Measurement in applied economics involves trade-offs between cost and quality. In the case of field experiments, researchers must often choose between cheap but noisy and possibly biased survey measures and expensive but more accurate measures of outcomes. We propose a way to reduce the budgeting burden of measurement by correcting for measurement error in the cheaper measure using validation data methods. We document a case in which non-classical measurement error in a key variable of interest would bias the estimated treatment effects in a randomized trial in Kenya. We explore the structure of this measurement error and demonstrate the effectiveness of our bias correction procedure. Finally, we conduct a series of Monte Carlo exercises to explore the statistical power implications of this procedure across a range of sample sizes and measurement error structures.

Abstract

Measurement in applied economics involves trade-offs between cost and quality. In the case of field experiments, researchers must often choose between cheap but noisy and possibly biased survey measures and expensive but more accurate measures of outcomes. We propose a way to reduce the budgeting burden of measurement by correcting for measurement error in the cheaper measure using validation data methods. We document a case in which non-classical measurement error in a key variable of interest would bias the estimated treatment effects in a randomized trial in Kenya. We explore the structure of this measurement error and demonstrate the effectiveness of our bias correction procedure. Finally, we conduct a series of Monte Carlo exercises to explore the statistical power implications of this procedure across a range of sample sizes and measurement error structures.
1 Introduction

Measurement in applied economics involves trade-offs. Many outcomes of interest, like total output by a firm or farmer, are costly to measure objectively. In practice, these outcomes are often measured using survey data. When designing a study or field experiment, researchers work under financial or time constraints and must choose how to collect the best data possible given these constraints. What if the treatment in an experiment affects not only an outcome of interest, but also the measurement of that outcome? And how are the effects of this measurement error exacerbated when final outcomes of interest are the product or ratio of several intermediate outcomes?

In this paper, we explore the implications of a particular form of measurement error: differential mis-reporting by treated and control participants in a randomized experiment. We document the presence of this form of measurement error in self-reported cultivated acreage by farmers in a recent field experiment in western Kenya (Deutschmann et al., 2019). We discuss some possible mechanisms for this error and demonstrate how our interpretation of the results of the experiment would change in the absence of GPS measures of cultivated acreage.

Next, we demonstrate the effectiveness of a validation data method adapted from Carroll et al. (2006) and Buonaccorsi and Tosteson (1993) to correct for this error. In this method, we use a validation sample in which we observe both the biased and non-biased measures to correct for the bias in a larger sample for which we observe only the biased measure. We repeatedly simulate the results of this procedure by randomly sampling without replacement a validation sample from the full set of experiment participants, and treating the remaining farmers as if we knew only their self-reported acreage. We show that, in this case, the distribution of average treatment effect (ATE) estimates recovered using this procedure is centered at the true ATE estimated using the full sample with the non-biased measure.

Finally, we extend this procedure to fully simulated data. We specify a range of sample sizes, measurement error magnitudes, and average treatment effects, and document the performance of the bias correction procedure in each case. We focus on the implications for statistical power and ask: given a particular measurement error structure and average treatment effect, how does the required validation sample size vary with total sample size in order to achieve 80% power? The goal of this exercise is to demonstrate how researchers could account for the possibility of collecting validation data and value the trade-off between a larger validation sample and a larger overall sample given the relative costs of data collection.

This is not the first paper to document measurement error in self-reported cultivated acreage. Recent work has demonstrated this measurement error is widespread in survey datasets and can bias estimates of the relationship between land and productivity (Carletto et al., 2013, 2015; Dillon et al., 2019). However, this work has typically been limited to

---

1 Measurement error in land size and crop yields may contribute to long-standing empirical puzzles in the
non-experimental contexts. In this paper, we show that misreported cultivated acreage would have meaningful impacts on the interpretation of the results of a major randomized evaluation.

As donors and partners increasingly demand experimental evidence on program effectiveness, NGOs and social enterprises are conducting high-stakes evaluations of their core programs (Wydick et al., 2018; Deutschmann and Tjernström, 2018). These evaluations, if they fail to generate positive results, can have real implications for the future of their organizations (Givewell, 2018, 2019). Growing concern among some practitioners about the high cost of randomized trials has also led to a push for smaller, more targeted experimentally-driven M&E (Gugerty and Karlan, 2018). In this paper, we argue that researchers and organizations seeking to evaluate programs should think carefully about the trade-off between added cost and improved measurement.

The problem of measurement error correlated with treatment status is not unique to agricultural settings. Baird and Özler (2012) show that over-reporting by control participants in a conditional cash transfer program would result in significantly understated results of the program on school attendance. Blattman et al. (2016) develop methods for correcting for measurement error when outcome variables are sensitive and reported with error, focusing on collecting qualitative validation data using focus groups to sign the potential bias. In general, given that field experiments in economics can rarely provide a “placebo” treatment to non-treated participants, we might expect that treatment status could impact the salience of certain outcomes, similar to a surveyor or Hawthorne effect. We contribute to this literature by showing how researchers can collect validation data on a subset of research participants and use this data to correct the biased measurement in the rest of the sample.

The remainder of the paper is organized as follows. In Section 2 we outline a simple econometric framework and introduce our bias correction procedure. In Section 3 we document the presence of measurement error in a real experimental context, and in Section 4 demonstrate how our procedure performs in correcting for this error. Section 5 extends the procedure to fully simulated data, and Section 6 concludes.

If both plot size and production measures are measured with non-classical error and correlated, correcting for only one may aggravate bias (Abay et al., 2019a). Importantly, this measurement error may occur for several reasons, including simple misreporting and respondent misperception, the latter of which may impact respondent input decisions (Abay et al., 2019b). Researchers increasingly rely on more costly GPS measurement of land to accurately measure land size, although even this measure is not perfect (Carletto et al., 2016; Dillon et al., 2019).

2 One exception is Abate et al. (2018), who show that non-classical measurement error in postharvest recall data and cultivated acreage can attenuate the estimated treatment effect of a comprehensive package for wheat farmers.

2 Econometric framework

In this section, we present a simple econometric framework for estimating the average treatment effect of a field experiment on an outcome of interest, and demonstrate how this estimated effect may be biased depending on how a mis-measured variable enters the equation. In what follows, we present the framework for a simple, individually-assigned RCT. Future work will extend this to account for clustering in treatment assignment and the inclusion of covariates or fixed effects.

Suppose we wish to estimate the impact of a randomly-assigned treatment \( T \) on true log total output, \( h^* \). In the simplest case, we would estimate the following equation:

\[
h^* = \beta_0 + \beta_1 T + \epsilon \quad (1)
\]

Instead of observing output \( h^* \) directly, we observe log per-acre yields \( y \) and log total acres cultivated \( a \).\(^4\) We then approximate total output as \( h = y + a \). Suppose we observe “true” log total acres cultivated \( a^* \) for a subset of our sample, whereas for all farmers in our sample we observe a self-reported measure \( a^s \).

We estimate the following equation and test whether the mean of the measurement error\(^5\) (defined as \( a^s - a^* \)) is affected by treatment status:

\[
a^s - a^* = \gamma_0 + \gamma_1 T + \mu \quad (2)
\]

If we then plug in this expression to equation 1 above and rearrange, we see that estimating using output approximated with self-reported acres \( (h = y + a^s) \) yields a coefficient on \( T \) which contains an extra term, \( \gamma_1 \):

\[
y + a^s = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1) T + (\epsilon + \mu) \quad (3)
\]

If \( \gamma_1 \neq 0 \), this will introduce bias into our estimated treatment effect. Depending on the context, we may reasonably expect \( \gamma_1 > 0 \) (for example, if treated farmers have some incentive to over-report) or \( \gamma_1 < 0 \) (for example, if farmers tend to overstate their land size but the treatment helps farmers estimate their land size more accurately). Additionally, the mechanisms that cause biased over- or under-reporting could have the opposite effect on the estimated treatment effect if the mis-measured variable enters in the denominator of the outcome. This could occur if, instead of yields and acreage, we measured total output and acreage and wished to estimate the treatment effect on per-acre yields.

In the absence of objective measures of farmer land size \( (a^*) \), it is not obvious in general

\(^4\)This could similarly be conceived in the context of a firm as per-machine output and total machine stock, or per-employee output and total employees, if we observed output for a subset of machines or employees and wished to estimate firm-level impacts.

\(^5\)In future work, we will build on this framework to demonstrate the implications if treatment also affects the variance of the measurement error.
how to deal with this bias. Theory or qualitative work could attempt to characterize whether a particular context should result in a positive or negative $\gamma_1$, and then one could simulate the impact of a reasonable degree of treatment-differential error on the estimated treatment effect.

Alternatively, if may be possible to collect objective measures of farmer land size in a (random) subset of farmers. If so, we propose that researchers can apply validation data methods and estimate a bias-corrected average treatment effect coefficient. The following is adapted from Carroll et al. (2006). Assume for some random subset $M$ of our experiment sample $N$, we observe both the self-reported land size measure $a^s$ and the true land size measure $a^*$. First, using the validation sample and the true land size measure, estimate the vectors $\hat{\beta}^v$ (the treatment effect on total output, equation 1 above) and $\hat{\gamma}^v$ (the treatment effect on measurement error, equation 2 above). Note that this $\hat{\beta}^v$ is an unbiased estimate of the treatment effect, but may be imprecisely estimated if the validation sample is small.

Next, using the values in the vector $\hat{\gamma}^v$, in the full sample generate an estimated unbiased value of $a^*$:

$$\hat{a}^* = a^s - \hat{\gamma}_0^v - \hat{\gamma}_1^v T$$  \hspace{1cm} (4)

Using this estimated $\hat{a}^*$, calculate $\hat{h}^*$ and again estimate equation 1 to obtain a second estimated treatment effect on total output, $\hat{\beta}^f$.

Then, use a bootstrap procedure to estimate the joint covariance matrix $\hat{\Sigma}$ of the two estimates of the treatment effect on output, $\hat{\beta}^v$ and $\hat{\beta}^f$. Use $\hat{\Sigma}$ to form the best weighted combination $\hat{\beta}$ of the two estimates:

$$\hat{\beta} = \left( J^t\Sigma^{-1}J \right)^{-1} J^t\Sigma^{-1} \begin{pmatrix} \hat{\beta}^v, \hat{\beta}^f \end{pmatrix}^t$$

where $J$ is a stacked identity matrix. Use $(J^t\Sigma^{-1}J)^{-1}$ as the estimated covariance matrix for the combined estimate $\hat{\beta}$.

We implement this procedure below in section 4, and explore how results may be sensitive to the size of the validation sample.

3 Case study: Kenya RCT

In this section, we apply the method outlined above in section 2 to data from a randomized evaluation of the One Acre Fund (1AF) small-farmer program in western Kenya. Deutschmann et al. (2019) provide significant detail about the implementation of the project and the primary results. This experiment is a useful context to demonstrate the potential pitfalls of measurement error in self-reported cultivated land size, having collected both farmer self-reported plot size and GPS measurement of plots. The data collected for the experiment is detailed and high-quality, and so provides some confidence that the measurement error we detect is due to errors in self-reporting, rather than problems in survey design.
or implementation.\(^6\) Yield data was collected using a crop-cut in two randomly-selected 40-square-meter boxes on farmer fields.\(^7\) This gives a precise, objective estimate of yields per acre, which is multiplied by cultivated acres to estimate total output.

First, we show that farmers in this dataset report their cultivated acreage with error, and that this error differs by treatment status. Table 1 characterizes the nature of the measurement error in self-reported land size.\(^8\) In column 1, we see that GPS measurement shows treated farmers cultivate more land during the experiment. Column 2 tells a different story, and would suggest that in general, treated farmers do not cultivate more land. Column 3 shows the treatment effect on measurement error itself, the difference between the outcomes in columns 1 and 2. We see that treatment reduces measurement error, and that on average control-farmers over-report their cultivated land size.

**Table 1: Measurement comparison by sample**

<table>
<thead>
<tr>
<th></th>
<th>(1) Log GPS Acres</th>
<th>(2) Log SR Acres</th>
<th>(3) Log SR - Log GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>0.18***</td>
<td>0.06</td>
<td>-0.11***</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.050)</td>
<td>(0.030)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1971</td>
<td>1971</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>-0.26</td>
<td>-0.17</td>
<td>0.030</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. Log GPS Acres is the log of total cultivated maize area of each farmer, measured using GPS. Log SR Acres is the log of farmer self-reported value of total cultivated maize acres. Log SR - Log GPS is the difference between the two values. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

Why do non-treated farmers mis-report their cultivated land size relative to treated farmers? We can speculate about mechanisms, although further work (like Abay et al. 2019) is needed to further understand why this error occurs. It could be that participating in the One Acre Fund program helps farmers more accurately estimate their land size, either by providing useful guidance for estimating it or simply making the size more salient, as farmers must decide how much land to enroll in the program. We see that, at baseline, treated and control farmers reported statistically indistinguishable cultivated acreage for the last three agricultural seasons. It could also be that non-treated farmers wished to strategically mis-report the land they had available, if they thought this might increase

---

\(^6\)Enumerators measured each plot three times, walking twice clockwise and once counterclockwise around the borders of the plot. Data collection on this experiment was also audited by an external firm (Intermedia Development Consultants, 2017) which should further increase confidence.

\(^7\)Unfortunately, we cannot perform a similar exercise for self-reported yields, as farmers were not asked about their maize output during the experiment.

\(^8\)Note: the data actually include three self-reported estimates of cultivated acres, from three different surveys. The estimate we use for results in this section comes from the Land Estimates survey, in which farmers were first asked how many acres they cultivated before the GPS measurement was taken. Results using the two other land size measures are broadly similar and presented in Appendix A.
their chances of being able to enroll in the next season.

Next, Table 2 illustrates the effects of this measurement error when we estimate the average treatment effect on log total maize output. If we regress log output calculated using GPS-measured maize acres, we estimate that the program had a positive and statistically-significant impact on total output. What if we had only collected self-reported acreage instead? As column 2 shows, we would find a treatment effect which was smaller in magnitude and less precisely estimated.\(^9\)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log Output</td>
<td>Log Output</td>
</tr>
<tr>
<td>Treated</td>
<td>0.36***</td>
<td>0.24***</td>
</tr>
<tr>
<td></td>
<td>(0.060)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1971</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>6.68</td>
<td>6.77</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. \(H_{GPS}\) is log total maize output estimated using crop-cut yields and GPS-measured cultivated acreage. \(H_{SR}\) is log total maize output estimated using crop-cut yields and self-reported cultivated acreage. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

In many contexts, it may be more realistic to receive a relatively well-measured estimate of total output. For example, farmers who produce crops for sale may obtain accurate estimates of total production by weight. In this case, we may instead wish to estimate a treatment effect on average productivity per acre. To demonstrate the impacts of measurement error in cultivated land size in this case, we divide total output (calculated using GPS-measured maize acres) by self-reported cultivated acreage to create an additional estimate of per-acre maize yields:

\[
Y_{SR} = \frac{\text{Total maize output (calculated with GPS land measure)}}{\text{Self-reported cultivated acres}}
\]

We compare this to maize yields measured directly using a crop cut.\(^10\) If we assume that the total output measure with GPS acreage is an accurate estimate of total farmer production, then this

---

\(^9\)As the comparable Table A.2 in Appendix A shows, if we relied instead on the first self-reported measure of cultivated acreage where the differential measurement error is more severe, we would fail to detect a treatment effect on \(Y_{SR}\).

\(^{10}\)Note: this variable is defined slightly differently than in Deutschmann et al. (2019). In that paper, when we considered the yield outcome we compared the yields on enrolled plots for treated farmers to the yields of control farmers. Here, for treated farmers, we divide total maize output by total maize acres to produce a weighted average estimate of maize yields. This allows us to focus on the error in measurement of one variable (total acreage). We do not see substantial differences in measurement error for enrolled and non-enrolled plots for treated farmers.
Table 3 presents the results of this exercise. Again, we see in column 1 that the treatment had a positive and significant effect on log maize yield. However, we now see in column 2 that we would estimate a much larger average treatment effect on yields if we relied on self-reported maize acres. In this case, the bias moves in the opposite direction because cultivated acreage enters into the denominator of the outcome rather than the numerator.

<table>
<thead>
<tr>
<th></th>
<th>(1) Log Crop-cut Yield $y_{GPS}$</th>
<th>(2) Log SR Yield $y_{SR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>0.18*** (0.030)</td>
<td>0.30*** (0.040)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1971</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>6.94</td>
<td>6.85</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. Log Crop-cut Yield $y_{GPS}$ is measured directly using a crop cut from a 40 square meter box on farmer fields. Log SR Yield $y_{SR}$ is calculated by dividing estimated total output by self-reported cultivated acreage. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

Of course, since in this study we collected GPS land measures, we need not rely on self-reported measurement at all. But what if, for logistical or budget reasons, it was not feasible to collect an objective measure for all participants? One possibility would be to collect both measures for a random validation sample, and collect only the (possibly biased) self-reported measure for the rest of the larger experimental sample. This is the approach we explore in Section 4.

4 Bias correction simulation with real data

In this section, we present results of the validation exercise discussed above in Section 2, with all outcome variables in natural log form.\textsuperscript{11} The goal of this section is to demonstrate that we can recover accurate estimates of the average treatment effect on our outcomes of interest, and that our simulation results appear centered at the “true” ATE. We then explore briefly how results vary with the size of the validation sample, which we explore more fully in the next section using fully simulated data.

We first present results from simulations in which we randomly sample 20% of farmers to use as a validation sample. Following the bias correction procedure, we use this validation sample to “correct” the self-reported variable in the larger sample, and calculate the resulting treatment effect estimate using information from both samples. Figure 1 plots the average treatment effects of program participation on log maize output, estimated using

\textsuperscript{11}\textsuperscript{See Appendix A for results presented using the two other self-reported cultivated land size variables, as well as Appendix B for results from simulations using levels of all outcomes rather than logs.}
GPS and self-reported acreage, and then show the mean and median result from the simulation exercise. Figures 2 show the analogous results for log maize yields per acre. These figures demonstrate that this procedure can return ATE estimates which are very close to the ATE estimated on the full sample, albeit with reduced precision.

![Figure 1: Simulation results, total maize output, full sample](image1)

Although the simulations produce ATE estimates which match the GPS estimates on average, if the distribution of estimates is wide, this kind of exercise may still not be very useful in practice. Figures 3 and 4 plot the distribution of ATE estimates on log total maize output and maize yields per acre from all simulations. In both cases, we see that the distribution of results is not only centered at or near the GPS estimates, but that the distribution is relatively tight and excludes the SR estimate.

Finally, a key issue in implementing this validation strategy in practice is the sample size required. If this data is very costly to collect, it’s important to understand how sensitive the point estimate and standard errors are to the size of the validation sample. Figures 5 and 6 show how the mean and median simulation result change when the validation sample...
Figure 3: Simulation result distribution, total maize output, full sample

Figure 4: Simulation result distribution, maize yield per acre, full sample
size is decreased to 10 or increased to 30%. Perhaps unsurprisingly, increasing the size of the validation sample improves the precision of the estimate.

Figure 5: Simulation results by validation sample size, total maize output, full sample

Figure 6: Simulation results by validation sample size, maize yield, full sample

5 Bias correction exercise with simulated data

In this section, we present evidence from a more general class of simulations, in which we specify a data-generating process, measurement error structure, and validation data sample size, and explore how statistical power varies by total sample size and validation sample size.

We draw $a \sim N(\mu_a, \sigma_a)$ as our intermediate outcome of interest (cultivated acreage) for
all $N$ individuals, and assume a constant treatment effect $\beta$ affects $a$ for treated individuals. We assume that individuals mis-report $a$ according to the following equation:

$$a^* = a + \gamma_0 + \gamma_1 T + \zeta$$

where $\zeta \sim \mathcal{N}(0,1)$ is a random noise parameter, $\gamma_0$ is a constant capturing mis-reporting by all individuals, and $\gamma_1$ represents differential mis-reporting by treated individuals.

Our first goal is to establish how our power to detect a treatment effect on the intermediate outcome ($a$) varies by total sample size, validation sample size, and various combinations of $\gamma_0, \gamma_1$, and $\beta$. Figure 7 plots simulation results for several parameter combinations. The lines show, for each parameter combination, the minimum validation sample size which resulted in at least an 80% rejection rate for each total sample size. These “iso-power” lines illustrate the trade-off between increasing validation sample size and increasing total sample size.

Figure 7: Trade-off between validation sample and full sample sizes required for 80% power to detect $\beta$ (average treatment effect on cultivated acreage)

Next, we demonstrate how this mis-measurement may impact analysis when the variable is part of a compound outcome, as above in the total output or yields per acre measure. We draw $y \sim \mathcal{N}(\mu_y, \sigma_y)$ as an additional intermediate outcome of interest (yields per acre) and assume again a constant treatment effect $\delta$ on $y$. Then, using the validation procedure described in section 2, we estimate the bias-corrected treatment effect across a range of parameter values. Figure 8 shows how the minimum validation sample size required to achieve an 80% rejection rate changes with total sample size. Again we see from the downward sloping iso-power lines that researchers face a trade-off between validation sample size.
and total sample size, and the final decision may depend on their expectations about the nature and magnitude of measurement error.

Figure 8: Trade-off between validation sample and full sample sizes required for 80% power to detect $\delta$ (average treatment effect on total output)

![Graph](image)

Each line shows for a given set of true parameter values, the minimum validation sample percentage that rejected the null at least 80% of the time for each sample size.

Future work will extend this exercise to account for heterogeneity in treatment effects (on intermediate outcomes and measurement error), as well as exploring how researchers can use stratified sampling for the validation sample to improve power.

6 Discussion

In this paper, we present preliminary results adapting validation data strategies from Buonaccorsi and Tosteson (1993) and Carroll et al. (2006) to the problem of treatment-differential measurement error in cultivated land size reporting. We provide additional Monte Carlo evidence of how the bias correction procedure performs under different relative magnitudes of treatment effect, common measurement error, and treatment-differential measurement error. These initial steps suggest this can be a useful toolkit for researchers seeking to design experiments where they face a trade-off between a costly, accurate measure and a cheap but possibly biased measure of an outcome of interest. Further work is needed to test this method in other contexts for which a similar measurement error problem exists, and ultimately to provide reliable guidelines to practitioners on how to implement this strategy to improve measurement in field experiments.
References


Appendix

A Tables and simulation results with alternative land size measures

A.1 First alternative land size variable: planting compliance survey

Table A.1: Measurement comparison by sample

<table>
<thead>
<tr>
<th></th>
<th>(1) Log GPS Acres</th>
<th>(2) Log SR Acres</th>
<th>(3) Log SR - Log GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>0.18***</td>
<td>0.00</td>
<td>-0.20***</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.050)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1852</td>
<td>1852</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>-0.26</td>
<td>-0.03</td>
<td>0.040</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. Log GPS Acres is the log of total cultivated maize area of each farmer, measured using GPS. Log SR Acres is the log of farmer self-reported value of total cultivated maize acres. Log SR - Log GPS is the difference between the two values. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

Table A.2: Log output comparison by sample

<table>
<thead>
<tr>
<th></th>
<th>(1) Log Output $H_{GPS}$</th>
<th>(2) Log Output $H_{SR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>0.36***</td>
<td>0.16**</td>
</tr>
<tr>
<td></td>
<td>(0.060)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1852</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>6.68</td>
<td>6.91</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. $H_{GPS}$ is log total maize output estimated using crop-cut yields and GPS-measured cultivated acreage. $H_{SR}$ is log total maize output estimated using crop-cut yields and self-reported cultivated acreage. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.
Table A.3: Log yield comparison by sample

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log Crop-cut Yield $y_{GPS}$</td>
<td>Log SR Yield $y_{SR}$</td>
</tr>
<tr>
<td>Treated</td>
<td>0.18*** (0.030)</td>
<td>0.39*** (0.050)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1852</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>6.94</td>
<td>6.69</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. Log Crop-cut Yield $y_{GPS}$ is measured directly using a crop cut from a 40 square meter box on farmer fields. Log SR Yield $y_{SR}$ is calculated by dividing estimated total output by self-reported cultivated acreage. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

Figure A.1: Simulation results, total maize output, full sample

Figure A.2: Simulation results, maize yield per acre, full sample
Figure A.3: Simulation result distribution, total maize output, full sample

Figure A.4: Simulation result distribution, maize yield per acre, full sample
Figure A.5: Simulation results by validation sample size, total maize output, full sample

Figure A.6: Simulation results by validation sample size, maize yield, full sample
A.2 Second alternative land size variable: input measurement re-survey

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log GPS Acres</td>
<td>Log SR Acres</td>
<td>Log SR - Log GPS</td>
</tr>
<tr>
<td>Treated</td>
<td>0.18***</td>
<td>0.11**</td>
<td>-0.07*</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.050)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1969</td>
<td>1969</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>-0.26</td>
<td>-0.03</td>
<td>0.040</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. Log GPS Acres is the log of total cultivated maize area of each farmer, measured using GPS. Log SR Acres is the log of farmer self-reported value of total cultivated maize acres. Log SR - Log GPS is the difference between the two values. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log Output $H_{GPS}$</td>
<td>Log Output $H_{SR}$</td>
</tr>
<tr>
<td>Treated</td>
<td>0.36***</td>
<td>0.28***</td>
</tr>
<tr>
<td></td>
<td>(0.060)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1969</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>6.68</td>
<td>6.71</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. $H_{GPS}$ is log total maize output estimated using crop-cut yields and GPS-measured cultivated acreage. $H_{SR}$ is log total maize output estimated using crop-cut yields and self-reported cultivated acreage. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.
Table A.6: Log yield comparison by sample

<table>
<thead>
<tr>
<th></th>
<th>(1) Log Crop-cut Yield $y_{GPS}$</th>
<th>(2) Log SR Yield $y_{SR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>0.18*** (0.030)</td>
<td>0.25*** (0.050)</td>
</tr>
<tr>
<td>Observations</td>
<td>1971</td>
<td>1969</td>
</tr>
<tr>
<td>Ctrl Mean</td>
<td>6.94</td>
<td>6.91</td>
</tr>
</tbody>
</table>

This table presents results from linear regressions of the variables named at the top of each column on a treatment dummy. Log Crop-cut Yield $y_{GPS}$ is measured directly using a crop cut from a 40 square meter box on farmer fields. Log SR Yield $y_{SR}$ is calculated by dividing estimated total output by self-reported cultivated acreage. Ctrl Mean is the mean of each outcome variable for the non-treated sample. Standard errors (in parentheses) are clustered at the treatment assignment (farmer group cluster) level.

Simulation results compared to full sample results

- ATE on log total maize output
  - GPS (full sample)
  - SR (full sample)
  - Simulation results (mean)
  - Simulation results (median)

Note: results from 1000 simulations with a 20% validation sample. Simulation standard errors bootstrapped with 100 replications.

Figure A.7: Simulation results, total maize output, full sample

Simulation results compared to full sample results

- ATE on log maize yields per acre
  - GPS (full sample)
  - SR (full sample)
  - Simulation results (mean)
  - Simulation results (median)

Note: results from 1000 simulations with a 20% validation sample. Simulation standard errors bootstrapped with 100 replications.

Figure A.8: Simulation results, maize yield per acre, full sample
Figure A.9: Simulation result distribution, total maize output, full sample

Figure A.10: Simulation result distribution, maize yield per acre, full sample
Simulation results (mean) compared to full sample results

Simulation results (median) compared to full sample results

ATE on log total maize output

Validation sample size

10% 20% 30%

Note: results from 1000 simulations. Simulation standard errors bootstrapped with 100 replications. Red vertical line represents ATE estimate from full sample.

Figure A.11: Simulation results by validation sample size, total maize output, full sample

Simulation results (mean) compared to full sample results

Simulation results (median) compared to full sample results

ATE on log maize yields per acre

Validation sample size

10% 20% 30%

Note: results from 1000 simulations. Simulation standard errors bootstrapped with 100 replications. Red vertical line represents ATE estimate from full sample.

Figure A.12: Simulation results by validation sample size, maize yield, full sample
B Simulation results with outcome variables as levels

Figure B.1: Simulation results, total maize output, full sample

Figure B.2: Simulation results, maize yield per acre, full sample
Note: results from 1000 simulations.

Figure B.3: Simulation result distribution, total maize output, full sample

Note: results from 1000 simulations.

Figure B.4: Simulation result distribution, maize yield per acre, full sample
Figure B.5: Simulation results by validation sample size, total maize output, full sample

Figure B.6: Simulation results by validation sample size, maize yield, full sample