heterogeneity and measurement error, the extent of misallocation from plot-level analysis remains quite large. Our findings suggest that the level of data aggregation matters in the assessment of misallocation and can lead to large discrepancies in the comparative literature that focuses on the farm as the main unit of analysis.

The large discrepancy in assessed misallocation suggests caution in extrapolating insights obtained from plot-level analysis, such as the relative importance of measurement error, to results obtained at the farm level. We also show that differences in estimates of production function parameters are important for the level of assessed misallocation, and hence its comparability across studies. We provide suggestive evidence from self-reported and GPS landholding areas that data granularity can lead to larger measurement error that is attenuated in aggregated data at the farm level.

Our findings have important implications for researchers and policy makers. First, they imply that estimates of misallocation at the plot and farm levels are not comparable. In particular, analysis at the plot level might lead to much larger estimates of misallocation. This issue should be taken into account when assessing the extent of misallocation in a country, making cross-country comparisons, or evaluating the impact of policies. Second, they suggest caution in generalizing insights from plot-level analysis to the farm level. For example, a researcher using plot-level data would observe that measurement error explains a large fraction of the productivity dispersion and wrongly conclude that misallocation in agriculture is not an important issue for productivity in less developed countries.

References

Abay, Kibrom A., Gashaw T. Abate, Christopher B. Barrett, and Tanguy Bernard, "Correlated non-classical measurement errors, 'Second best' policy inference, and the inverse size-productivity relationship in agriculture," *Journal of Development Economics*, 2019, *139*, 171 – 184.

Abay, Kibrom A, Leah EM Bevis, and Christopher B Barrett, "Measurement Error Mechanisms Matter: Agricultural intensification with farmer misperceptions and misreporting," *American Journal of Agricultural Economics*, 2021, *103* (2), 498–522.

Adamopoulos, Tasso and Diego Restuccia, "The Size Distribution of Farms and International Productivity Differences," *American Economic Review*, June 2014, *104* (6), 1667–97.

_ and _ , "Land Reform and Productivity: A Quantitative Analysis with Micro Data," *American Economic Journal: Macroeconomics*, July 2020, *12* (3), 1–39.

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

Aragón, Fernando M., Diego Restuccia, and Juan Pablo Rud, "Are small farms really more productive than large farms?," *Food Policy*, 2022, *106*, 102168.

Bils, Mark, Peter J Klenow, and Cian Ruane, "Misallocation or mismeasurement?," *Journal of Monetary Economics*, 2021, *124*, S39–S56.

Carletto, Calogero, Sara Savastano, and Alberto Zezza, "Fact or artifact: The impact of measurement errors on the farm size–productivity relationship," *Journal of Development Economics*, 2013, *103*, 254–261.

_ , Sydney Gourlay, and Paul Winters, "From guesstimates to GPStimates: Land area measurement and implications for agricultural analysis," *Journal of African Economies*, 2015, *24* (5), 593–628.

_ , _ , Siobhan Murray, and Alberto Zezza, "Cheaper, faster, and more than good enough: is GPS the new gold standard in land area measurement?," in "Survey Research Methods," Vol. 11 2017, pp. 235–265.

Chen, Chaoran, Diego Restuccia, and Raül Santaeulàlia-Llopis, "The effects of land markets on resource allocation and agricultural productivity," *Review of Economic Dynamics*, 2022, *45*, 41–54.

_ , _ , and _ , "Land misallocation and productivity," *American Economic Journal: Macroeconomics*, 2023, *15* (2), 441–465.

Desiere, Sam and Dean Jolliffe, "Land productivity and plot size: Is measurement error driving the inverse relationship?," *Journal of Development Economics*, 2018, *130*, 84–98.

Gollin, Douglas and Christopher Udry, "Heterogeneity, measurement error, and misallocation: Evidence from african agriculture," *Journal of Political Economy*, 2021, *129* (1), 000–000.

Gourlay, Sydney, Talip Kilic, and David B Lobell, "A new spin on an old debate: Errors in farmer-reported production and their implications for inverse scale-Productivity relationship in Uganda," *Journal of Development Economics*, 2019, *141*, 102376.

Hopenhayn, Hugo A, "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1992, pp. 1127–1150.

Janvry, Alain De, Marcel Fafchamps, and Elisabeth Sadoulet, "Peasant household behaviour with missing markets: some paradoxes explained," *The Economic Journal*, 1991, *101* (409), 1400–1417.

Jr, Robert E Lucas, "On the size distribution of business firms," *The Bell Journal of Economics*, 1978, pp. 508–523.

Judge, George and Laura Schechter, "Detecting problems in survey data using Benfordâs Law," *Journal of Human Resources*, 2009, *44* (1), 1–24.

Manysheva, Kristina, "Land Property Rights, Financial Frictions, and Resource Allocation in Developing Countries," *mimeo*, 2021.

Ndiritu, S Wagura, Menale Kassie, and Bekele Shiferaw, "Are there systematic gender differences in the adoption of sustainable agricultural intensification practices? Evidence from Kenya," *Food Policy*, 2014, *49*, 117–127.

Restuccia, Diego, "The Impact of Land Institutions and Misallocation on Agricultural Productivity," Technical Report March 2020.

_ **and Richard Rogerson**, "Policy Distortions and Aggregate Productivity with Heterogeneous Plants," *Review of Economic Dynamics*, October 2008, *11* (4), 707–720.

_ **and** _ , "The causes and costs of misallocation," *Journal of Economic Perspectives*, 2017, *31* (3), 151–74.

Shandal, Monica, Sandeep Mohapatra, and Prakashan Chellattan Veettil, "Pareto efficiency in intrahousehold allocations: Evidence from rice farming households in India," *Oxford Development Studies*, 2022, *50* (2), 158–176.

Shenoy, Ajay, "Market failures and misallocation," *Journal of Development Economics*, 2017, *128*, 65–80.

Slavchevska, Vanya, "Gender differences in agricultural productivity: The case of Tanzania," *Agricultural economics*, 2015, *46* (3), 335–355.

Udry, Christopher, "Gender, Agricultural Production, and the Theory of the Household," *Journal of Political Economy*, October 1996, *104* (5), 1010–46.

ONLINE APPENDIX

A Evidence from Tanzania

We replicate the baseline analysis using data from Tanzania. We use four rounds of the National Panel Survey (years 2008, 2010, 2012 and 2014).[11] This survey is also collected as part of the World Bank's LSMS-ISA project and contains similar data to the Uganda dataset.

We use the same procedure described in Section 2 to construct measures of agricultural output, inputs and productivity, aggregate the plot-level data, and calculate efficiency gains.

The main difference is that we do not use Gollin and Udry (2021)'s 2SLS estimates for Tanzania, but instead those obtained using the instrumental variables correlated random coefficients estimator (IVCRC). The reason is that the 2SLS estimates effectively imply constant returns to scale ($\gamma = 1.01$) which leads to a corner solution in the efficient allocation, and an undefined value of efficiency gains. The IVCRC estimates are $\alpha_L = 0.61$ and $\alpha_X = 0.26$. In terms of our production function (2), these estimates imply values of $\alpha = 0.70$ and $\gamma = 0.87$.

Table A.1 presents the estimated measures of misallocation.

[11] This dataset is also available in Gollin and Udry (2021)'s supplementary material at `https://www.journals.uchicago.edu/doi/suppl/10.1086/711369`.

Table A.1: Analysis of misallocation in Tanzania

	Gollin and Udry (2021) estimates			Macro estimates	
	Plot level	Plot level (adjusted)	Farm level	Plot level	Farm level
	(1)	(2)	(3)	(4)	(5)
A. Efficiency gains (%)					
Nationwide	4,022	664	1,336	205	142
Region	925	253	394	151	104
Parish (Village)	283	118	132	91	58
B. Productivity dispersion					
Var($\ln s_i$)	1.19	0.47	0.93	1.15	0.93
C. Input-productivity elasticities					
Land	−0.22	−0.35	−0.09	0.04	0.12
Labor	−0.08	−0.13	0.00	0.02	0.08
Efficient	7.69	7.69	7.69	3.33	3.33
D. Parameters					
α	0.70	0.70	0.70	0.57	0.57
γ	0.87	0.87	0.87	0.70	0.70
Number of production units	14,535	14,535	8,293	14,535	8,293

Notes: Efficiency gains are the increase in aggregate output from changing the actual allocation to the efficient one, expressed in percentage change, and averaged over growing seasons. All columns use the same plot-level data, except columns (3) and (5) which aggregate data at the farm-level. Columns (1)-(3) use estimates of the parameters of the production function from Gollin and Udry (2021), while columns (4) and (5) use estimates from the macroeconomics literature. Column (1) uses as production unit productivity (s_i) the unadjusted plot-level productivity estimated by Gollin and Udry (2021), while column (2) uses the plot-level productivity adjusted for measurement error and unobserved heterogeneity. Columns (3) to (5) calculate productivity as the residual $\ln s_i = \ln y_i - \gamma\alpha \ln \ell_i - \gamma(1-\alpha)\ln x_i]$. Input-productivity elasticities are obtained by regressing $\ln$ input on $\ln s_i$ and season-yer fixed effects.

Table A.2: Dispersion in observed and true distortions in Tanzania

	Trimming 1%		No trimming	
	(1)	(2)	(3)	(4)
Value of α	0.70	0.70	0.57	0.57
$\frac{\text{Var}(\ln\tau)}{\text{Var}(\ln\text{TFPR})}$	0.98	0.94	0.89	0.86
$\text{Var}(\ln\tau)$	0.77	0.73	0.84	0.81
$\text{Var}(\ln\text{TFPR})$	0.78	0.77	0.95	0.94

Notes: The table shows the log-variance of the true dispersion (τ), revenue productivity (TFPR), and their ratio estimated using the correction proposed by Bils et al. (2021). TFPR is calculated using the same values of α as in Table A.1. Columns 1 and 2 trim the 1% tails of the TFPR distribution.

B Additional tables

Table B.1: Estimates of β_k

	Decile (k)									
	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}_k$	1.109	0.977	0.883	0.868	0.828	0.737	0.728	0.800	0.701	0.640
	(0.055)	(0.051)	(0.046)	(0.044)	(0.044)	(0.041)	(0.041)	(0.040)	(0.038)	(0.034)

Notes: The table presents estimates of β_k used to calculate $\text{Var}(\tau)$ in column 1 of Table 3. k refers to the decile of the TFPR distribution. Standard errors are in parentheses.