Correlated Non-Classical Measurement Errors and 'Second Best' Policy Inference: The Case of The Inverse Size-Productivity Relationship in Agriculture

Kibrom A. Abay, University of Copenhagen
Gashaw T. Abate, IFPRI
Christopher B. Barrett, Cornell University
Tanguy Bernard, IFPRI

STAARS Workshop
African Development Bank Headquarters,
Abidjan, Côte d'Ivoire
December 12, 2017

Introduction

- Measurement drives analysis. But GIGO.
- Much research now on improved measurement of key variables
- Widespread prevalence of non-classical measurement error in key variables relevant to agriculture and beyond.

Research questions

- Yet often multiple variables suffer non-classical measurement error (NCME)
- Not all such variables are amenable to correction.
- What are the consequences for inference, especially if those measurement errors are correlated?
- Does correction for one, but not both, otherwise-mismeasured variables reduce bias and improve inference?

What we do

- Study correlated NCME in multiple variables generally

- Apply empirically to the size-productivity relationship (SPR) in agriculture
  - Long a metaphor for agricultural development policy
  - Earlier studies attribute empirical regularity of an inverse SPR to factor market imperfections (e.g. Sen, 1966; Feder, 1985; Barrett, 1996) or omitted land attributes (Benjamin, 1995; Assuncao and Braido, 2007; Barrett et al., 2010)
  - Recent studies attribute it to measurement error in land (Carletto et al., 2013; Carletto et al., 2015) or production (Gourlay et al., 2017; Desiere and Jolliffe, 2018) variables

- Key findings:
  - Bias due to NCME is analytically ambiguous
  - With correlated NCME, correcting for measurement error in just one variable can aggravate rather than attenuate bias in the SPR estimate. A ‘second best’ result.

Measurement Errors in Household Surveys

- Most micro research relies on self-reported, recall-based data prone to measurement error
- Recent studies show non-classical measurement error in self-reported land area (Carletto et al., 2013; Carletto et al., 2015) and production (Gourlay et al., 2017; Desiere and Jolliffe, 2018) and show that such inaccuracies can affect the estimated SPR
- No study has analyzed the implication of measurement errors in both metrics

Analytical Framework

- Consider the relationship between log-production and log-land area

$$Y^* = \theta X^* + \varepsilon$$

- Plot size-productivity (log yield-log area) relationship:

$$Y^* - X^* = \beta X^* + \varepsilon ; \beta \equiv \theta - 1$$

Effects of Measurement Error

- Assume additively entering log-transformed measurement error in self-reported area and production $Y = Y^* + u, X = X^* + v$

***Case 1**: Non-classical measurement error in dep variable (production)*

- Assume $u = \delta Y^* + \omega, \quad Y = (1+\delta)Y^* + \omega$
- Then $\beta^{OLS} = (1+\delta)\beta$
- For $\delta, \beta < 0$, NCME → SPR est. biased *upward* (attenuates inverse relationship)

Effects of Measurement Error

- Assume additively entering log-transformed measurement error in self-reported area and production $Y = Y^* + u, X = X^* + v$

Case 2*: Non-classical measurement error in dependent variable (production), error correlated with independent variable*

- Assume $u = \lambda X^* + \zeta$
- Then $\beta^{OLS} = \beta + \lambda$
- For β,λ<0, NCME → SPR est. biased *downward* (exaggerates inverse relationship)
- Empirical results: Desiere and Jolliffe (2018), Gourlay et al. (2017)

Effects of Measurement Error

- Assume additively entering log-transformed measurement error in self-reported area and production $Y = Y^* + u$, $X = X^* + v$

Case 3: *Non-classical measurement error in independent variable (land area)*

- Assume $v = \alpha X^* + \ell \Rightarrow X = (1+\alpha)X^* + \ell$
- Then $$\beta^{OLS} = \frac{(\beta-\alpha)(1+\alpha)\rho x_*^2}{(1+\alpha)^2 \rho x_*^2 + \rho_\ell^2} = \frac{\beta(1+\alpha)\rho x_*^2}{(1+\alpha)^2 \rho x_*^2 + \rho_\ell^2} - \frac{\alpha(1+\alpha)\rho x_*^2}{(1+\alpha)^2 \rho x_*^2 + \rho_\ell^2}$$

  where ρ^2 is a variance of that variable

- For α=0, this is classical ME → SPR estimate attenuated for $\rho_\ell^2 > 0$.
- More generally, α<0, NCME → *bias of ambiguous sign* in SPR estimate.

Effects of Measurement Error

- Assume additively entering log-transformed measurement error in self-reported area and production $Y = Y^* + u, X = X^* + v$

***Case 4**: Correlated non-classical measurement errors*

- Assume $Cov(u, v) = \pi$
- Then $$\beta^{OLS} = \frac{(\beta - \alpha)(1 + \alpha)\rho x_*^2 + \lambda \rho x_*^2 + \pi}{(1 + \alpha)^2 \rho x_*^2 + \rho_\ell^2}$$

- Generalizes other cases. For α=π=λ=0, this is classical ME → SPR est attenuated.
- Even if π=0, NCME (α ≠ 0 ≠ λ) means correcting one does not eliminate bias.
- Even if ME is classical (α=λ=0), π ≠ 0 leads to bias in SPR estimate of sign(π).
- Esp. if π>0, correcting just one NCME can exaggerate bias in SPR estimate. *It may be better to ignore both NCMEs than to correct just one.*

Effects of Measurement Error

Table 1: Summary of Analytical Results

Source of non-classical measurement error	Key Parameters				Estimated SPR	Direction of bias on the SPR
	δ	λ	α	π		
No error	0	0	0	0	β	No bias
Error in production	<0	0	0	0	$(1+\delta)\beta$	Underestimation of ISPR
Error in production	*	<0	0	0	$\beta+\lambda$	Overestimation of ISPR
Error in plot size	*	0	<0	0	$\beta(1+\alpha)\Phi-\alpha(1+\alpha)\Phi$	Ambiguous
Error in both	*	<0	<0	0	$\beta(1+\alpha)\Phi-\alpha(1+\alpha)\Phi-\lambda\Phi$	Ambiguous
Error in both	*	<0	<0	>0	$\beta(1+\alpha)\Phi-\alpha(1+\alpha)\Phi-\lambda\Phi+\pi\Phi$	Ambiguous

Empirical Demonstration: Data

- Data collected from rural wheat farmers in Ethiopia
- 36 *kebeles* from 18 *woredas* in Oromia, Amhara, and Tigray regions
- Random sample of one random wheat plot/farmer 2013/14 *meher* season
- Crop-cut measurement in November and December 2013 by experts from the Central Statistical Agency (CSA).
- Household survey conducted in February and March 2014
- Land area measures: self-reported, compass-and-rope
- Production measures: self-reported, crop-cut

Empirical Results: Measurement Errors

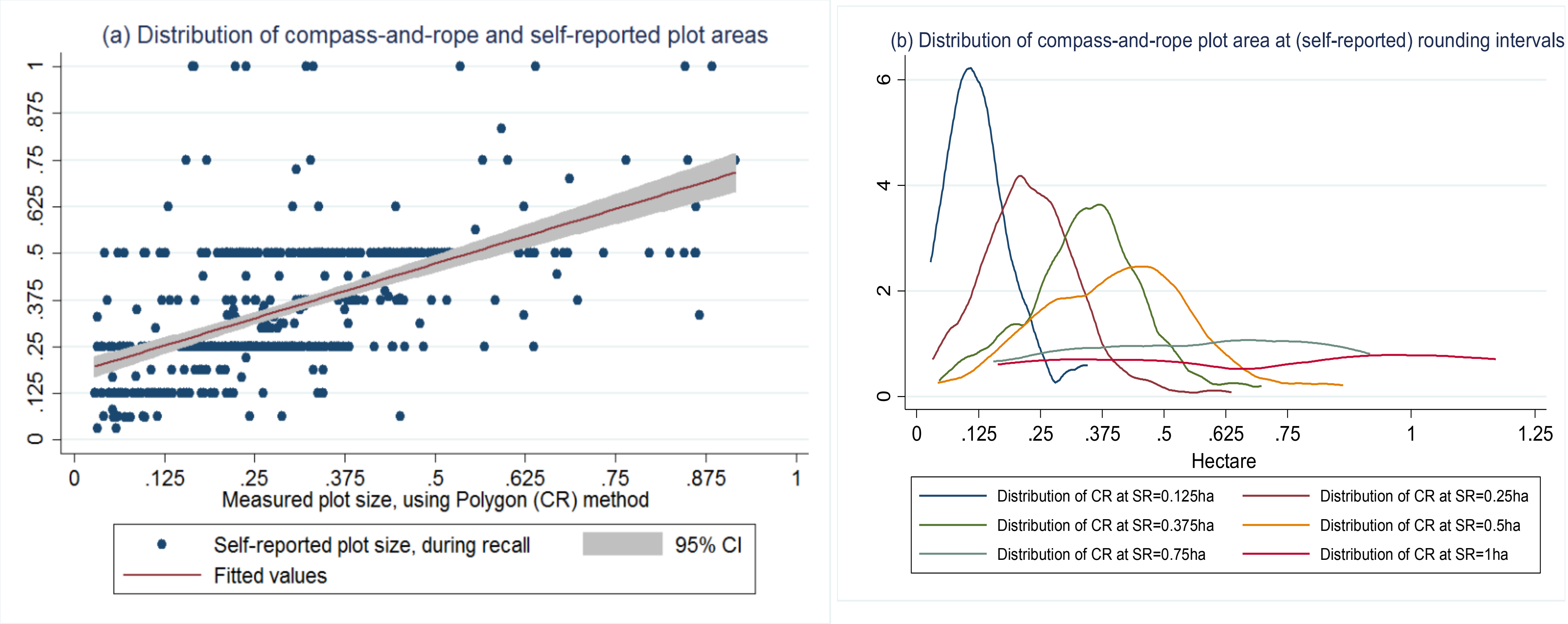
Table 3: Discrepancy between measured (CR) and self-reported (SR) plot size

Plot size group (CR)	Number of obs.	Self-Reported (SR) (1)	Compass-and-Rope (CR) (2)	Bias (SR) – (CR)		Difference in mean (p-value) (5)
				Bias=(1)-(2) (3)	%Bias=(3)/(2) (4)	
≤0.125 ha	70	0.20	0.08	0.12	150%	0.000
0.125–0.25 ha	132	0.31	0.19	0.12	63%	0.000
0.25–0.375 ha	125	0.38	0.30	0.08	27%	0.000
0.375–0.5 ha	74	0.46	0.44	0.02	5%	0.350
0.5–0.75 ha	46	0.60	0.58	0.02	3%	0.783
0.75–1 ha	12	0.64	0.85	–0.21	–25%	0.005
>1.0 ha	24	1.22	1.70	–0.48	–28%	0.019
Total	483	0.42	0.37	0.05	14%	0.002

Implication: $\alpha<0$

Empirical Results: Measurement Errors

Fig 2: Distribution of compass-and-rope (CR) and self-reported (SR) plot area

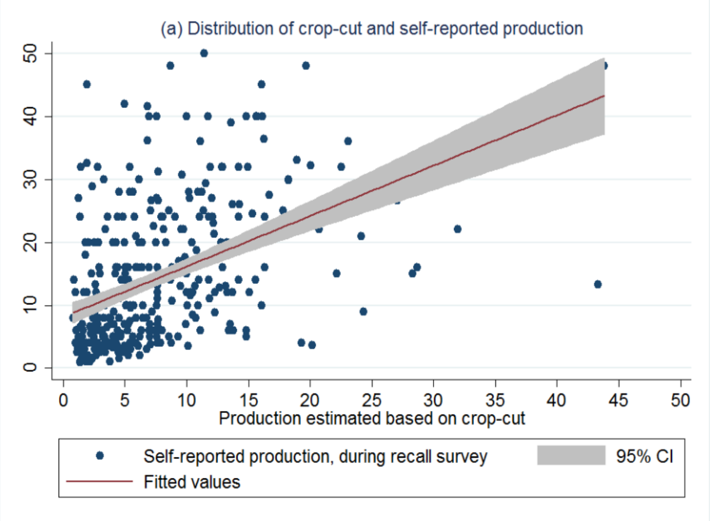


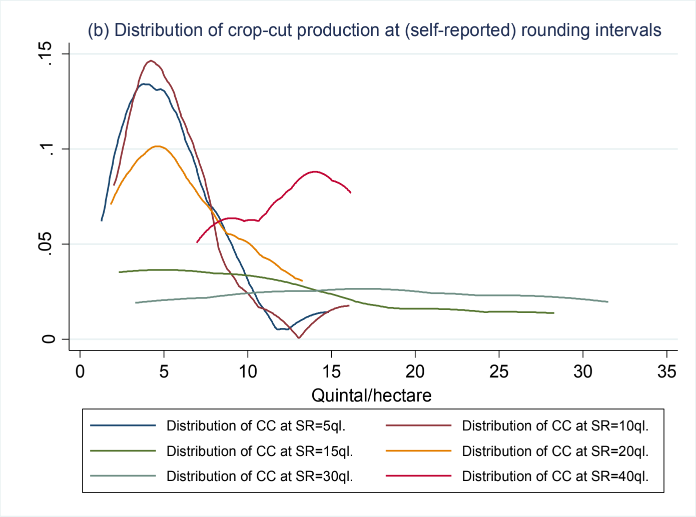
- Rounding: horizontal bunching of self-reported plot area around values that correspond to the conversion factor between the common local unit and hectare (e.g., 1 *oxen day*=0.25 ha) . *Implication: V(X)<V(X*).*

Empirical Results: Measurement Errors

Table 4: Discrepancy between crop cut (CC) and self-reported (SR) ouitput

Plot size group (CR)	Number of obs.	Self-reported (SR) (1)	Crop-cut (CC) (2)	Bias (SR) – (CC)		Difference in mean (p-value) (5)
				Bias=(1)-(2) (3)	%Bias=(3)/(2) (4)	
≤0.125 ha	59	9.1	2.6	6.5	250%	0.000
0.125–0.25 ha	108	13.9	5.6	8.3	148%	0.000
0.25–0.375 ha	87	16.3	7.7	8.6	111%	0.000
0.375–0.5 ha	50	19.1	11.7	7.4	63%	0.000
0.5–0.75 ha	33	26.1	13.6	12.5	91%	0.000
0.75–1 ha	9	24.2	21.8	2.3	10%	0.800
>1.0 ha	19	46.5	32.2	14.3	44%	0.064
Total	365	17.5	8.9	8.5	95%	0.000





Implication: $\delta<0$

Empirical Results: Measurement Errors

Table 5: Characterizing measurement errors in production

Explanatory variables	Dependent variable: ln (self-reported production/crop-cut production)		
	(1)	(2)	(3)
ln (crop-cut production)	-0.656***	-0.658***	-0.670***
	(0.055)	(0.043)	(0.055)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	0.987***	0.324	0.739
	(0.093)	(0.502)	(0.552)
Observations	365	365	360
R-squared	0.609	0.617	0.635

- Mean-reverting error in production
- *Implication: $\delta<0$, upward bias in (attenuation of) true SPR (case 1)*

Empirical Results: Measurement Errors

Table 6: Characterizing measurement errors in production

Explanatory variables	Dependent variable: ln (self-reported production/crop-cut production)		
	(1)	(2)	(3)
ln (compass-and-rope plot size)	-0.596***	-0.590***	-0.558***
	(0.073)	(0.058)	(0.080)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	-1.327***	-1.793***	-1.379**
	(0.147)	(0.574)	(0.608)
Observations	365	365	360
R-squared	0.495	0.501	0.516

- *Implication: $\lambda<0$, downward bias in (exaggeration of) true SPR (case 2)*
- The patterns in Table 6 drive those in Table 5, so case 2 should dominate

Empirical Results: Measurement Errors

Table 7: Characterizing measurement errors in land area

Explanatory variables	Dependent variable: ln (self-reported area/compass-and-rope plot size)		
	(1)	(2)	(3)
ln (compass-and-rope plot size)	-0.550***	-0.540***	-0.532***
	(0.045)	(0.044)	(0.042)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	-0.889***	-1.162***	-0.981**
	(0.090)	(0.341)	(0.415)
Observations	365	365	360
R-squared	0.463	0.494	0.518

- *Implication: α<0, ambiguous bias (case 3)*

Empirical Results: Measurement Errors

Table 8: Correlation between both types of measurement errors

Explanatory variables	Dependent variable: ln (self-reported production/crop-cut production)		
	(1)	(2)	(3)
ln (land area bias)	0.623***	0.542***	0.492***
	(0.091)	(0.105)	(0.114)
ln (CR Plot size)		-0.310**	-0.246*
		(0.125)	(0.126)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	-0.263***	-0.194***	-0.076
	(0.020)	(0.039)	(0.499)
Observations	365	365	360
R-squared	0.481	0.494	0.521

- *Implication: $\pi>0$, ambiguous bias (case 4)*
- A farmer who under-reports land area would likely do so production

Estimating the Size—Productivity Relationship

Table 9: Benchmark results: plot size-productivity relationship *(correcting for both area and production measurement errors)*

Explanatory variables	Dependent variable: ln (**crop-cut production/compass-and-rope plot size)**		
	(1)	(2)	(3)
ln (compass-and-rope plot size)	-0.083**	-0.086*	-0.104
	(0.040)	(0.042)	(0.063)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	3.542***	3.247***	3.351***
	(0.082)	(0.263)	(0.426)
Observations	365	365	360
R-squared	0.518	0.525	0.562

- Yield is truly statistically invariant to area or mild inverse relationship

Estimating the Size—Productivity Relationship

Table 10: Plot size-productivity relationship (correcting for area measurement only)

Explanatory variables	Dependent variable: ln (self-reported production/**compass-and-rope plot size**)		
	(1)	(2)	(3)
ln (compass-and-rope plot size)	-0.679***	-0.675***	-0.662***
	(0.079)	(0.083)	(0.074)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	2.215***	1.454**	1.972***
	(0.160)	(0.572)	(0.608)
Observations	365	365	360
R-squared	0.576	0.587	0.607

- **Correction for one variable only yields most biased ISPR estimate**

Estimating the Size—Productivity Relationship

Table 11: Plot size-productivity relationship *(correcting for production measurement only)*

Explanatory variables	Dependent variable: ln (**crop-cut production**/self-reported plot size)		
	(1)	(2)	(3)
ln (self-reported plot size)	-0.410***	-0.404***	-0.578***
	(0.067)	(0.070)	(0.077)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	2.752***	2.921***	1.682***
	(0.121)	(0.389)	(0.443)
Observations	365	365	360
R-squared	0.403	0.424	0.535

- **Exaggerated ISPR estimated when correct other variable only**

Estimating the Size—Productivity Relationship

Table 12: Plot size—productivity relationship *(with no correction of measurement errors)*

Explanatory variables	Dependent variable: ln (self-reported production/self-reported plot size)		
	(1)	(2)	(3)
ln (self-reported plot size)	-0.154**	-0.155**	-0.204***
	(0.062)	(0.061)	(0.073)
Household characteristics	No	Yes	Yes
Plot characteristics	No	No	Yes
Village level dummies	Yes	Yes	Yes
Constant	7.693***	7.128***	7.060***
	(0.111)	(0.476)	(0.566)
Observations	365	365	360
R-squared	0.459	0.465	0.476

- **Biased (exaggerated) ISPR estimate but statistically significantly lower bias than when correct just one variable.**

Table 13: Summary of Empirical Relationships

Source of non-classical measurement error	Key empirically estimated parameters				Estimated SPR	Relative implication on the SPR
No error	NA	NA	NA	NA	-0.104 (0.063)	Insignificant ISPR estimated
Error in production	-0.670*** (0.055)	-0.558*** (0.080)	NA	NA	-0.659*** (0.074)	Strongest ISPR estimated
Error in plot size	NA	NA	-0.532*** (0.042)	NA	-0.578*** (0.077)	Strong ISPR estimated
Error in both	-0.670*** (0.055)	-0.558*** (0.080)	-0.532*** (0.042)	0.492*** (0.114)	-0.204*** (0.073)	Weaker ISPR estimated

Compared to the benchmark:

- **Strong ISPR estimated when we ignore either error**
- **Weak ISPR estimated when we ignore both**
- **Consistent with our generic analytical predictions**

Key Results and Concluding Remarks

- We analytically and empirically study correlated non-classical measurement errors
- We show that the signs and magnitude of resulting biases are ambiguous
- We show that accounting for measurement error in only one of the variables may worsen the bias, yielding results that are misleading for informing policy.
- Correction of just one of several NCMEs may be inferior to a "second best" approach based on multiple variables measured with error.

Thank you for your interest and comments

Variable	Description	Mean	Std. Dev.	Min	Max	Obs.
Area SR	Self-reported area size for reference plot (ha)	0.42	0.36	0.03	4.00	488
Area measured	Measured area size during crop-cut (ha)	0.37	0.39	0.03	3.80	483
Production SR	Self-reported production for reference plot (qt.)	21.05	19.18	0.50	120.00	488
Production measured	Estimated production based on crop-cut (qt.)	8.98	9.91	0.81	101.5	365
Yield SR	Self-reported (production/area), (qt./ha)	30.69	18.18	1.00	96.00	488
Yield measured	Measured (production/area), (qt./ha)	28.23	15.05	2.78	95.38	366
Age of HH head	Age of the household head in completed years	45.67	10.84	20.00	77.00	488
Gender of HH head	Gender of the household head	0.86	0.34	0.00	1.00	488
HH size	Number of household members	6.79	2.39	1.00	16.00	488
Literacy of HH head	=1 if the household head is literate	0.64	0.48	0.00	1.00	488
No. of corners	Number of corners of the reference plot	8.74	4.88	4.00	23.00	484
Closure error	Closure error in plot area measurement	1.09	0.89	0.02	4.50	483
Area unit†	=1 if farmers used ha for SR area measurement	0.39	0.49	0.00	1.00	488
Total owned area†	Total farm land owned by sample farmers	2.31	2.14	0.00	20.00	488
Crop-cut to edge	Distance between the crop-cut and shortest or closest plot edge (meters)	25.83	18.57	1.40	148.00	374
Production unit	=1 if farmers used kg for SR production measurement	0.59	0.49	0.00	1.00	488
Total wheat produced†	Total wheat production during 2013/14 *meher*	46.64	75.26	0.95	755.00	488
Soil fertility†						
High	=1 if the fertility of the reference plot is high	0.44	0.49	0.00	1.00	488
Medium	=1 if the fertility of the reference plot is medium	0.49	0.50	0.00	1.00	488
Poor	=1 if the fertility of the reference plot is poor	0.07	0.26	0.00	1.00	488
Soil color†						
Red	=1 if the color of the reference plot is red	0.26	0.44	0.00	1.00	488
Black	=1 if the color of the reference plot is black	0.54	0.49	0.00	1.00	488
Grey/sand	=1 if the color of the reference plot is grey or sandy	0.20	0.40	0.00	1.00	488
Distance to plot†	Walking distance between the dwelling and the plot	30.98	9.94	0.00	120.00	488
Plot ownership	=1 if the reference plot owned by the HH	0.82	0.38	0.00	1.00	488