Title

Data-driven Development: Essays on the Use of Mobile Phone Data and Information to Measure and Reduce Poverty

Author

On, Robert

Publication Date

2018

Peer reviewed|Thesis/dissertation

Data-driven Development: Essays on the Use of Mobile Phone Data and Information to Measure and Reduce Poverty

by

Robert On

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Information Management and Systems

and the Designated Emphasis

in

Development Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joshua E. Blumenstock, Chair

Professor Michael Kremer

Professor Edward A. Miguel

Professor Hal R. Varian

Fall 2018

Abstract

Data-driven Development: Essays on the Use of Mobile Phone Data and Information to Measure and Reduce Poverty

By

Robert On

Doctor of Philosophy in Information Management and Systems

Designated Emphasis in Development Engineering

University of California, Berkeley

Professor Joshua E. Blumenstock, Chair

Mobile phone ubiquity in much of the developing world has turned from a question of when rather than if. Some of the poorest and most remote parts of the world are being connected to the global telecommunications network to enable an unprecedented ability to both observe and interact with previously hard-to-reach populations at scale. While many mobile phone owners adopt this technology for basic phone use, the connectedness this expansive ownership enables presents an opportunity to the study and practice of economic development that extend beyond simple peer-to-peer communication.

The modern information technology sector and its underlying network infrastructure presented this same opportunity during its own formation. The network was not only valuable for the communication it enabled, but also for the data it produced from those who utilized its services. It also serves as a platform for a deluge of information systems and services that have become a part of our everyday lives and has spurred significant economic growth over the past few decades. This "data revolution" is well underway in the developed economies but is diminishing in its returns, solving increasingly marginal problems. This same transformation is relatively nascent in developing economies where more salient challenges, such as poverty, have yet to be overcome. In this dissertation, we explore a data-driven approach that leverages mobile phone technology to better measure and address poverty in sub-Saharan Africa.

Our approach starts with the identification of a problem: in this case, poverty. In the first chapter, we apply novel machine learning methods to analyze roughly ten terabytes of data of mobile phone use from Rwanda's largest telecommunications operator to measure poverty at a national scale. We demonstrate that an individual's history of mobile phone usage can be used to infer his or her socioeconomic status. Using this individual model of mobile phone use and socioeconomic status, we can predict poverty and wealth across the entire network and accurately reconstruct national and regional distributions of wealth. Once we obtain this measure of poverty, we can then focus our efforts in regions that are most afflicted.

The second chapter helps moves us from diagnosis to a potential cure. Predictions may be helpful to provide some guidance on which regions or populations to target but does not provide much in the way of what to do to have impact. In three years of field research in poor

regions of rural Kenya and Rwanda, it was clear that much of the world's poor thrive and survive on subsistence agriculture, but many of these farmers also own mobile phones. Having such a platform enabled the ability to provide potentially welfare-improving information at scale. This chapter presents the research design and analyzes the results of of six randomized controlled trials testing the welfare effects of sending hundreds of text message formulations encouraging agricultural experimentation to over 500,000 farmers in Kenya and Rwanda. Targeting farmers with the right messaging and delivery characteristics was a focus of these trials. We find statistically significant effects on agricultural technology adoption and high rates of return on welfare outcomes by providing information over this medium. This mirrors the digital advertising industry in many developed economies and reminds us that advertisements as information can have very large welfare effects in poor information environments.

The third chapter dives deeper into one of the six studies where the research design focused on information spillover in Rwanda where mobile phone ownership was about half of what it was in Kenya. We find that information does indeed spillover onto other farmers within the same group, and those farmers who don't have phones experience the largest percentage increases in adoptions when others within the same group receive a text message. This has large implications on the effectiveness and cost efficiency of information treatments to regions with lower mobile phone adoption. Not only were these interventions effective, they were also very inexpensive and resulted in network effects, further improving agricultural technology adoption, increasing food production and reducing poverty.

The chapters in this dissertation develop a theory and methods for understanding how to leverage mobile technologies to measure and reduce poverty. It serves as a guide for both research and practitioners to approach solving problems in development that is grounded in measurement, data, collaboration, impact and scale.

1 Predicting Poverty and Wealth from Mobile Phone Metadata

with Joshua Blumenstock and Gabriel Cadamuro[1]

1.1 Abstract

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer sources of big data exist. We show that an individual's past history of mobile phone use can be used to infer his or her socioeconomic status. Furthermore, we demonstrate that the predicted attributes of millions of individuals can, in turn, accurately reconstruct the distribution of wealth of an entire nation or to infer the asset distribution of microregions composed of just a few households. In resource-constrained environments where censuses and household surveys are rare, this approach creates an option for gathering localized and timely information at a fraction of the cost of traditional methods.

1.2 Article

Reliable, quantitative data on the economic characteristics of a country's population are essential for sound economic policy and research. The geographic distribution of poverty and wealth is used to make decisions about resource allocation and provides a foundation for the study of inequality and the determinants of economic growth (1, 2). In developing countries, however, the scarcity of reliable quantitative data represents a major challenge to policy-makers and researchers. In much of Africa, for instance, national statistics on economic production may be off by as much as 50% (3). Spatially disaggregated data, which are necessary for small-area statistics and which are used by both the private and public sector, often do not exist (4, 5).

In wealthy nations, novel sources of passively collected data are enabling new approaches to demographic modeling and measurement (6–8). Data from social media and the "Internet of Things," for instance, have been used to measure unemployment (9), electoral outcomes (10), and economic development (8). Although most comparable sources of big data are scarce in the world's poorest nations, mobile phones are a notable exception: They are used by 3.4 billion individuals worldwide and are becoming increasingly ubiquitous in developing regions (11).

Here we examine the extent to which anonymized data from mobile phone networks can be used to predict the poverty and wealth of individual subscribers, as well as to create high-resolution maps of the geographic distribution of wealth. That this may prove fruitful is motivated by the fact that mobile phone data capture rich information, not only on the frequency and timing of communication events (12) but also reflecting the intricate structure of an individual's social network (13, 14), patterns of travel and location choice (15–17), and histories of consumption and expenditure. Regionally aggregated measures of phone penetration and use

[1] Blumenstock: Information School, University of Washington; Cadamuro: Department of Computer Science and Engineering, University of Washington. Corresponding author: joshblum@uw.edu

have also been shown to correlate with regionally aggregated population statistics from censuses and household surveys (8, 18, 19).

Our approach is different from prior work that has examined the relation between regional wealth and regional phone use, as we focus on understanding how the digital footprints of a single individual can be used to accurately predict that same individual's socioeconomic characteristics. This distinction is a scientific one, which also has several important implications: First, it allows for the method to be used in contexts for which recent census or household survey data are unavailable. Second, when an authoritative source of data does exist, it can be used to more objectively validate or refute the model's predictions. This limits the likelihood that the model is overfit on data from a single source, which is otherwise difficult to control, even with careful cross-validation (20). Third, our approach allows for a broad class of potential applications that require inferences about specific individuals instead of census tracts. As we discuss in the supplementary materials (section 6), future iterations of this approach could help to improve the targeting of humanitarian aid and social welfare, disseminate information to vulnerable populations, and measure the effects of policy interventions.

For this study, we used an anonymized database containing records of billions of interactions on Rwanda's largest mobile phone network and supplemented this with follow-up phone surveys of a geographically stratified random sample of 856 individual subscribers. Upon contacting and surveying each of these individuals, we received informed consent to merge their survey responses with the mobile phone transaction database. The surveys solicited no personally identifying information but contained questions on asset ownership, housing characteristics, and several other basic welfare indicators. From these data, we constructed a composite wealth index using the first principal component of several survey responses related to wealth (21, 22) (supplementary materials section 1D). For each of the 856 respondents, we thus have ~75 survey responses, as well as the historical records of thousands of phone-based interactions such as calls and text messages (Table 1).

	Phone survey	**Call detail records**	**DHS (2007)**	**DHS (2010)**
Number of unique individuals	856	1.5 million	7,377	12,792
Data collection period	July 2009	May 2008 - May 2009	Dec. 2007- Apr. 2008	Sep. 2010 – Mar. 2011
Number of questions in survey	75	N/A	1615	3396
Primary geographic units	30 districts	30 districts	30 districts	30 districts
Secondary geographic units	300 cell towers	300 cell towers	247 clusters	492 clusters

Table 1. Summary statistics for primary data sets. Phone survey data were collected by the authors in Kigali, in collaboration with the Kigali Institute of Science and Technology. Call detail records were collected by the primary mobile phone operator in Rwanda at the time of the phone survey. Demographic and Health Survey (DHS) data were collected by the Rwandan National Institute of Statistics. N/A, not applicable.

We use the merged data from this sample of 856 phone survey respondents to show that a mobile phone subscriber's wealth can be predicted from his or her historical patterns of phone use (Fig. 1A) (cross-validated correlation coefficient r = 0.68).

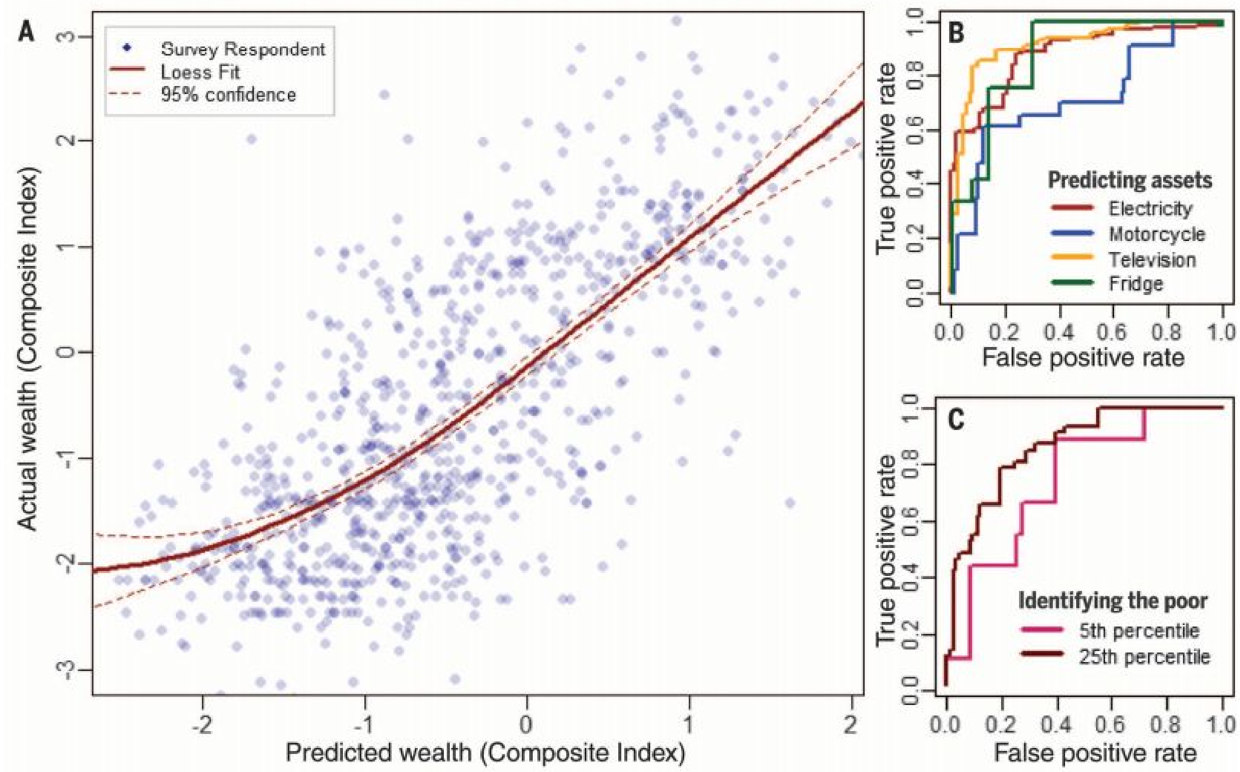


Fig. 1. Predicting survey responses with phone data. (A) Relation between actual wealth (as reported in a phone survey) and predicted wealth (as inferred from mobile phone data) for each of the 856 survey respondents. (B) Receiver operating characteristic (ROC) curve showing the model's ability to predict whether the respondent owns several different assets. AUC values for electricity, motorcycle, television, and fridge, respectively, are as follows: 0.85, 0.67, 0.84, and 0.88. (C) ROC curve illustrates the model's ability to correctly identify the poorest individuals. The poor are defined as those in the 5th percentile (AUC = 0.72) and the 25th percentile (AUC = 0.81) of the composite wealth index distribution.

Our approach to modeling combines feature engineering with feature selection by first transforming each person's mobile phone transaction logs into a large set of quantitative metrics and then winnowing out metrics that are not predictive of wealth. The first step employs a structured, combinatorial method to automatically generate several thousand metrics from the phone logs that quantify factors such as the total volume, intensity, timing, and directionality of communication; the structure of the individual's contact network; patterns of mobility and migration based on geospatial markers in the data; and so forth. The second step uses "elastic net"regularization to eliminate irrelevant phone metrics and select a parsimonious model that is more likely to generalize (23). We use cross-validation to limit the possibility that the model is overfit on the small sample on which it is trained. In the supplementary materials (section 3B), we provide details on these methods and show that comparable results are obtained under a variety of alternative supervised-learning models, including tree-based ensemble regressors and classifiers (24). We also show that this two-step approach to feature engineering and model selection performs significantly better than a more intuitive approach based on a small number of hand-crafted metrics (table S1).

In addition to predicting composite wealth, this same approach can be used to estimate, with varying degrees of accuracy, how a phone survey participant will respond to any question, such as whether the respondent owns a motorcycle or has electricity in the household (Fig. 1B and table S1). Cross-validated area-under-the-curve (AUC) scores—which indicate the probability that the model will rank a randomly chosen positive response higher than a randomly chosen negative one—range from 0.50 (no better than random) to 0.88 (quite effective). An analogous method can be used to accurately identify the individuals in the sample who are living below a relative poverty threshold (AUC = 0.72 to 0.81) (Fig. 1C). With further refinement, such methods could prove useful to policy-makers and organizations that target resources to the extreme poor (25) (supplementary materials section 6).

For each of these prediction tasks, we use the two-step procedure to select a different model with different metrics and parameters. Although not the focus of our analysis, we note discernible patterns in the set of features identified as the best joint predictors of these different response variables. For instance, features related to an individual's patterns of mobility are generally predictive of motorcycle ownership, whereas factors related to an individual's position within his or her social network are more useful in predicting poverty and wealth (fig. S3). These results suggest that our approach might be generalized to predict a broader class of survey responses, such as the subjective opinions and perceptions of mobile subscribers.

Having fit and cross-validated the model on the phone survey sample—a sample drawn to be representative of all active mobile phone users—we next generate out-of-sample predictions for the characteristics of the remaining 1.5 million Rwandan mobile phone users who did not participate in the survey. Combined with the rich geospatial markers in the phone data, the predicted attributes of millions of individual subscribers enable us to study the geographic distribution of subscriber wealth at an extremely fine degree of spatial granularity (Fig. 2). Whereas public data from Rwanda are only accurate at the level of the district (of which there are 30), the phone data can be used to infer characteristics of each of Rwanda's 2148 cells, as well as small microregions of just a few mobile subscribers (Fig. 2, bottom right inset).

The accuracy of these micro-regional wealth estimates cannot be directly verified, because no other data set provides wealth information with sufficient geographic resolution. However, when further aggregated to the district level, we can compare the distribution of wealth predicted from the call records of mobile subscribers (Fig. 3A) to the distribution of wealth measured with "ground truth" data collected by the Rwandan government (Fig. 3B).

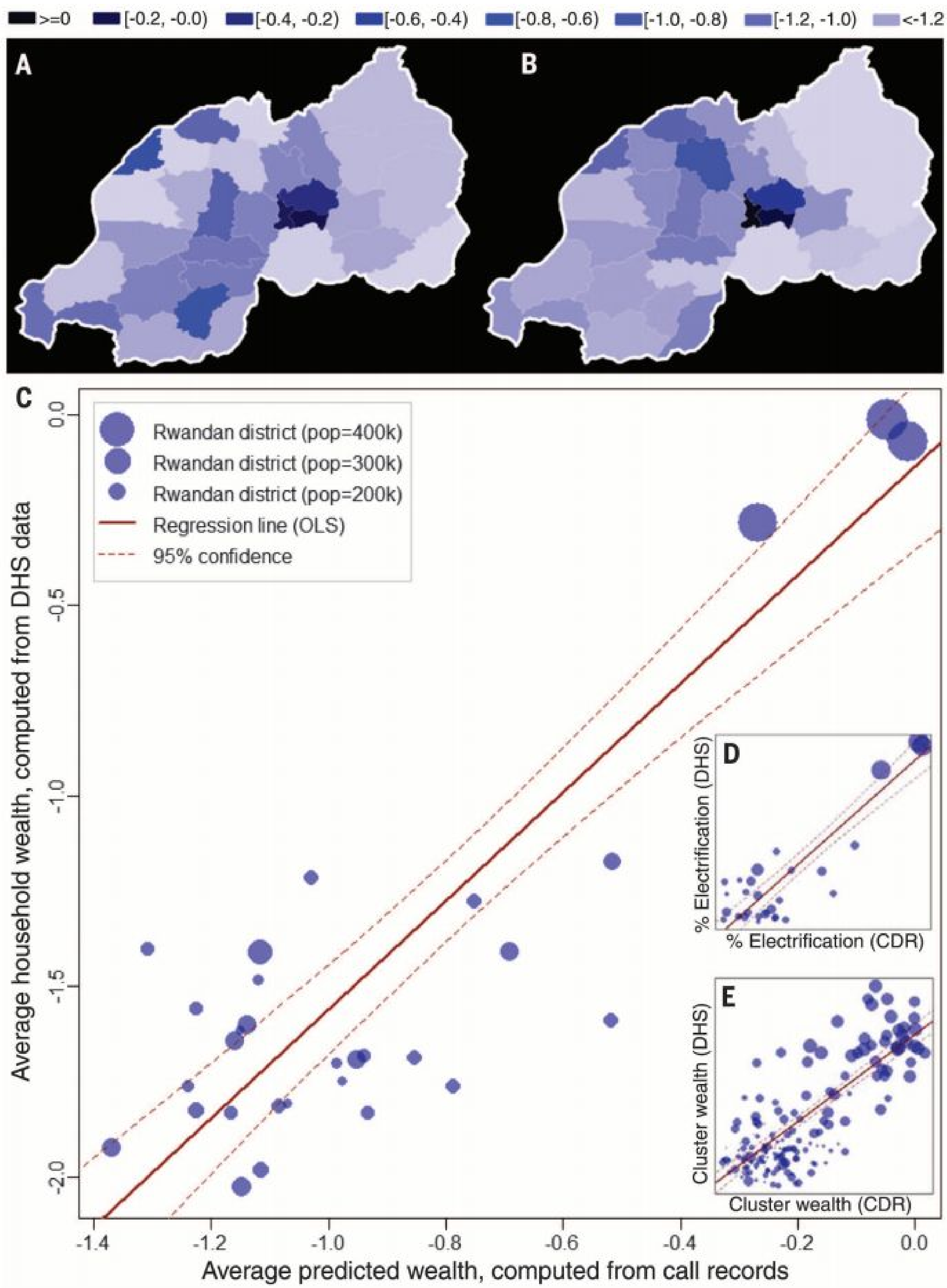


Fig. 3. Comparison of wealth predictions to government survey data. (A) Predicted composite wealth index (district average), computed from 2009 call data and aggregated by administrative district. (B) Actual composite wealth index (district average), as computed from a 2010 government DHS of 12,792 households. (C) Comparison of actual and predicted district wealth, for each of the 30 districts, with dots sized by population. (D) Comparison of actual and predicted rates of electrification, for each of the 30 districts. (E) Comparison of actual and predicted cluster wealth, for each of the 492 DHS clusters.CDR, call detail records.

The former estimates are computed by averaging predicted wealth across the thousands of individual mobile phone–based predictions in each of Rwanda's 30 districts; the latter estimates are calculated using data from a nationally representative Demographic and Health Survey (DHS) of 12,792 households, conducted in person by the National Institute of Statistics of Rwanda (26). The strong correlation between these two predictions is evident in Fig. 3C and exists whether the ground truth is estimated from only those DHS households that report owning a mobile phone ($r = 0.917$) or from all households in the survey ($r = 0.916$). As we discuss in the supplementary materials (section 5A), the first correlation shows that the model's out-of-sample predictions are representative of the population of Rwandan mobile phone owners. The second correlation indicates that in countries like Rwanda, where patterns of mobile phone adoption are similar across regions, this method can provide a close approximation of the distribution of wealth of the full national population. Similar results are obtained when the analysis is disaggregated to the level of the DHS "cluster" ($r = 0.79$) (Fig. 3E), a geographic unit designed to be comparable to a village. These strong correlations are partially driven by the stark differences between urban and rural areas in Rwanda, but the correlations persist even when comparing clusters within urban or rural areas (fig. S6).

This same approach can be used to predict more than just the average wealth of a district. For instance, rates of district electrification estimated from phone records are comparable to those reported in the DHS survey ($r = 0.93$) (Fig. 3D). In the urban capital of Kigali, we also find a correlation ($r = 0.58$) between satellite estimates of night light intensity in 0.55-km2 grid cells (fig. S7B) and the predicted distribution—based on phone data and the methods described earlier—of responses to the question "Does your household have electricity?" (fig. S7C).

How might such methods be used in practice? In addition to small-area estimation, one promising application is as a source of low-cost, interim national statistics. In many developing economies, long lag times typically occur between successive national surveys. In Angola, for instance, the most recent census before 2014 was conducted in 1970. In that 44-year period, the official population grew by more than 400%. Rwanda has better resources for data collection, and the DHS preceding the 2010 DHS was conducted in 2007. However, even in that relatively short period, the distribution of wealth in Rwanda shifted slightly. Thus, we find that the 2010 distribution of wealth is more accurately reflected in projections based on our analysis of phone data from 2009 than in estimates based on the 2007 DHS (fig. S8). This implies that a policy-maker tasked with targeting the poorest districts in Rwanda would obtain more accurate information from estimates based on mobile phone data than from estimates based on 2007 DHS data (supplementary materials section 6A).

In developing economies, where traditional sources of population data are scarce but mobile phones are increasingly common, these methods may provide a cost-effective option for measuring population characteristics. Whereas a typical national household survey costs more than $1 million and requires 12 to 18 months to complete (27), the phone survey we conducted cost only $12,000 and took 4 weeks to administer. Looking forward, the greatest challenge to such work lies in identifying protocols that enable analysis of similar data while respecting the privacy of individual subscribers and the commercial concerns of mobile operators (28, 29).With careful consideration, however, many compelling (and some speculative) applications are within reach, including population monitoring in remote and inaccessible regions, real-time policy evaluation, and the targeting of resources to those with the greatest need.

1.3 Supplementary Material

1.3.1 Data Description and Construction

1.3.1.1 Phone survey administration

In Summer 2009, we coordinated a phone survey of a geographically stratified group of Rwandan mobile phone users. Using a trained group of enumerators from the Kigali Institute of Science and Technology (KIST), a short, structured interview was administered to roughly 900 active mobile phone subscribers. The survey instrument contained approximately 80 questions that focused on basic socioeconomic and demographic information, including asset ownership and household and housing characteristics (Table 1). Several of these questions were drawn directly from the survey instruments used by the National Institute of Statistics of Rwanda in their Demographic and Household Surveys (DHS), which is described in greater detail below. Aside from the phone number of the respondent, we did not solicit any personally identifying information such as first name, last name, or address.

Full details on the administration of this phone survey are discussed in (30). In brief, the survey population was intended to be a representative sample of active subscribers on Rwanda's largest mobile phone network. At the time, the operator had roughly 90 percent market share, and 1.5 million registered Subscriber Identification Modules (SIM cards). However, since the number of registered SIMs greatly exceeds the number of active subscribers, we eliminated numbers which had not been used at least once in each of the three most recent months for which mobile phone data was available (October through December 2008). Each of the remaining 800,000 numbers was assigned to a geographic district based on the location of the phone for the majority of calls made (see SM Section IVA for details). The final sample was a geographically stratified random set of these numbers, with sampling weights determined by the distribution of active subscribers across districts (30).

Enumerators made three attempts to contact each respondent, on different days and at different times of day. Respondents were compensated RWF500 (roughly US$1) for participating in the survey, which took between 10 and 20 minutes to administer. Survey enumerators requested informed consent from each respondent, in which the goals of the study were described and oral permission was received to merge survey responses with anonymized call records, in accordance with the protocols of our university's ethical review board.

The contact rate was roughly 61%; non-contacts were largely the result of phones that were turned off or disconnected. The cooperation rate was 97%; almost everyone who picked up the phone was enthusiastic to participate in a study with university researchers, with whom they generally had little prior contact. We thus interpret the survey sample as representative of the population of active mobile phone subscribers, who we assume are systematically different from both the population of mobile phone subscribers and the general Rwandan population. In SM Section V, we discuss in greater detail the extent to which the non-representativeness of the phone survey sample affects our results.

1.3.1.2 Mobile phone call detail records

From Rwanda's near-monopoly mobile phone operator, we obtained a complete historical log of call detail records (CDR), which contain basic metadata on all transactions mediated by the mobile phone network. The logs included all domestic and international calls, as well as every text message (SMS) sent and received on the network, from early 2005 to mid-2009. For each of these transactions, we observe the time and date of the call, the anonymized but unique identifier of the calling and receiving party, the duration of the call, as well as the cellular towers through which the call was routed. As described in greater detail in SM Section IVA, information on these cellular towers can be used to infer the approximate location of both the caller and the receiver at the time of the call. For the sample of phone survey respondents who completed the survey, information from the mobile operator was provided to match the true phone number to the anonymized identifier in the CDR dataset.

1.3.1.3 Demographic and Health Surveys (DHS)

To provide further validation of the external validity of this method, we compare out-of-sample wealth predictions to "ground truth" Demographic and Health Surveys (DHS) collected by the National Institute of Statistics of Rwanda. Two rounds of these surveys are used in our analysis: DHSV, which was conducted between December 2007 and April 2008 on a sample of 7,377 households; and DHSVI, conducted between September 2010 and March 2011 on a sample of 12,792 households.

The DHS surveys are conducted with a nationally representative sample of households. Villages were selected with probability proportional to village size, and households are given survey weights to allow for reconstruction of nationally representative statistics (31). DHSV contained 247 village clusters, while DHSVI contained 492. The geographic coordinates of each cluster's centroid are also provided with the DHS data. However, as noted in the DHS documentation, "the data are randomly displaced up to 5 kilometres in rural areas and up to 2 kilometres in urban areas. A further 1 percent of rural clusters are displaced up to 10 kilometres." These displacements add considerable measurement error to subregional estimates of wealth, but should not estimates aggregated at the district level, as is the case in most of our analysis.

1.3.1.4 Composite wealth index construction

In Rwanda, as in most developing countries, it is difficult to estimate the socioeconomic status of a survey respondent with a single survey question. Instead, household surveys typically rely on a large number of questions which can be used to infer the consumption or permanent income of the respondent (32). The Rwandan DHS, for instance, contains roughly seventy questions related to household assets, characteristics, and expenditures. The first principal component of these responses is commonly treated as a proxy indicator of the respondent's unobserved wealth (21).

In our phone surveys, which were designed to be very short, we did not have the option of asking such a large number of questions related to assets and housing characteristics. Instead, we selected the subset of questions that, in the DHS data, were most highly correlated with the first principal component of the full set of DHS responses. We further excluded questions that would be difficult to administer in a phone survey (e.g., in our piloting we found that most respondents were unable to quickly ascertain how much land they owned). The final set of asset-related questions is listed in Table S1B. We also include the size of the household and the number of children, but all results are robust to the exclusion of these factors.

We compute the "composite wealth index" as the first principal component of the asset and household characteristics questions in our phone survey (21). The basis vectors W of the covariance matrix are estimated using weighted principal component analysis on the normalized data from the 856 phone survey respondents, where the weights are determined as described in SM Section 1A above (33). The first principal component captures 26 percent of the total variation in assets and household characteristics. When we later validate the phone-based predictions against data collected through government surveys (SM Section IV), we use the same basis vectors W computed on the phone survey data to project each DHS household's asset responses onto an analogous composite wealth index.

1.3.1.5 Satellite data

We validate the phone-based predictions of regional electrification using data on satellite "night lights" using average radiance composite images from the Visible Infrared Imaging Radiometer Suite Day/Night Band (VIIRS-DNB). The VIIRS-DNB imagery recognizes wavelengths from green to near-infrared, and is preprocessed by the National Oceanic and Atmospheric Administration to remove stray light and emphasize light from cities. The satellite data is provided at the resolution of 0.742km x 0.742km grid cells, and is measured in units of nanowatts/cm 2 /square radian (34).[2]

1.3.2 Feature engineering

Our goal in engineering features is to transform an individual's mobile phone transaction logs into a set of quantitative metrics that in turn can be used to infer that same individual's economic state. In the related literature, the most common approach has been to carefully construct a small number of intuitive indicators from the phone metrics, and compare regional aggregates of those phone metrics to regional socioeconomic indicators. In such work, for instance, there is evidence that the geographic diversity and reciprocal nature of social relationships are both correlated with economic outcomes (8, 35–38).

Our goal is different. We seek to develop measures of poverty and wealth that maximize predictive accuracy, possibly at the expense of the interpretability of the model. Thus, instead of devising a parsimonious set of metrics based on intuition, we take a brute force to feature engineering that is designed to capture as much variation as possible from the raw call detail records. Specifically, we develop a method based on a deterministic finite automaton (DFA) (39) to generate a large number of potentially correlated metrics, and then rely on regularization and related techniques to eliminate redundant metrics from the model. The primary advantage of using the DFA is that is restricts the number of degrees of freedom in the hands of the researcher; rather than specifying hundreds or thousands of "features" one by one, the DFA allows the researcher to specify a small number of different operations, which are then recursively applied to generate a large number of features.

[2] April 2012 version. These data were obtained from the NOAA National Geophysical Data Center, Earth Observation Group.

1.3.2.1 Baseline models: Single feature and top-5 features

In addition to the combinatoric deterministic finite automaton (DFA) described below, we implement two simple approaches to establish baselines for comparison. The first is a an "intuitive" model, which consists of five hand-picked features based loosely on related work (8, 35–38), and which are chosen to capture a variety of the behaviors reflected in mobile phone transaction logs. These features are: (i) the total number of calls in which the individual is involved (outgoing + incoming); (ii) the total number of text messages; (iii) the total number of international calls; (iv) the degree centrality of the individual (i.e., the total number of unique contacts with whom the individual interacts); and (v) the Radius of Gyration, a measure of the average travel distance of the individual (15). A cross-validated ordinary least squares model using these five features explains 20% of the variation in composite wealth index in the sample of 856 survey respondents (Table S1A).

The second baseline uses the single feature, generated by the DFA, which is empirically determined to be most strongly correlated with the wealth composite index in the sample of 856 survey respondents. More precisely, for each of the 5,088 features generated by the DFA, we use 5-fold cross-validation to divide the set of 856 respondents into five different training and testing sets (with an 80%-20% split). For each training set, we fit a linear regression of the response variable on the single feature, and compute the R^2 for the corresponding test set; we average the test R^2 across these five folds, and select the feature that has the highest average test R^2. The single best predictor, which has an average test R^2 of 0.39, indicates, for an individual i, the weighted average of all of i's first-degree neighbors "day of week entropy" of outgoing SMS volume, where the weights are determined by the frequency of interaction between i and the neighbor. Roughly, this is an indication of the extent to which there is a high degree of predictability in the days of the week on which i's friends and family tend to send text messages.

While this single feature performs surprisingly well, we do not expect that this could have been foreseen in advance or that it would be as informative in other contexts (see SM Section VB).

1.3.2.2 Deterministic finite automaton (DFA)

Our deterministic finite automaton takes as input a list of call detail records (CDRs), where each element in the list is a transaction record containing a tuple of fields (date, time, userID, and so forth). From this initial state, the data transitions to subsequent states, where each transition defines a legal operation that transforms the data input to the state into a different dataset output from the state. The final output from the DFA is a single numerical value, which is equivalent to a single behavioral metric, or "feature." Thus, any feature used in our analysis can be generated by a complete traversal of the automata. The DFA used for feature generation is shown in SM Figure 1, and is defined by:

- A set of states $Q = \{q_0, q_1, ..., q_{23}\}$
- The start state q_0
- The accepting state q_3
- The alphabet $\Sigma = \{CDRs, Fields, Field, Value\}$
- The transition function $\delta : Q \times \Sigma \Rightarrow Q$

SM Figure 1 depicts the transition function. Note that we assume that any element of $Q \times \Sigma$ not pictured results in a transition from state q_i back to q_i. For example, $\delta(q_{13}, a') = q_3$ while $\delta(q_{13}, f) = q_{13}$ since it is not legal. The transition function is specified as:

- $f(\cdot) : Q \times CDRs \Rightarrow CDRs$: "Filter" operations on a set of CDRs that select a subset of the CDR tuples.
  Example: Filter all rows that are not incoming calls.
  Legal transitions: calls over 60 seconds; calls made during the working week (Monday - Friday, 9am -5pm); calls not made during the working week; incoming activity; outgoing activity; international activity; text messages (SMS).
- $m(\cdot), m'(\cdot), m''(\cdot) : Q \times CDRs \Rightarrow list(CDRs)$: "Group By" operations that transform a dataset of type D into a map from the attribute to subsets of D, where subsets are defined by the attribute, which may be the identity of the subscriber ("ego"), the identity of the subscriber's contacts ("alter"), or a time period attribute.
  Example: Group all CDRs by ego and week of the year.
  Legal transitions: group by ego; group by alter; group by week.
- $s(\cdot) : Q \times CDRs \Rightarrow Fields$: "Select" operations that transform a set of rows into a set of values. This can be any operation on a set of rows that maps each row to a number.
  Example: Select a single field from a row (such as "duration of call")
  Legal transitions: select duration of event; select geo-coordinates of ego at time of event; select day of week; select hour of day;
- $a(\cdot), a'(\cdot), a''(\cdot) : Q \times Field \Rightarrow Value$: "Aggregate" operations that aggregate a set of numbers into a single number. These convert a mapping from some attribute to a set of values to a mapping from an attribute to a single number. Example a: Compute mean of a list of numbers; compute radius of gyration of a set of geocoordinates. Example a': Computes aggregation over first-degree network properties, e.g., average PageRank of first degree neighbors of an individual.
  Example a": Computes aggregation over time, e.g., trend over time in calls per week.
  Legal transitions: mean; maximum; minimum; standard deviation; sum; radius of gyration; count of unique values.
- $r(\cdot), r''(\cdot) : Q \times CDRs \Rightarrow CDRs$: "Reduce" operations that groups a multi-level mapping by one level.
  Example: Subsets grouped by user-time are aggregated into subsets grouped by user.
  Legal transitions: Values mapping to (ego, time period) tuples are grouped in sets identified by the ego and the mapping from egos to their sets are returned.

Not depicted in the DFA, but also included in feature engineering, are simple transformations (log and quadratic) of the DFA traversals. We also experimented with including features that do not fall neatly into this framework, such as PageRank, but in practice this has little effect on the results.

As an example, the following traversal of the DFA will produce a feature that indicates the standard deviation of the weekly average call duration during working hours.

- Start state q_0
- $\delta(q_0, f_{workday} \in f) = q_0$: filters out calls made on weekends or outside 9am-5pm
- $\delta(q_0, m) = q_1$: groups all calls by subscriber ("ego")
- $\delta(q_1, m'') = q_{21}$: groups all calls by week, calls are now grouped by subscriber-week

- $\delta(q_{21}, s_{duration} \in s) = q_{22}$: converts groups of calls to groups of call-durations
- $\delta(q_{22}, a_{mean} \in a) = q_{23}$: computes the mean of each group of call-durations; at this point, each subscriber is represented by a set of weekly averages
- $\delta(q_{23}, a''_{sd} \in a) = q_3$: computes the standard deviation of weekly averages.

1.3.2.3 Feature categorization

While the DFA is effective in constructing a large number of features from a relatively parsimonious grammar, the quantity of resultant features complicates interpretation. This is a clear disadvantage of the DFA relative to more parsimonious models based on intuitive features. As noted above, however, our primary goal is predictive accuracy, not model interpretation. Nonetheless, to inform the subsequent analysis, we label each feature with a "type" by grouping the features according to approximate function. Alternative partitionings of the feature space are equally plausible, but the partition we choose roughly follows the broad classes of features discussed in related literature (40).

- SMS activity (ego): Metrics reflecting SMS-based activity of the subscriber, including
- volume, variance, variation over time, etc.
- SMS activity (alter): Metrics reflecting SMS-based activity of the subscriber's first-degree network (FDN).
- Call activity (ego): Metrics of call-based activity of the subscriber.
- Call activity (alter): Metrics of call activity of the subscriber's FDN.
- International communications (ego): International call activity of the subscriber.
- International communications (alter): International call of the subscriber's FDN.
- Movement (ego): Information on the pattern of locations visited by the subscriber.
- Movement (alter): Information on the locations visited by the subscriber's FDN.
- Local network structure (ego): Simple properties describing the subscriber's position within his or her FDN.
- Local network structure (alter): properties of the subscriber's FDN's social networks.
- Global network structure: Structural properties describer the subscriber's position within the entire graph, such as PageRank and clustering coefficients.

1.3.3 Model fitting and out-of-sample prediction

1.3.3.1 Supervised Learning

From the several thousand behavioral metrics constructed by the DFA, we used supervised learning techniques to identify a smaller subset of features that are the best joint predictors of the response variable, using the sample of 856 survey respondents to train the model. Specifically, we use elastic net regularization (41) to penalize model complexity and reduce the likelihood that the model is "overfit" on the small number of training instances. For each possible model parameter β_j , the elastic net imposes a penalty equal to

$$\lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)\left|\beta_j\right| ,$$

This penalty linearly combines a lasso (L_1) penalty for variable selection with variable shrinkage as in ridge regression (L_2), where higher values of λ produce more parsimonious models. As noted in SM Section IIA, we compare the elastic net model to models using lasso and ridge regression separately, and find only modest differences in performance from the elastic net. Similar results obtain when using nonlinear tree-based ensemble regressors to predict the continuous-valued composite wealth index, and random forest classifiers to predict asset ownership and housing characteristics (24) – these results are presented in Table S1A.

For each model, we use cross-validation to help ensure that the model will generalize beyond the small sample upon which it is fit. Specifically, we use 5-fold cross-validation to select model parameters that maximize average R^2 on the held-out test data across 5-folds.[3] Each fold is selected with a weighted bootstrap, where the weights are determined as described in SM Section SIA to help ensure that the model is representative of the total population of mobile phone subscribers (43).[4]

SM Figure 4A illustrates how model performance depends on the choice of the regularization parameter λ. For large values of λ, the model selects a very small number of features, and the average performance on both the training and testing data is quite poor. (For extreme values of λ, performance is also considerably worse than the unregularized single-predictor model). As λ is decreased, a larger number of features enter the model, and performance on both the training and testing data increases until the optimal model selects 101 features. Additional increases in λ yield improved performance on the training data, but performance on the test data degrades as the model is overfit to the training instances.

1.3.3.2 Improving model performance

While model performance appears to be only marginally affected by the choice of the learning algorithm, we find that predictive performance is significantly impacted by the relatively small number of independent observations available. This issue is illustrated in SM Figure 2, where we show the performance that would have been achieved if we had trained on a smaller number of independent observations. These hypothetical scenarios are determined by drawing a random subset of m observations from the full set of 856 respondents, then re-training the model as if only those observations were available. We interpret the monotonic increase with sample size, and the continued positive slope at the maximum where m=856, as evidence that further performance gains could be achieved by expanding the sample of phone survey respondents. In our case, the size of the survey sample was determined by a financial constraint; increasing the sample size would likely produce noticeable improvements in predictive accuracy.

1.3.3.3 Interpreting supervised learning models

The original set of 5,088 features contains several behavioral metrics that are unconditionally correlated with the socioeconomic data collected in phone surveys, and a large number of features that are uncorrelated (SM Figure 3A). SM Figure 3B shows the ten features which are most highly (unconditionally) correlated with the wealth composite index; many of these features are correlated with each other, and have to do with the temporal entropy of the

[3] Cross-validation is a common method for model selection and validation. The data is first randomly divided into K random subsets, called "folds". Then, each fold is removed from the dataset, one at a time; the model is fit on the remaining data, and evaluated on the held-out fold. This process is repeated for each fold, and the model performance is reported as the average across all of the held-out folds (42). In our case, we repeat this entire process for all possible values of λ and α, then select the model that performs best (across held-out folds).
[4] In practice, the weighted bootstrap sample selection has little impact on results relative to a naïve selection process that evenly divides the sample into five non-overlapping sets of training (80%) and testing (20%) instances.

communications behavior of an individual's first-degree network. SM Figure 3C uses the feature partitioning described in SM Section IIC to show the distribution of the 5,088 separate R^2 values by feature type, separately for the task of predicting the composite wealth index and for the task of predicting whether the respondent owns a motorcycle. While the two sets of distributions are visually similar, the correlations are generally higher for wealth than for motorcycle ownership. Comparing the relative importance of different classes, it appears that features related to the movement patterns of an individual's social network are predictive of motorcycle ownership, whereas factors related to text messaging are most useful in predicting wealth. While it is not difficult to rationalize these observed trends ex post (for instance, it may be that text messaging is related to literacy, which is in turn correlated with wealth), we are wary of interpreting these correlations too literally.

The supervised learner described earlier optimizes the joint predictive ability of a set of features, where regularization and other methods for model selection are used to eliminate features that are not predictive or redundant. SM Figure 4A shows how model performance depends on the number of features in the model, which is in turn determined by the regularization parameter λ. SM Figure 4B illustrates how the set of features in the final model also changes as a function of the regularization parameter. When model complexity is highly penalized, few features are selected and they are initially all from the class of features that are unconditionally correlated with the response variable (in this case, the features related to the temporal entropy of the communications behavior of an individual's first-degree network.). As the penalty is reduced and more features enter the model, a more diverse set of features is selected. The optimal model includes features from a large number of different feature groups.

1.3.4 Validation with independent sources of "ground truth" data

1.3.4.1 Assignment of individual mobile phone subscribers to geographic location

Each mobile phone transaction in the call detail records is tagged with a geographic identifier corresponding to the mobile phone cell tower nearest the subscriber at the time of the transaction. Combined with a separate database containing the GPS coordinates of each cell tower, this allows us to approximately locate each individual at the time when the transaction occurs. The set of locations associated at which an individual is observed can in turn be used to infer that individuals approximate "home" location (17, 44). The primary method we employ to locate an individual is to calculate the modal evening tower, defined as the single tower which the subscriber is observed to use most frequently between the hours of 8pm and 6am.[5] In developing the high-resolution visualizations (Figure 2), we additionally compute each subscriber's "center of gravity", defined as the weighted Euclidean centroid of all locations observed by the subscriber (17).[6] In practice, our results are not sensitive to the exact manner in

[5] More precisely, we compute, for each hour of the day, the most frequently used tower in that hour (the "modal tower-hour"). We then compute, for each evening, the most frequently observed modal tower-hour (the "modal tower-evening"). Finally, we compute the most frequently observed modal tower-evening across all evenings in the dataset, and use that as the subscriber's "home" location. This approach is designed to capture the location at which the subscriber spends the majority of his or her hours, rather than the location from which a majority of calls are made.

[6] Specifically, if an individual i with an modal evening tower mti is observed at N_i (non-unique) locations ($r_{i1}, \ldots, r_{iNi}$), we define the center of gravity as $(1/N_i)\sum r_{it} COG_i * \mathbf{1}(r_{it} - mt_i < k)$, where the indicator function restricts the

which locations are computed: choosing "home" location by looking at all towers used at all hours of the day, for instance, yields nearly identical results. At the finest level of spatial granularity presented (Figure 2D), we show average locations of groups of 5-15 subscribers, where groups are determined using k-means clustering on the subscribers' centers of gravity, in order to add a layer of anonymity to the high-resolution maps.

1.3.4.2 Geographic aggregation: matching cell tower locations to DHS locations

When comparing the predicted wealth composite measures derived from the call records to the "ground truth" data found in the Demographic and Health Surveys, we require a comparable method of geographically aggregating data from the two sources. Our analysis uses two such levels of aggregation: district-level aggregation and "cluster"-level aggregation.

When aggregating estimates at the district level, each mobile phone subscriber is assigned to a modal evening tower as described in Section IVA above. As shown in SM Figure 5, the set of unique tower locations form a voronoi division of Rwanda. We compute the average composite wealth of each voronoi division Y_v^{CDR} as the mean of the composite wealth values of all subscribers i whose modal evening tower is v, i.e. $Y_v^{CDR} = \frac{1}{N_v} \sum_{i \in v} \hat{y}_i$, where $\hat{y}_i$ is the predicted wealth of subscriber i and N_v is the number of subscribers in v. The average predicted composite wealth of district d is then computed as the weighted average of all towers falling within the district borders, $Y_d^{CDR} = \frac{1}{\Sigma w_{dv}} \Sigma_v w_{dv} * Y_v^{CDR}$ Y, where w_{dv} indicates the proportion of the tower's voronoi cell that lies within the district boundary (SM Figure 5, inset).

Our validation estimates compare these Y_d^{CDR}, the estimates of district wealth based on mobile phone data, to the "true" wealth of the district, Y_d^{DHS}, which is computed from the DHS data as $Y_d^{DHS} = \frac{1}{\Sigma_{j \in d} w_j} \Sigma_{j \in d} w_j * y_j$, or simply the weighted average of all households j in district d, where w_j is the sampling weight given to j in the DHS. Y_d^{DHS} is computed separately for all households in a district, and for just the subset of household who own a mobile phone, which we later refer to as Y_d^{DHS-MP}. Correlations are weighted by population expansion factors to fit the regression line more closely to regions with large populations (4).

1.3.4.3 Cluster-level validation

We follow an analogous procedure when aggregating wealth estimates at the cluster level. Clusters are meant to approximate villages in Rwanda, and are defined in the data by the GPS locations of cluster centers collected during DHS survey collection (31). SM Figure 5 provides an example of how the aggregated composite wealth index is computed for a single cluster. The red dot indicates the cluster's center, and the pink shaded area represents the voronoi cell covered by the cluster. The blue dots indicate the locations of mobile phone towers, and the blue lines indicate the implied voronoi division, where dots are only shown for towers where the tower's voronoi cell overlaps with the cluster's voronoi cell. The numbers indicate wdv, the proportion of the cluster's cell covered by the tower's cell. Thus, the CDR-imputed wealth value for the pink cluster will be the weighted sum of the average composite wealth predictions of each

---

weighted average to include towers within k kilometers of mti, to remove the influence of outliers (such as a weekend trip or short vacation). In the figures that rely on the center of gravity, we set k =10, but qualitatively similar results are obtained for a variety of reasonable thresholds (including k=∞)

of the labelled blue cells, where the weight is given by the black number in the cell. As noted in SM Section IC, the cluster centroids are randomly displaced by up to 10km by the DHS administrators. These displacements are intended to protect the identity of individual households, and add considerable measurement error to our ability to match DHS data to mobile phone data. The DHS documentation thus advises against disaggregating geospatial analysis below the district level (31).[7] For this reason, the results we emphasize in the main text that rely on DHS data use district-level aggregation.

These caveats notwithstanding, we compare phone-based estimates of average cluster wealth to DHS averages, for each of the 492 clusters in the 2010 DHS (Figure 3E). In general, the correlation at the cluster level (r = 0.79) is weaker than at the district level (r = 0.92), though for the reasons noted above this is not surprising. The primary advantage of the cluster-level analysis is that it makes it possible to analyze within-district variation, to see whether the phone-based approach picks up on differences between clusters within a district that are observed in the DHS data. SM Figure 6 thus disaggregates the results of Figure 3E by urban and rural regions. The original relationship (r = 0.79) is attenuated, but a correlation is still observed within both urban (r = 0.64) and rural (r = 0.50) districts.

1.3.4.4 Satellite night lights

Recently, a small body of work has used night-time luminosity data collected by satellites to measure economic output and growth (45, 46). A key advantage of satellite data is that it is pervasive and publicly available. SI Figure 6 compares data collected by satellites on the nighttime luminosity in Rwanda with estimates of electrification based on mobile phone data. The night-light imagery, collected by the National Oceanic and Atmospheric Administration, provides a resolution of 15 arc-seconds (equivalent to a 0.74km x 0.74km grid), which is shown for the country of Rwanda (SI Figure 6A) and enlarged for the region surrounding the capital city of Kigali (SI Figure 6B). As can be seen in SI Figure 6A, there is very little variation in luminosity data in poor, rural regions. Indeed, outside of the capital city of Kigali, most of the country of Rwanda appears dark and unlit.

By contrast, the approach based on phone data captures a great deal of variation even in the most rural parts of the country, and allows for the distinction between households that have access to electricity and households that are brightly lit at night (Figure 3). We use the method described in the paper to predict how each of the 1.5 million subscribers would respond to the survey question, "Does your household have electricity?" using methods analogous to those used to predict composite wealth. Each subscriber's center of gravity is used to place the individual in a grid cell, and the average predicted response is computed across all subscribers. These values are then used to construct a map of predicted electrification in the Kigali region (SI Figure 6C). While the two images are visually similar, they are designed to capture slightly different phenomena: the night light imagery is optimized "to observe dim signals such as city lights, gas

---

[7] Excerpted from the DHS documentation (at http://dhsprogram.com/faq.cfm, accessed October 2015): "Can I calculate indicator estimates for areas smaller than the [district]? The survey design for DHS is not conducive for small area estimation. Households and respondents were selected in order to produce representative population estimates at the national and [district] level only. Any sub-[district] estimates are highly unreliable and likely to result in large standard errors. Is it possible to do spatial analysis of DHS at the individual cluster level? No, the sample frame is designed to ensure that the data are representative at the national and district level only."

flares, auroras, wildfires, and reflected moonlight"; the mobile phone-based predictions are constructed to map household electrification. In urban settings like Kigali, we presume these to be strongly correlated, but in more rural regions the distinction is more important.

1.3.5 Generalizability and external validity

The results in Figure 1 illustrate how our method can be used to infer individual characteristics (in our case, phone survey responses) from passively-generated transactional data (mobile phone records), for the population of individuals who generate such data (the population of active mobile phone subscribers). This method, we believe, should generalize to a wide range of contexts where it is possible to supplement large transactional datasets with targeted surveys. SM Section VI provides several examples of possible applications of this method that extend far beyond the population of Rwandan mobile phone owners, which we hope we and other researchers can improve upon in future work.

1.3.5.1 Population inference from a sample of mobile phone subscribers

The model fit on the sample of 856 respondents is then used to generate out-of-sample predictions for the population of 1.5 million mobile phone subscribers in Rwanda. To validate the accuracy of these predictions, we compare the aggregated output of this model to DHS data aggregated at the same geographic level. In performing this validation, we observe two distinct results. First, we find that the average wealth of a district, as predicted by the mobile phone data (Y_d^{CDR}), is strongly correlated (r = 0.917) with the average wealth of mobile-phone owning households in that district (Y_d^{DHS-MP}), as reported in the 2010 DHS.[8] This provides objective validation that our method can reconstruct the distribution of wealth of a population for whom we expect it to be representative, i.e., mobile phone owners. Since our estimate of Y_d^{CDR} was constructed "in a vacuum" and without access to the DHS data, there is no possibility that the relationship is mechanical or that the model was overfit to the DHS target. Second, as shown in Figure 3, we observe an equally strong correlation (r = 0.916) between the phone-based estimates of district wealth (Y_d^{CDR}) and the average wealth of all households in the district (Y_d^{DHS}). This result indicates that, at least in Rwanda, our method can approximate the distribution of wealth of the full national population. This is true despite the fact that Y_d^{DHS} is constructed from a sample that is representative of the population of all Rwandans, while Y_d^{CDR} is constructed from a sample that is representative of the population of active mobile phone subscribers. And it is true despite the fact that, as we have shown in prior work (30), these two populations are different: mobile phone subscribers in general are wealthier, better educated, and more likely to be male.

1.3.5.2 Generalizing to other contexts

In other contexts, it is possible that one could accurately reconstruct the wealth of phone owners from phone records (as we do in Figure 1), but not be able to accurately reconstruct the distribution of wealth of the full population from out-of-sample inferences about mobile subscribers (as we do in Figure 3). In the general case, assume the researcher has conducted a targeted survey with a sample of individuals (POP^{survey}), who we assume are a random,

[8] In this DHS, mobile phones are owned by approximately 42% of the sample or 5,315 households.

representative sample of the population of individuals for whom the researcher has transactional data (POP^{data}),[9] who in turn constitute a subset of the full population (POP^{full}). As a broad heuristic, the more representative POP^{data} are of POP^{full}, the more effective we expect this approach will be; if POP^{data} are not representative, then validating estimates against external data on POP^{full}, as we have with Figure 3, is a critical step.

In Rwanda, there are several possible explanations for why we are able to reconstruct the distribution of wealth of POP^{full} from POP^{data} even though we know the latter is not a representative sample of the former. The simplest explanation, however, is the fact that in Rwanda, Y_d^{DHS-MP} is closely correlated with Y_d^{DHS} (r = 0.984). In other words, there exists a strong correlation between the average wealth of region's population and the average wealth of a region's mobile phone-owning population. In situations where the selection process into mobile phone ownership is uniform across regions, this result is likely to generalize.

More broadly, as mobile phones are quickly adopted in developing countries (11), it may become more tenable to predict wealth and poverty from mobile phone data in a broad range of geographic contexts. In general, however, POP^{data} may not be representative of POP^{full}, and the ability to infer properties of POP^{full} from POP^{data} will depend heavily on the context of the application. In Rwanda, for instance, our analysis was facilitated by the unusual concentration of the mobile phone market. In more fragmented markets, the approach might need to be adapted if there is systematic selection of subscribers into mobile phone network providers, unless the researcher can obtain data from all relevant operators.[10] Similarly, the near-ubiquitous coverage and high density of cellular towers in Rwanda (SM Figure 5) made it possible to include remote regions in POP^{survey}, which in turn allowed us to construct high-resolution estimates for the entire country (Figure 2).

Related, our analysis focuses on predicting the composite wealth of a subscriber ($\hat{y}_i$), where the composite wealth is defined the first principal component of the assets and characteristics of the household. This target variable was well-suited to the Rwandan context, where many phones are shared within households (30), income is typically pooled among household members, and the majority of households rely on subsistence agriculture. In other contexts, where phone use is more individual and it is more common to earn a fixed wage, individual income may be a more natural target prediction variable. However, one limitation of the approach we propose is that it is designed to model response variables that can be elicited through short, structured phone interviews. Thus, it would be difficult to use this method to predict consumption or expenditures, which typically require extensive survey modules, or more sensitive topics that respondents do not feel comfortable discussing over the phone.

Other idiosyncrasies of the Rwandan context, such as the dominance of prepaid accounts and the per-second billing structure, likely impacted the set of features engineered and selected through supervised learning. A fragmented market would also affect the model fit on POP^{survey} , as a single operator's call detail records would only capture partial information for a competitor's subscribers. However, we do not expect that such idiosyncrasies would necessarily weaken one's ability to train a model on POP^{survey}, or imply non-representativeness of POP^{data}. In other words,

---

[9] Our efforts to ensure to draw a sample for POPsurvey that was representative of POPdata are described in SM Section 1A.

[10] Here, an intriguing possibility is governments would require, or other institutions would provide incentives, to operators to make data available for humanitarian use (47).

while the fitted model would change, the process for fitting the model would remain the same, and any changes in goodness of fit are hard to predict ex ante.

1.3.6 Applications and extensions

The focus of this paper has been on predicting poverty and wealth from mobile phone data. However, with minimal changes, an analogous approach could be used to predict a much broader set of characteristics (not just wealth and poverty) by supplementing other large datasets (not just mobile phone records) with other targeted data collection (not just phone surveys). We conclude with a discussion of several ways in which the methods presented in this paper could be further extended.

1.3.6.1 Interim national statistics

One compelling use case for the phone-based predictions of poverty and wealth is as a source of interim national statistics. The thought experiment we imagine is a policymaker who needs to make a decision that requires knowledge of the distribution of wealth. If the policymaker does not have the resources to collect original data, in many countries she would likely rely on data from the most recent nationally-representative survey. As we have noted in the main text, in many developing countries, such data is woefully out of date (3).

Rwanda, in this sense, is unrepresentative of much of sub-Saharan Africa, as multiple nationally-representative surveys have been conducted in Rwanda in recent years. Even so, if our policymaker were in Rwanda in 2010, it is likely that she would use the results of the 2007 DHS, as the results from the 2010 DHS were not made public until mid-2011. As can be seen in SM Figure 8, however, the correlation between estimates of wealth based on mobile phone data and 2010 DHS data ($r = 0.91$) is in fact greater than the correlation between the two successive rounds of DHS data ($r = 0.84$). Thus, if she were to use the 2007 DHS data to identify the districts with below-average wealth, as defined by the first principal component of 2007 DHS assets, she would correctly identify 14 of the 20 districts (70%) which had below-average wealth in 2010, defined by the first principal component of 2010 DHS assets. By contrast, if she were to use the estimates of district wealth compute from the call records, she would correctly identify 17 of the 20 districts (85%). In countries where longer lags exist between successive survey waves, these differences could be quite meaningful.

1.3.6.2 Targeting individuals

The method we describe makes it possible to predict the characteristics of millions of individual mobile phone subscribers. This creates obvious opportunities for profit, if firms wish to target advertising or promotional content to specific demographics. It may also facilitate new methods for targeting target resources to individuals with the greatest need, or for providing information to individuals likely to be at risk. As currently developed, the method focuses on predicting a composite asset index, but in principle a similar approach could be used to estimate consumption as in a proxy means test (48). Relative to the more common asset-based proxy means test, a method based on phone (or other digital transactions) data has certain advantages: it could be targeted to individuals rather than to households; the observed characteristics, derived from call data, can be observed with little marginal cost once the fixed cost of data access is paid; the highly nonparametric process for fitting the target variable to observed metrics could allow for more accurate targeting; and the allocation rule could be made difficult to game.

Yet any implementation of such a system will also face significant obstacles. Many individuals, and particularly the most vulnerable, still do not generate a digital transaction log, and may remain "off the grid" for the foreseeable future. Even if the goal were to only reach mobile phone owners, there would be significant barriers to obtaining the necessary data on phone use, particularly in markets with multiple operators. Finally, as we discuss below, it is likely that the function mapping phone use to the target variable will change over time, either through natural shifts in patterns of device use or through deliberate actions of individuals who wish to alter their behavior to become eligible for benefits. One can imagine possible solutions to these challenges – for example by distributing phones to potential beneficiaries, government-mandated data sharing regulations, or frequent model rebasing – but the path forward is not trivial.

1.3.6.3 Measuring changes over time, and impact evaluation

Perhaps most compelling, and also most speculative, is the possibility that related methods could be used to detect changes over time in the social, economic, or mental state of an individual or small region. A large body of work indicates that events in the real world have unique fingerprints in transactional data (6, 49, 50), and it is easy to imagine that a sudden period of hunger or a bout of depression would be manifest in the phone records of the affected. If a derivative approach could be used to reliably estimate changes in welfare over time, it would enable new approaches to impact evaluation and program monitoring, among other applications.

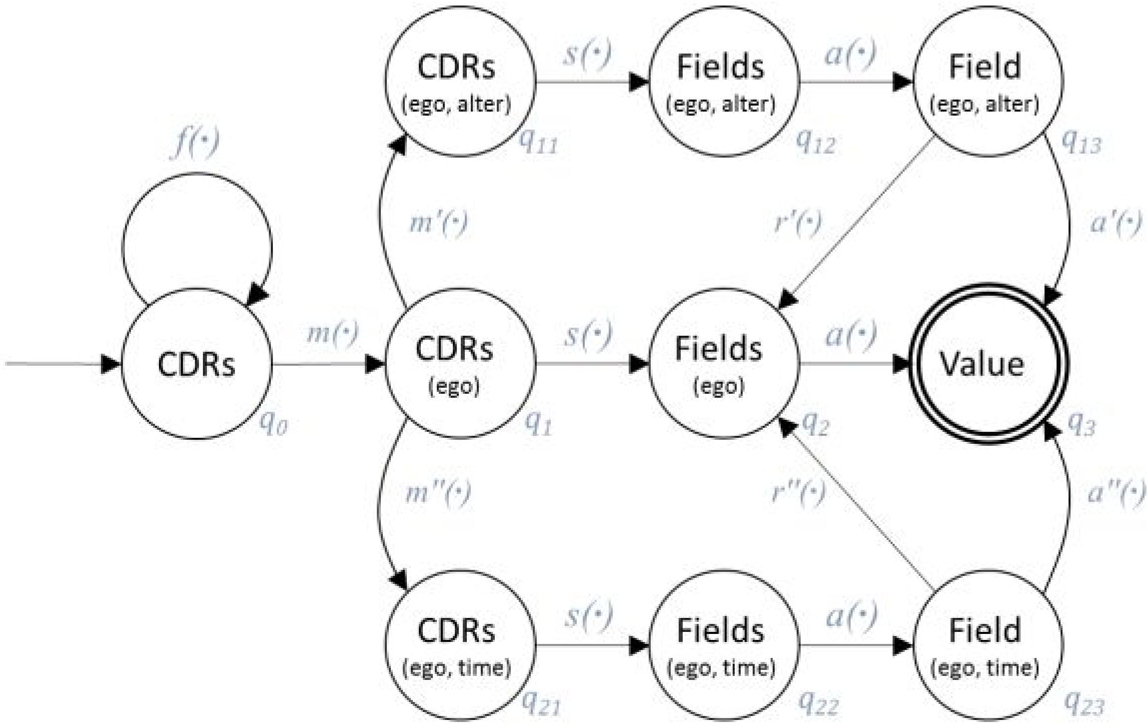
As we have stated repeatedly, however, we do not assume that a model trained on a specific population at a specific point in time could be used to draw inferences about a different population or a different time period. Rather, we expect that the true mapping from digital data to welfare outcomes is context-dependent, and that the model estimated in one time period would deteriorate as time passes from the moment at which it is fit (51). An interesting avenue to pursue here would be to periodically rebase the model by conducting additional surveys to update the model parameters, possibly using online machine learning methods to determine when new surveys are needed and with which populations.

1.3.7 Tables and Figures

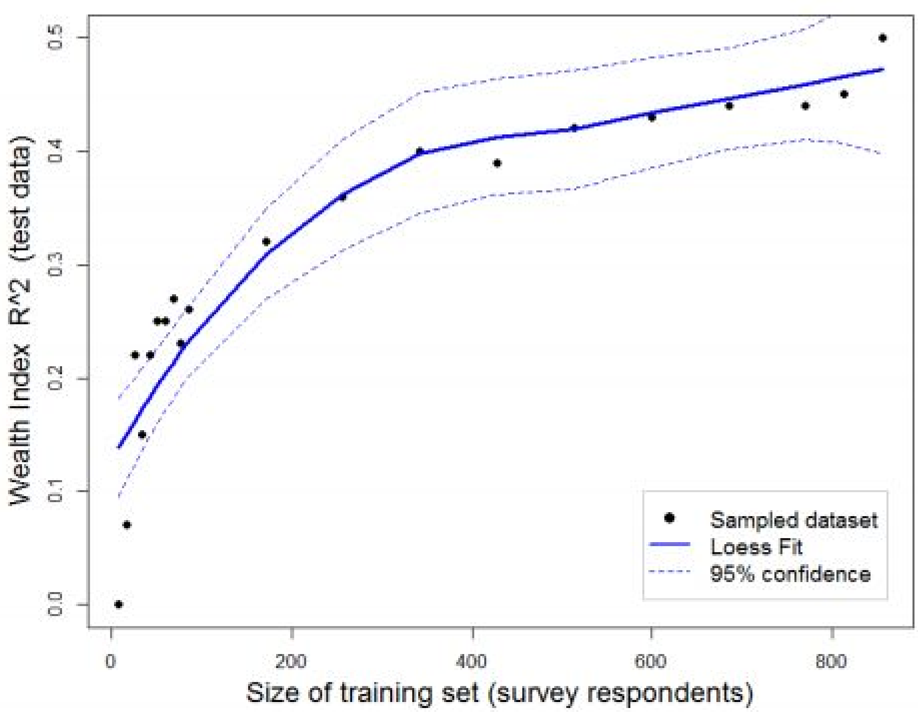
A	Elastic Net		Random Forest	
	r	R^2	r	R^2
Optimal DFA-based model	0.68	0.46	0.63	0.40
"Intuitive" 5-feature model	0.44	0.20	0.37	0.14
Single-feature model	0.61	0.38	0.46	0.22

B	Accuracy	AUC	F score	Baseline
Owns a refrigerator	0.75	0.88	0.40	0.11
Household has electricity	0.72	0.85	0.74	0.60
Owns a television	0.73	0.84	0.72	0.49
Owns a bicycle	0.64	0.68	0.47	0.30
Owns a motorcycle/scooter	0.72	0.67	0.20	0.11
Owns a radio	0.92	0.50	0.96	0.96

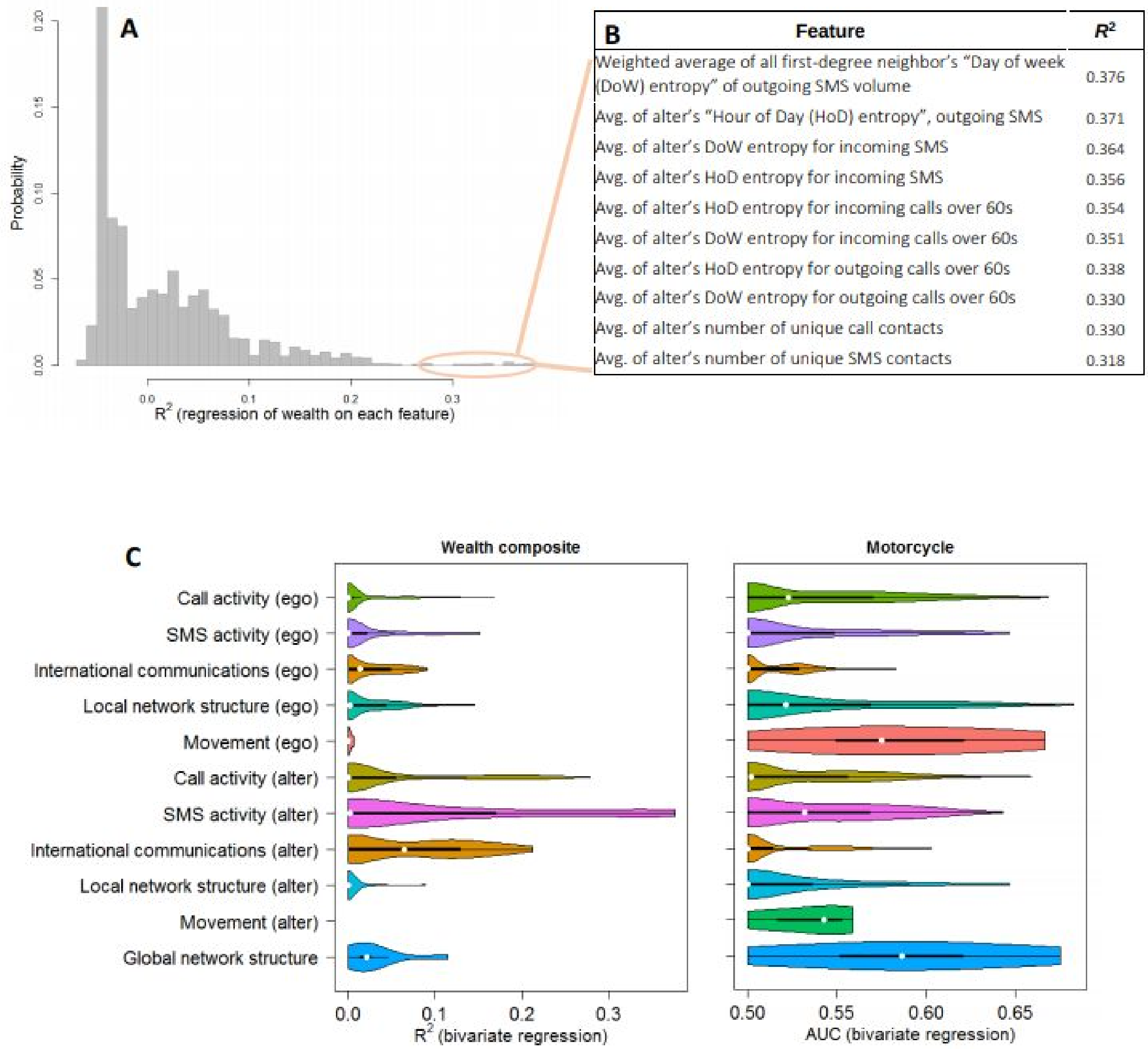
Table S1. Cross-validated performance of predictive models. The models are trained using 5-fold cross-validation on the set of 856 survey respondents. (A) Measures of goodness of fit (correlation coefficient and R 2) for two optimized models: the elastic net which selects 101 features, and a random forest regressor. For comparison, we show performance measures trained on set of five features commonly cited in the literature (total call volume, total SMS volume, total international call volume, radius of gyration, degree centrality); and for a model with the single most predictive feature (the weighted average of all first-degree neighbor's "Day of week (DoW) entropy" of outgoing SMS volume). (B) Performance measures and a naïve baseline for predicting binary survey responses. Accuracy indicates the fraction of correct predictions from regularized logistic regression; Area under curve (AUC) indicates the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, which helps account for the fact that some assets are quite common while others are quite uncommon; the F score provides a performance measure that balances precision and recall; the Baseline is the fraction of respondents who report owning the asset.



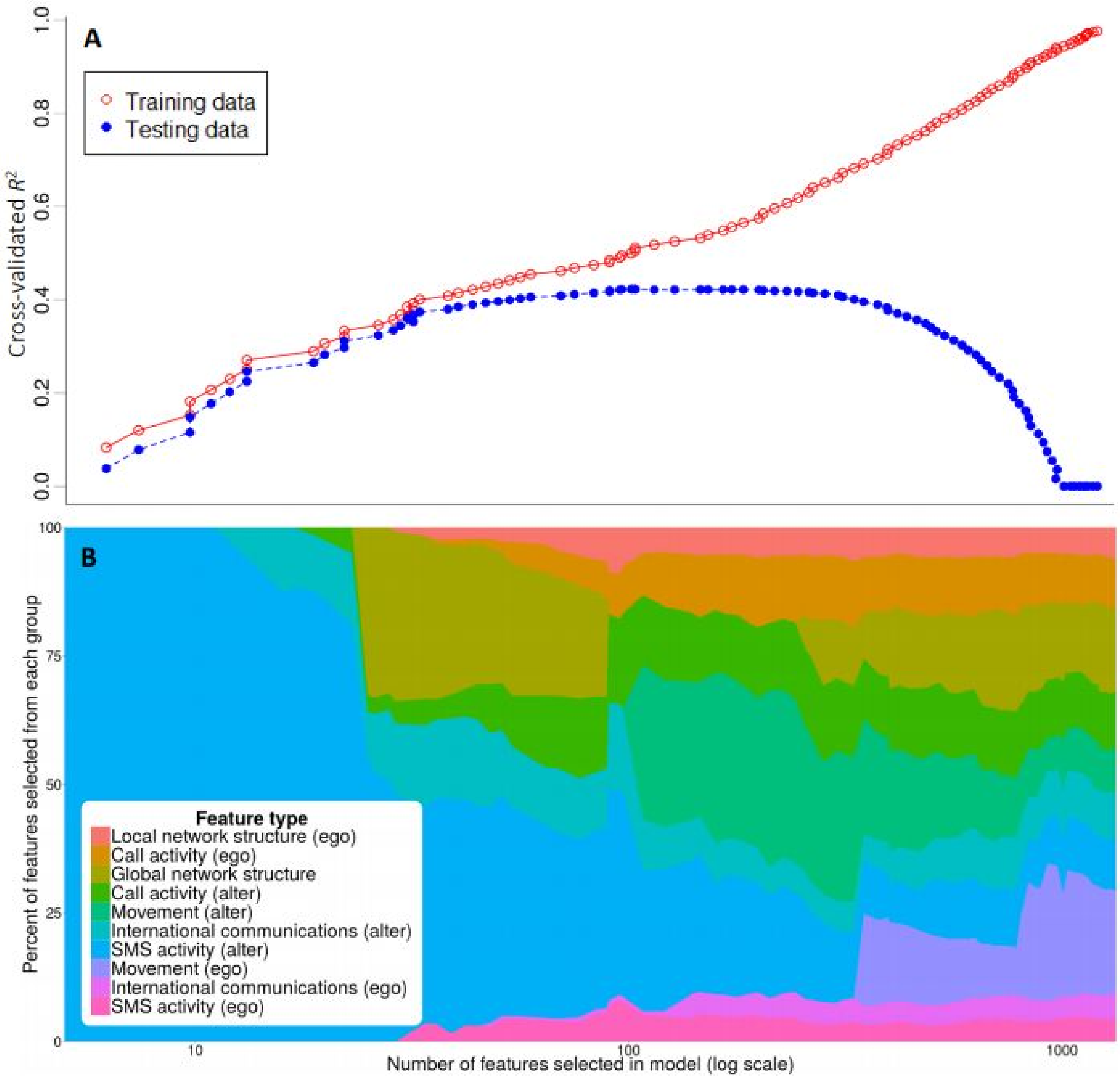
SM Fig. 1. Deterministic finite automaton used for feature engineering. Circles represent states and arrows represent legal transitions, where q0 is the start state and q3 is the accepting (end) state. The final output from the DFA is a single numerical value, which is equivalent to a single behavioral metric, or "feature."



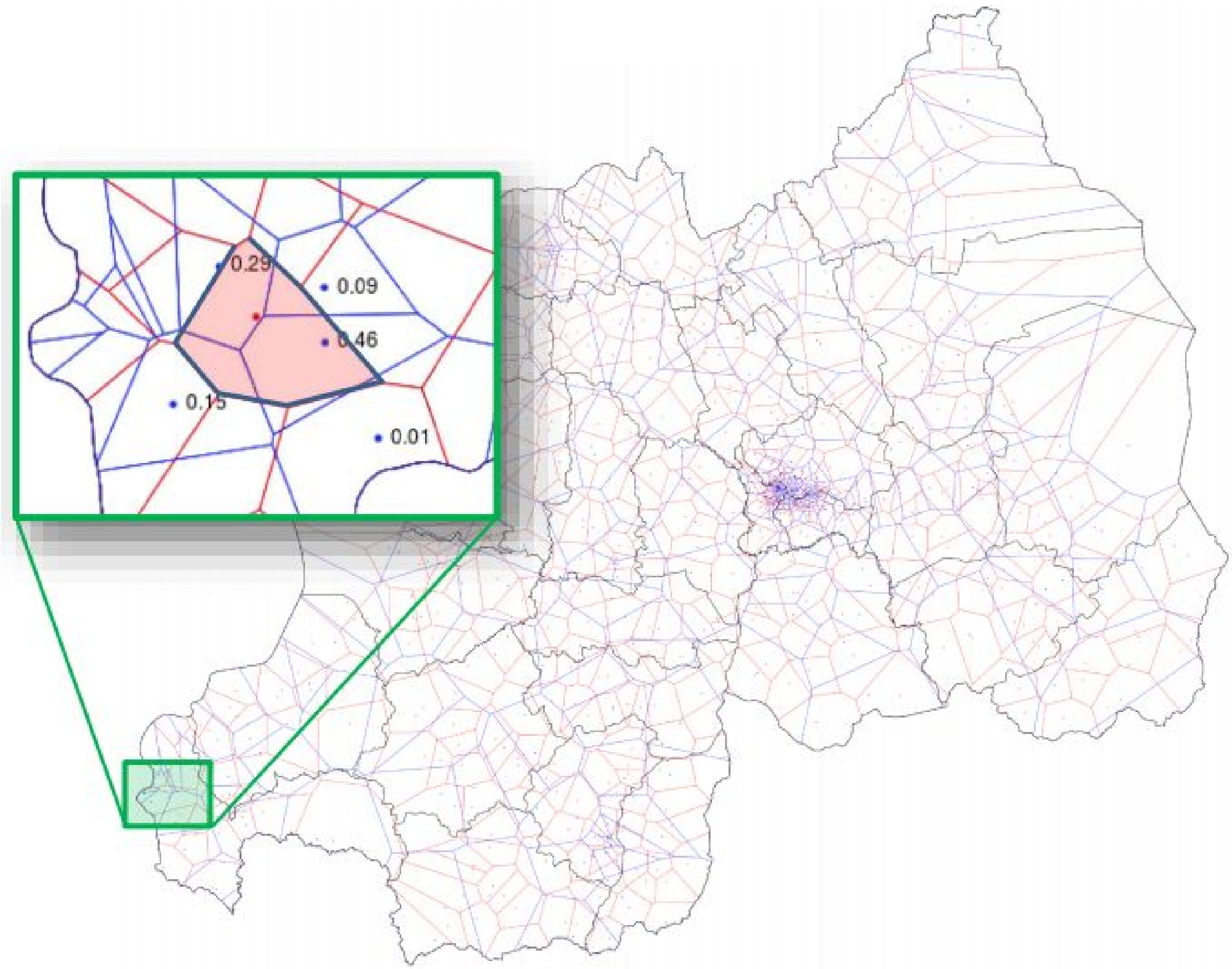
SM Fig. 2. Model performance. As the number of training instances increases, the performance of the model steadily improves. Adding additional respondents would likely enable continued increases in predictive accuracy.



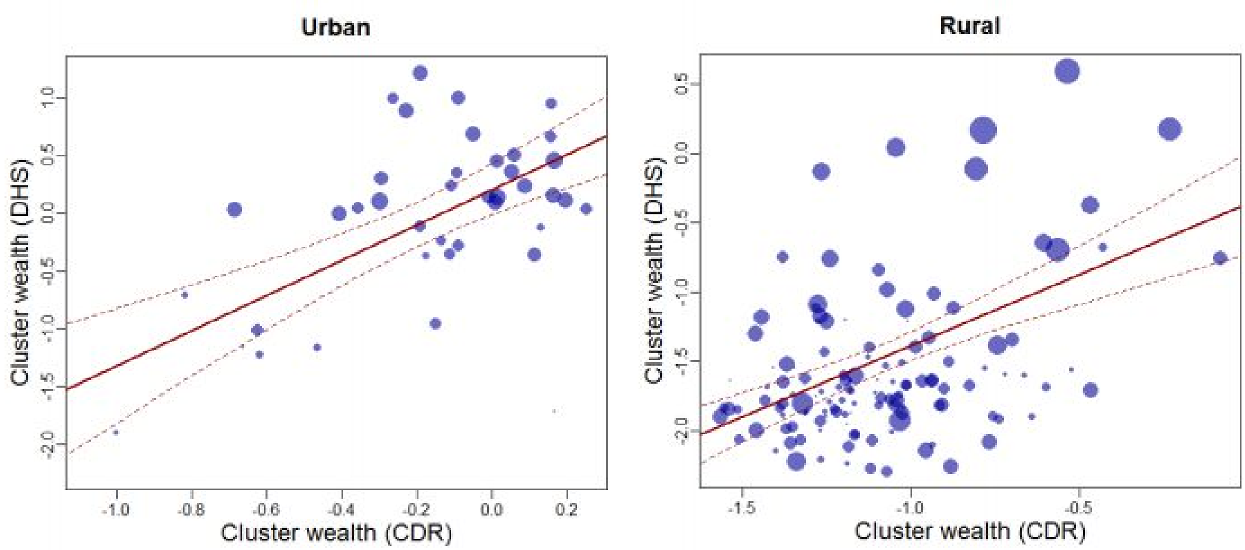
SM Fig. 3. Metrics of phone use that correlate with survey responses. (A) The distribution of R^2 values from 5,088 separate regressions of the wealth composite index on each feature, showing average accuracy on the test set after 5-fold cross validation. (B) Representative list of features strongly correlated with the composite wealth index. (C) Distribution of R^2 values by feature class, for different response variables.



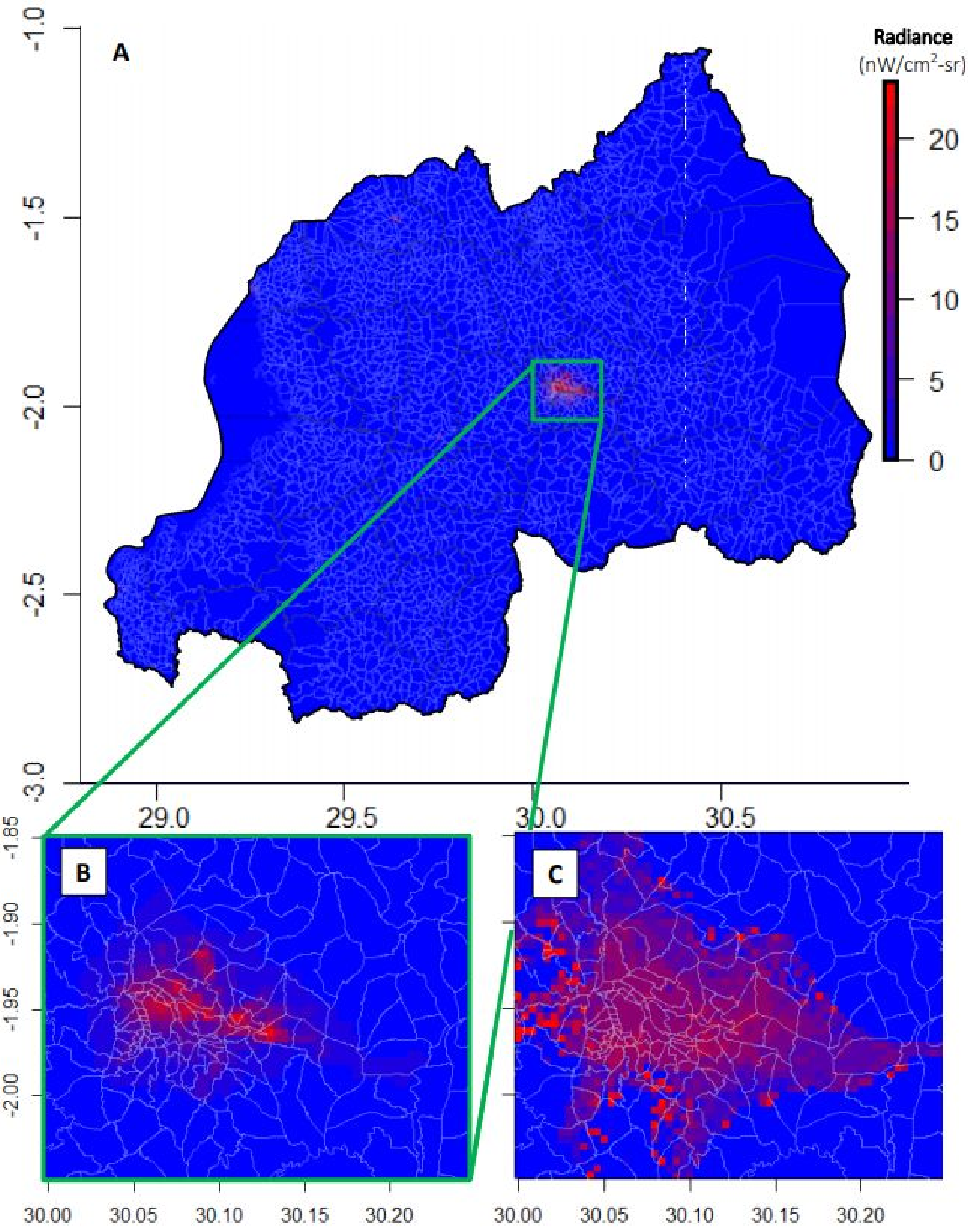
SM Fig. 4. The impact of regularization on model performance and feature selection. (A) Average cross validated performance, showing average R 2 across 5 random training folds and testing folds. Increasing the regularization parameter produces more parsimonious models with fewer features. The optimal regularized model includes 101 features. Including additional features causes the model to overfit on the set of training instances, while excluding features degrades predictive accuracy. (B) Composition of features selected for models of varying complexity. When model complexity is highly penalized, few features are selected and they are all initially from the same class (SMS activity of the ego's first-degree network of "alters"). As the penalty is reduced and more features enter the model, a more diverse set of features is selected.



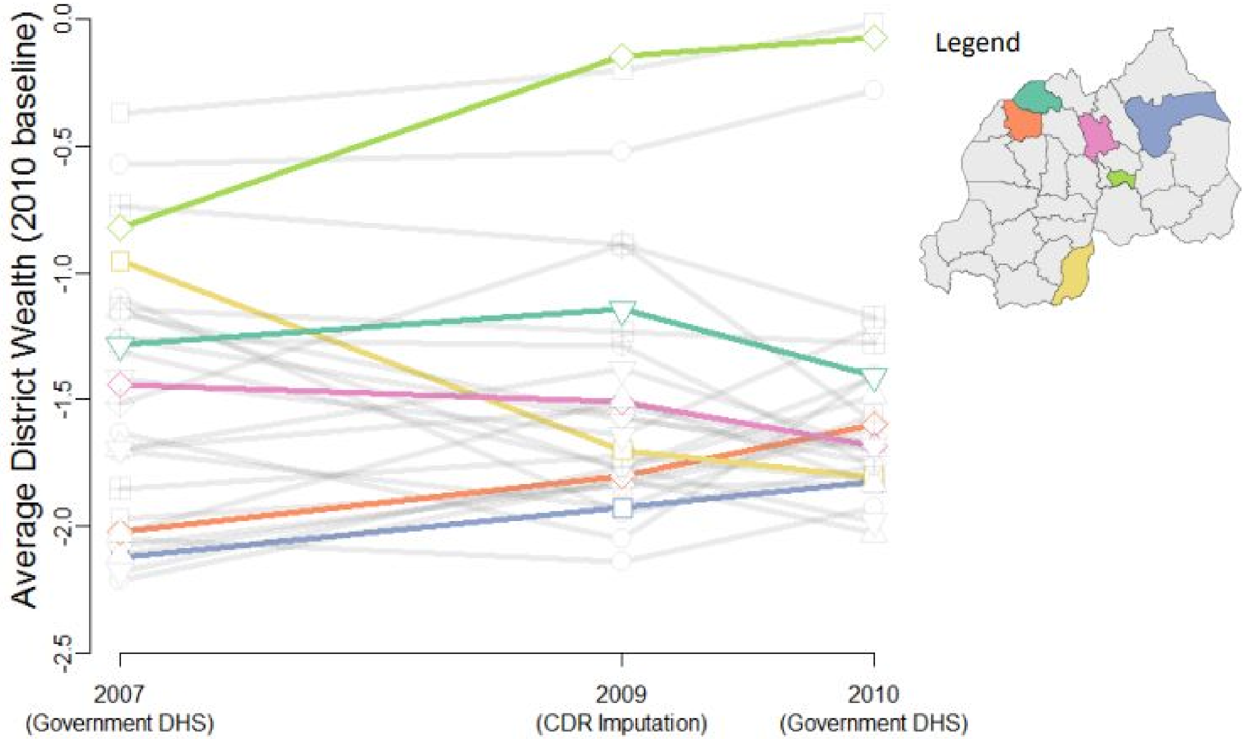
SM Fig. 5. Matching locations of mobile phone subscribers to geographic regions in household survey data. Rwanda is comprised of 30 administrative districts, shown with black borders. In 2009, Rwanda contained roughly 300 unique mobile phone towers, indicated with blue dots. The voronoi tessellation formed by these towers is shown with blue lines. The 2010 DHS sample frame used 492 clusters, the centroids of which are indicated with red dots, and the voronoi tessellation with red lines. The inset figure illustrates how the areas of overlap between the two voronoi divisions are used to compare information aggregated within mobile phone towers to information aggregated within DHS clusters.



SM Fig. 6. Comparison of wealth predictions to government survey data, separately for urban and rural areas. The left figure restricts the analysis to DHS clusters within the urban capital of Kigali; the right panel includes only clusters outside of Kigali. Solid and dashed red lines indicate the regression line and 95% confidence intervals.



SM Fig. 7. Comparison of satellite night-light data to phone-based estimates of electrification. (A) Map of Rwanda showing night-time luminosity, as captured by satellites orbiting the earth (NOAA National Geophysical Data Center). (B) Enlargement of satellite imagery in the region near Kigali, the capital of Rwanda. (C) Predicted household electrification, based on call records, using a model fit on how 856 survey respondents answered the question, "Does your household have electricity?" and projected onto the full population of mobile subscribers.



SM Fig. 8. Phone-based wealth predictions accurately interpolate between traditional rounds of household surveys. Each of Rwanda's 30 districts is represented as a line, where the values in 2007 and 2010 are calculated using household survey data from the Rwandan Demographic and Health Surveys (DHS) of 7,377 and 12,792 households, respectively. The value in 2009 is computed from the mobile phone call detail records (CDR) of roughly 1.5 million subscribers in Rwanda, using a predictive model calibrated on a sample of 856 survey respondents. Every fifth district (ordered by predicted wealth in 2009) is colored to highlight changes over time of six different districts.

1.4 References

1. S. Kuznets, Economic growth and income inequality. Am. Econ. Rev. 45, 1–28 (1955).
2. G. S. Fields, Changes in poverty and inequality in developing countries. World Bank Res. Obs. 4, 167–185 (1989). doi:10.1093/wbro/4.2.167
3. M. Jerven, Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It (Cornell Univ. Press, Ithaca, NY, 2013).
4. C. Elbers, J. O. Lanjouw, P. Lanjouw, Micro-level estimation of poverty and inequality. Econometrica 71, 355–364 (2003). doi:10.1111/1468-0262.00399
5. M. Ghosh, J. N. K. Rao, Small area estimation: An appraisal. Stat. Sci. 9, 55–76 (1994). doi:10.1214/ss/1177010647
6. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van
7. Computational social science. Science 323, 721–723 (2009). Medline doi:10.1126/science.1167742
8. G. King, Ensuring the data-rich future of the social sciences. Science 331, 719–721 (2011). Medline doi:10.1126/science.1197872
9. N. Eagle, M. Macy, R. Claxton, Network diversity and economic development. Science 328, 1029–1031 (2010). Medline doi:10.1126/science.1186605

10. H. Choi, H. Varian, Predicting the present with Google Trends. Econ. Rec. 88, 2–9 (2012). doi:10.1111/j.1475-4932.2012.00809.x
11. W. Wang, D. Rothschild, S. Goel, A. Gelman, Forecasting elections with non-representative polls. Int. J. Forecast. 31, 980–991 (2015). doi:10.1016/j.ijforecast.2014.06.001
12. "The mobile economy 2014" (GSMA Intelligence, 2014); www.gsmamobileeconomy.com/GSMA_ME_Report_2014_R2_WEB.pdf.
13. J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabási, Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A 41, 224015 (2008). doi:10.1088/1751-8113/41/22/224015
14. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A. L. Barabási, Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. U.S.A. 104, 7332–7336 (2007). Medline doi:10.1073/pnas.0610245104
15. G. Palla, A. L. Barabási, T. Vicsek, Quantifying social group evolution. Nature 446, 664–667 (2007). Medline doi:10.1038/nature05670
16. M. C. González, C. A. Hidalgo, A.-L. Barabási, Understanding individual human mobility patterns. Nature 453, 779–782 (2008). Medline doi:10.1038/nature06958
17. X. Lu, E. Wetter, N. Bharti, A. J. Tatem, L. Bengtsson, Approaching the limit of predictability in human mobility. Sci. Rep. 3, 2923 (2013). Medline doi:10.1038/srep02923
18. J. E. Blumenstock, Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. Inf. Technol. Dev. 18, 107–125 (2012). doi:10.1080/02681102.2011.643209
19. V. Frias-Martinez, J. Virseda, in Proceedings of the Fifth International Conference on Information and Communication Technologies and Development (Association for Computing Machinery, New York, 2012), pp. 76–84; http://doi.acm.org/10.1145/2160673.2160684.
20. P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, A. J. Tatem, Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. U.S.A. 111, 15888–15893 (2014). Medline doi:10.1073/pnas.1408439111
21. G. C. Cawley, N. L. C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079–2107 (2010).
22. D. Filmer, L. H. Pritchett, Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. Demography 38, 115–132 (2001). Medline
23. J. Blumenstock, N. Eagle, Divided we call: Disparities in access and use of mobile phones in Rwanda. Inf. Technol. Int. Dev. 8, 1–16 (2012).
24. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320 (2005). doi:10.1111/j.1467-9868.2005.00503.x
25. L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and Regression Trees (Chapman and Hall/CRC Press, New York, ed. 1, 1984).
26. B. Abelson, K. R. Varshney, J. Sun, in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery, New York, 2014), pp. 1563–1572; http://doi.acm.org/10.1145/2623330.2623335.
27. National Institute of Statistics of Rwanda (NISR) [Rwanda], Ministry of Health (MOH) [Rwanda], ICF International, "Rwanda Demographic and Health Survey 2010," DHS Final Reports (publication ID FR259, NISR, MOH, and ICF International, Calverton, MD, 2012).

28. M. Jerven, "Benefits and costs of the data for development targets for the post-2015 development agenda," in Data for Development Assessment Paper (Copenhagen Consensus Center, 2014).
29. Y.-A. de Montjoye, L. Radaelli, V. K. Singh, A. S. Pentland, Unique in the shopping mall: On the reidentifiability of credit card metadata. Science 347, 536–539 (2015). Medline doi:10.1126/science.1256297
30. A. Wesolowski, C. O. Buckee, L. Bengtsson, E. Wetter, X. Lu, A. J. Tatem, Commentary: Containing the ebola outbreak - the potential and challenge of mobile network data. PLOS Curr. 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e (2014). Medline doi:10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e
31. J. Blumenstock, N. Eagle, "Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda," in Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (Association for Computing Machinery, New York, 2010), article no. 6; http://doi.acm.org/10.1145/2369220.2369225.
32. Ministry of Health (MOH) [Rwanda], National Institute of Statistics of Rwanda (NISR), and ICF Macro, "Rwanda DHS, 2007-08 - Rwanda Interim Demographic and Health Survey (English)," DHS Final Reports (publication ID FR215, MOH, NISR, and ICF Macro, Calverton, MD, 2009).
33. A. Deaton, S. Zaidi, Guidelines for Constructing Consumption Aggregates for Welfare Analysis (World Bank Publications, 2002).
34. H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, "A general framework for increasing the robustness of PCA-based correlation clustering algorithms," in Scientific and Statistical Database Management, B. Ludäscher, N. Mamoulis, Eds. (Lecture Notes in Computer Science Series, Springer, Berlin, Heidelberg, 2008), pp. 418–435.
35. D. Hillger, T. Kopp, T. Lee, D. Lindsey, C. Seaman, S. Miller, J. Solbrig, S. Kidder, S. Bachmeier, T. Jasmin, T. Rink, First-light imagery from Suomi NPP VIIRS. Bull. Am. Meteorol. Soc. 94, 1019–1029 (2013). doi:10.1175/BAMS-D-12-00097.1
36. V. Frias-Martinez, J. Virseda, E. Frias-Martinez, "Socio-economic levels and human mobility," paper presented at the QualMeetsQuant Workshop at the 4th International Conference on Information and Communication Technologies and Development, London, 13 to 16 December 2010.
37. J. Blumenstock, Y. Shen, N. Eagle, "A method for estimating the relationship between phone use and wealth," paper presented at the QualMeetsQuant Workshop at the 4th International Conference on Information and Communication Technologies and Development, 13 to 16 December 2010.
38. A. Decuyper et al., http://arxiv.org/abs/1412.2595 (2014).
39. T. Gutierrez, G. Krings, V. D. Blondel, http://arxiv.org/abs/1309.4496 (2013).
40. M. O. Rabin, D. Scott, Finite automata and their decision problems. IBM J. Res. Develop. 3, 114–125 (1959). doi:10.1147/rd.32.0114
41. V. D. Blondel, A. Decuyper, G. Krings, http://arxiv.org/abs/1502.03406 (2015).
42. R. Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288 (1996).
43. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics, Springer, New York, 2009), vol. 2.

44. B. Efron, R. Tibshirani, Improvements on cross-validation: The 632+ bootstrap method. J. Am. Stat. Assoc. 92, 548–560 (1997).
45. R. Ahas, S. Silm, O. Järv, E. Saluveer, M. Tiru, Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. 17, 3–27 (2010). doi:10.1080/10630731003597306
46. J. V. Henderson, A. Storeygard, D. N. Weil, Measuring economic growth from outer space. Am. Econ. Rev. 102, 994–1028 (2012). Medline doi:10.1257/aer.102.2.994
47. X. Chen, W. D. Nordhaus, Using luminosity data as a proxy for economic statistics. Proc. Natl. Acad. Sci. U.S.A. 108, 8589–8594 (2011). Medline doi:10.1073/pnas.1017031108
48. Y.-A. de Montjoye, J. Kendall, C. F. Kerry, "Enabling humanitarian use of mobile phone data," in Issues in Technology Innovation (Brookings Center for Technology Innovation, 2014); http://dspace.mit.edu/handle/1721.1/92821.
49. V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, J. Tobias, Targeting the poor: Evidence from a field experiment in Indonesia. Am. Econ. Rev. 102, 1206–1240 (2012). Medline doi:10.1257/aer.102.4.1206
50. J. P. Bagrow, D. Wang, A.-L. Barabási, Collective response of human populations to large-scale emergencies. PLOS ONE 6, e17680 (2011). Medline doi:10.1371/journal.pone.0017680
51. J. E. Blumenstock, N. Eagle, M. Fafchamps, "Risk sharing and mobile phones: Evidence in the aftermath of natural disasters," Working paper (2014); www.jblumenstock.com/files/papers/jblumenstock_mobilequakes.pdf.
52. D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google Flu: Traps in big data analysis. Science 343, 1203–1205 (2014). Medline doi:10.1126/science.1248506

2 Can SMS-extension increase farmer experimentation? Evidence from Six RCTs in East Africa

with Raissa Fabregas, Michael Kremer, Matthew Lowes, and Giulia Zane[11]

2.1 Abstract

This project provides new evidence on the effects of SMS-based agricultural extension programs on the likelihood that farmers will react to the advice. We present results from six RCTs conducted with farmers in Kenya and Rwanda. All six programs encouraged farmers to experiment with agricultural lime, an input that can reduce soil acidity and increase yields. Four programs also encouraged farmers to experiment with certain types of fertilizers. Programs varied in their design, informational content and target populations. To interpret the findings, we use meta-analytic techniques to combine the results. Our most conservative odds ratio estimates for the effects of the programs on purchases of agricultural lime is 1.19 (95% CI: 1.12,1.26) and 1.02 (95% CI: 0.86,1.22) for purchases of fertilizer. Repeating the same messages had a statistically significant impact on adoption of inputs.

2.2 Introduction

There are many cases in which a principal, such as a government, firm, or NGO, wishes to affect the behavior of agents and can try to do so by delivering information rather than by altering incentives. The spread of mobile phones allows for the possibility of delivering information at scale, in a way that can reach individuals in a timely manner, targeting their specific circumstances and at a very low cost.

Relative to the hundreds of digital public and private sector initiatives that have been deployed in developing and emerging economies, we only have evaluations for a fraction of them (Aker, 2017). A number of programs that rely on short message services (SMS), one of the cheapest ways to deliver information, have been shown to improve educational outcomes (Ksoll et al., 2014; Aker et al., 2012; Cunha et al., 2017), encourage certain health behaviors (Hall et al., 2015; Head et al., 2013) and increase civic engagement (Aker et al., 2017). Yet, other evaluations have failed to find evidence of effects on individual behavior. For instance, so far the track record of m-Agriculture initiatives has been described as mixed, with evidence of some positive and some null results (Aker et al., 2017; Nakasone et al., 2014).

Illiteracy, inability to understand messages, and the cognitive cost of sorting through messages could limit the impacts of information delivery through SMS. Additionally, informational interventions assume that behavior change is not completely constrained by the lack of access to other markets -such as credit or output- and that there are binding informational gaps (Aker et al., 2016). An additional consideration is that the cost to carriers to transmit a marginal message is close to zero. Therefore, even if the impacts of SMS-based programs are

---

[11] Fabregas: UT Austin, LBJ School of Public Affairs, rfabregas@utexas.edu; Kremer: Harvard Economics Department and NBER, mkremer@fas.harvard.edu; Lowes: One Acre Fund, matt.lowes@oneacrefund.org; On: University of California, Berkeley, roberton@gmail.com; Zane: Inter-American Development Bank, giuliaz@iadb.org

modest they can still be cost-effective. One question is whether studies are sufficiently powered to detect the magnitude of effects that would still make these interventions attractive.

In this paper, we present evidence of the impacts of six different SMS-based agricultural extension programs on the use of recommended agricultural inputs. Each program targeted farmers with different profiles and the exact content that was delivered varied, but they were all designed to increase experimentation with certain agricultural inputs, in particular agricultural lime. Experimentation with inputs can be appropriate if there is local variation in unobservable factors -such as soil types- that influence responsiveness to these technologies. We use meta-analytic techniques to systematically summarize the results, increase statistical power and learn about the extent of impact heterogeneity that exists across studies.

Traditional in-person public extension efforts have been criticized for being deficient, expensive and affected by governance issues (Anderson and Feder, 2007, 2004). Therefore, learning more about the role that digital services could play in this sector appears to be promising. We have only a few studies that focus on advisory services and measure changes in farmer behavior. A study of a SMS-extension program offered to sugar cane farmers in Kenya found positive yield impacts in one trial but no effects on a second trial with a different sample (Casaburi et al., 2014). Larochelle et al. (2016) study a SMS-based program for potato farmers in Ecuador. They find that the intervention increased knowledge and self-reported adoption of integrated soil management practices. Cole and Fernando (2016) evaluate an advisory service targeted at cotton farmers in India and find increases in adoption of recommended agricultural inputs for cotton cultivation. However, unlike our the interventions in this paper, that service was delivered through a hotline and not through text-messages. Two other evaluations have used SMS to reach farmers but have mainly focused on the effects of crop price and weather information rather than extension. The service evaluated by Fafchamps and Minten (2012), however, also contained some crop advisory messages. The authors do not find any evidence indicative of farmers changing behavior or practices. A different weather and price information service implemented in Colombia, found improvements on knowledge but not on prices farmers' received or profits (Camacho and Conover, 2010).

We contribute to this list with evidence for six newly evaluated programs. The first experiment was implemented by a Kenyan public agency, the Kenya Agriculture and Livestock Research Organization (KALRO), and it targeted a randomly selected sample of farmers in Western Kenya. These farmers had little existing access to extension services. The second and third programs we evaluate were implemented by Innovations for Poverty Action (IPA), a non-profit research and policy organization, in partnership with Precision Agriculture for Development (PAD), a non-for profit organization that focuses on supporting and delivering agricultural information to smallholder farmers over the phone. One program targeted farmers who appeared in existing (NGO and producer) databases. Another one, recruited customers of agricultural supply dealers. The fourth, fifth, and sixth programs were implemented by One Acre Fund (1AF), an organization that provides credit and agricultural training to farmers. Their SMS-programs aimed to supplement existing high-touch in-person extension activities for clients.

An additional feature of this study is that we have evidence of the effects of these programs on actual measures of behavior change -e.g. acquisition of recommended inputs- instead of only relying on self-reported data. We find that the odds ratio for the effects of being enrolled in the SMS extension programs on lime purchases is 1.19 (95% CI 1.12 to 1.26) in the administrative data and 1.58 (95% CI: 1.35,1.83) with self-reported data. The effect for purchasing fertilizers was 1.02 (95% CI: 0.86,1.22) in the administrative data and 1.31 (95% CI 1.19 to 1.45) in the survey data. However, for fertilizer the significant effect appears to be driven by the inclusion of a single study. For both lime and fertilizers, we cannot reject the null hypothesis of homogenous treatment effects across studies.

Focusing on individual studies, we fail to reject the null that farmers assigned to receive KALRO's SMS program increased their knowledge or use of lime or fertilizers. However, point estimates are positive and statistically significant for the programs implemented by IPA/PAD and 1AF. Farmers who participated in those programs and who received messages were between 1 and 6 percentage points more likely to report following the lime recommendation they were given. In three of the four programs that recommended fertilizer, we do not detect a statistically significant effect. Perhaps because fertilizers are a well-known technology and information about them might not be the primary constraint for adoption.[12] The 1AF program in which we can detect a positive impact for fertilizer use had lower (14%) baseline rates of use.

This paper is organized as follows: Section 2 describes the context that we study, the design of each program and their evaluation design. Section 3 discusses the empirical strategies to estimate effects of each individual program and to summarize the effects across programs using meta-analytic techniques. Section 4 discusses the results from individual programs and the meta-analysis. Section 5 provides a discussion of the results. We present cost-effectiveness estimates in section 6. We conclude in section 7.

2.3 Background on SMS-Based Extension & the Programs

In this section, we describe the agricultural context in which the program took place and describe how each program was implemented and evaluated. Table 1 presents a summary of each program and experiment. To keep the description readable, we only present the main features of each program and key aspects of the evaluation, but further details about each program and their evaluation design are presented in Appendix B.

2.3.1 Maize Farming in East Africa

The projects discussed in this paper targeted maize farmers across Rwanda and western Kenya (see Figure 1) between 2014 and 2017.[13] In these regions, as in many parts of Sub-Saharan Africa, smallholder yields have remained low, partly because of issues of soil degradation and nutrient depletion, soil acidity, and low adoption of productivity-enhancing technologies. It is estimated that over one third of sub-Saharan African soils are acidic (Pauw, 1994). High soil acidity, corresponding to pH levels below 5.5, can dramatically reduce crop yields by limiting nutrient availability to the plants (The et al., 2006; Tisdale et al., 1990; Brady and Weil, 2004).[14] In addition, high soil acidity can cause aluminum and manganese toxicities,

---

[12] In almost all programs, at baseline over 65% of farmers report having used it or order it in the past.

[13] All programs targeted a specific agricultural season. In both countries maize is farmed twice a year. In Kenya, the primary agricultural, the long rain season, takes place from March until August and a secondary agricultural season, the short rains season, takes place from September until December. In Rwanda, the main agricultural season takes place from September to January and the secondary season takes place from March to August.

[14] In particular phosphorus becomes less available to the plant. This can also imply that the use of fertilizers is less efficient on these soils.

which can inhibit plant development (NAAIAP, 2014). The application of agricultural lime to the soil is one of the cheapest and most widely recommended methods to increase soil pH. Experimental plots conducted in Kenya suggest that lime application can increase maize yields by 5-75% (Kisinyo et al., 2015; Gudu et al., 2005; 1AF, 2014).[15]

Currently, several public agencies and NGOs have advocated for the use of lime in these regions. Yet, few farmers are aware that lime is a way to address the problem of soil acidity. For instance, in Kenya, at baseline only 25% of farmers participating in the second IPA/PAD program knew that lime could be used to reduce soil acidity and only 9% of them reported using it in the past. Whether to use lime and the optimal quantity to apply depend on soil chemistry.[16] Few smallholder farmers in this region conduct their own soil chemistry tests as they are not easily accessible and individually costly.[17] To address this issue, the IPA/PAD and the first 1AF programs used area-level soil information rather than field-level to predict soil acidity as a way to provide better targeted recommendations to farmers at a lower cost. There are a number of different efforts seeking to generate and compile localized soil data with the goal of improving management of soils.[18] In a separate project, using soil data, we document that using arealevel means rather than global means reduced the mean squared error of the prediction by 12% for pH (Fabregas et al., 2017b).

In addition to lime, some (but not all) programs also advised on the use of chemical fertilizers. A large body of work suggests that chemical fertilizer can substantially raise agricultural yields (Evenson and Gollin, 2003), and previous research in the region suggests that, certain fertilizers are profitable if used in the right quantity (Duflo et al., 2008). There is a range of different fertilizers available in this area which differ in chemical composition, soil and crop suitability, and price.

2.3.2 KALRO's Program

The Kenya Agriculture and Livestock Research Organization (KALRO) is a public agency with the mandate to promote agricultural research and dissemination in Kenya.[19] In 2014 and 2015, KALRO’s Kakamega office implemented two extension programs aimed at encouraging smallholder farmers to adopt inputs and management practices that could address some of the soil deficiencies in the region. The design of these programs reflected their goal of reaching a large number of farmers at a lower cost than that of individual farm visits. Their first approach was to organize farmer field days (FFDs), one-day events in which a large number of farmers can observe demonstration plots and receive information from extension agents. The second approach consisted of delivering agricultural messages to farmers via SMS. This paper

[15] These estimates reflect results for trials with and without combining lime with other inputs, particularly fertilizers containing nitrogen and phosphorous.

[16] The optimal level of pH is between 5.5 and 7 (NAAIAP, 2014), applying too much agricultural lime is not only inefficient, but it can lead to alkaline soils which cannot sustain crops (Kiplagat et al., 2014).

[17] A wet soil test for all nutrients through public agency is at least \$11, a soil test through private company can reach over \$20. This does not account for other costs such as transportation of samples, materials, etc.

[18] For instance, a number of projects have been launched to gather soil data, e.g. africasoils.net, soilmap.org, soilgrids.org, etc.

[19] KALRO manages a range of agriculture-related programs, and works closely with the Ministry of Agriculture to offerselected extension services.

focuses on the results from the second approach , but we discuss how the effects of both interventions compared in Section 6 and in Appendix B.[20]

KALRO's SMS program consisted of sending 20 different agriculture-related text messages to maize farmers' mobile phones. The content of the messages was developed by the Ministry of Agriculture, Livestock and Fisheries (MoALF) and the delivery was managed by KALRO.[21] Each message provided broad advice on best practices but most messages did not provide actionable advice on agricultural practices. For instance they encouraged farmers to "buy recommended certified maize and legume seed from approved agrodealers" and "obtain information on favorable market prices before you sell your harvest". One message advised farmers to test their soil's pH, and another one recommended farmers to use lime if their soil is acidic, stating: "if the soil is acidic (pH less than 5.5), apply recommended rate of agricultural lime at least 30 days before planting" it also provided farmers with phone numbers where they could inquire about purchasing a soil test to assess their farms' pH. Appendix B provides additional details and lists all the messages sent during the intervention.

KALRO's extension programs were evaluated in partnership with IPA. To recruit farmers into the program the evaluation team conducted a census of farmers in the Ugenya and Mumias sub-counties using specific walking rules to visit a representative sample of households.[22] Farmers who owned a mobile phone, had grown maize or legumes during the previous year, and were in charge of farming activities in the household were then invited to participate in the project.[23]

In September 2014, farmers completed an in-person baseline survey and were then randomized into the SMS treatment (415 farmers) or a comparison group (417 farmers).[24] Panel A in Table 2 reports summary statistics, showing balance for selected characteristics (additional summary statistics can be found in Appendix Table A2). The text-message service between July 2015 and November 2015, in the period corresponding to the short rains season. An in-person endline survey, asking information about input use and knowledge, was completed with 92% of the baseline sample by January 2016. We do not find evidence of differential attrition by treatment group (Appendix Table A1, Panel A).

At the end of the endline survey, all farmers received two (paper) discount coupons that they could redeem at selected agricultural supply dealers in their nearest market center. The coupons were devised as a way to collect information on input choices and reduce concerns about enumerator demand effects since purchasing decisions were made at a later time when farmers were not directly observed by any member of the research or KALRO team. The first discount coupon was redeemable for a 50% discount for agricultural lime. The second coupon was redeemable for a 50% discount for any chemical fertilizer of their choice (NPK, DAP, CAN, Urea or Mavuno).[25] Coupons could be redeemed until March 2016, which corresponded with the

[20] Further details of impacts of FFDs can be found in Fabregas et al. (2017a).

[21] Since 2014 the MoALF has announced plans to roll out an e-extension system to reach over 7 million farmers, by providing phone-based support to extension workers who would then advise farmers. The version of the program that was evaluated was a pilot program that tried to deliver information directly to farmers. In July 2018, the Kenyan Ministry of Agriculture and Irrigation (MoAI), in partnership with PAD and Safaricom, launched an SMS service (MoA-Info) aimed at providing agricultural advice to farmers across the country

[22] Enumerators completed a total of 1,330 surveys following these protocols.

[23] Approximately 94% of those recruited during census activities met these criteria.

[24] A third group was randomized into the FFD program.

[25] Both coupons had an upper limit discount of approximately $10 USD)

start of the subsequent 2016 long rain agricultural season. Coupons had a unique number that could be linked to the individual farmers. Participating agricultural supply dealers were instructed (and incentivized through a small payment) to keep clear records on input choices and quantities purchased by farmers who redeemed coupons.[26] Therefore, the questions of the endline survey measure behavior that occurred during the season when the program was implemented, whereas the coupons measure purchasing behavior that occurred the following agricultural season.

2.3.3 IPA & PAD Programs

PAD is a non-profit organization that supports the provision of phone-based customized agricultural information services to smallholder farmers in developing countries. PAD supported two agricultural extension research projects in Western Kenya that were implemented and evaluated by IPA. Both programs aimed to test different approaches to providing agricultural advice based on local soil information and encourage in experimentation with agricultural lime and fertilizer.

2.3.3.1 Program 1 (IPA/PAD1-K)

Throughout the 2016 short rains season, IPA, with support from PAD, sent selected farmers SMS messages with information on agricultural inputs (including lime and chemical fertilizers) as well as general agronomic recommendations on maize farming. Farmers who participated in this program were recruited from two sources: from IPA records of farmers who had previously participated in the organization's activities (47% of the final sample) and from administrative records of a large sugar cane company in the region, that worked with farmers who were also planting maize.[27] In July 2016, a sample of farmers from both databases were contacted over the phone to invite them to participate in the study and complete a short baseline survey to determine eligibility. Farmers who were planning to plant maize in the 2016 short rains season, had a farm located within the intervention area, and expressed interested in receiving agricultural information over their phone were invited to participate.[28] The phone call-based baseline survey asked questions about farmers' demographic characteristics, knowledge about soil acidity and previous input use.

Two types of messages were tested: messages with general advice for the program area that did not refer to local soil data (e.g. "Lime reduces soil acidity and makes nutrients such as phosphorous available to your maize") and messages that provided information from local soil tests (e.g. "Based on soil tests performed around [primary school/catchment area] we recommend you: apply [quantity] bottletop of lime and cover with soil and then apply [quantity] of DAP"). Among farmers receiving these specific messages, those who lived in areas that had median pH

---

[26] Incentives were paid on the basis of having both the physical coupon and a record of the purchase in their logbooks.Since these are fairly large market centers, it is unlikely that agricultural supply dealers would have known who was in the sample, unless farmers showed them the coupons.

[27] The Mumias Sugar Company ran a contract farming model with sugar cane farmers in the region up to 2015. However,the vast majority of farmers plants maize in addition to many other crops so the company supported delivery of maize extension messages. The farmers who appeared in the IPA database were mainly recruited through large school meetings, as discussed in Duflo et al. (2018).

[28] From 2,255 interviewed farmers, 2,131 consented to participate in the baseline. From that set 1,897 (89%) met the criteria for selection.

of more than 5.5 did not receive message about lime (18% of the sample). Both groups of farmers also received messages about a chemical fertilizer that is widely available to apply as top-dressing (urea). Farmers received between 24 and 28 messages. Appendix B provides additional details and lists all the messages sent.

A final sample of 1,897 farmers was randomized into three groups: receiving the general messages, receiving the specific messages, and a control group. In addition, during the following agricultural season (long rains 2017) both treatment groups received five additional messages promoting the use of agricultural lime (both groups received messages based on local soil characteristics). We show selected summary statistics, pooling both treatment arms in Panel B of Table 2 and a full list of balance checks in Appendix Table A3. We do not find evidence of systematic statistically significant differences between control and treatment groups at baseline.

We measured impacts through redemption of input discount coupons and a phone-survey. Electronic discount coupons were sent via SMS to all participating farmers at the beginning of the season after the initial set of recommendations were sent.[29] All farmers, including those in the control group, received these coupons. These electronic coupons could be redeemed at any point during the 2016 short rains season. Two coupons were sent. The first coupon gave farmers a choice of either 10 kg of lime or 1 bar of soap. By allowing farmers to choose between lime and another common product of the same value, we intended to capture farmers' input choices without liquidity constraints. The second coupon, sent mid-season, provided a 30% discount on one type of top-dressing fertilizer (urea, CAN or Mavuno), up to a pre-discount amount of 500 Ksh (approximately $ 5 USD). To redeem coupons, each farmer was assigned to an agricultural supply dealer in their preferred market center (selected during baseline). To measure effects over a second season all farmers received a second round of lime coupons for the 2017 long rain season. This coupon provided a 15% discount on the first seven 10-kg bags of agricultural lime.

The phone endline survey was conducted mid-2017 long rain season with the full sample of farmers participating in the experiment. The survey included questions about input use during the 2016 and 2017 agricultural seasons and farmers' general agricultural knowledge. Enumerators were able to survey 80% of farmers in the sample, and we do not find evidence of differential survey completion by treatment group (Appendix Table A1, Panel B).

2.3.3.2 Program 2 (IPA/PAD2-K)

A second program was implemented the following season that incorporated lessons from first program and that also aimed to focus on a population that already purchased agricultural inputs and to test a low-cost way to recruit new farmers into the system. In this version, farmers were first recruited by agricultural supply dealers. This method offered several advantages. First, it was a low-cost and quick method to recruit farmers: in a period of two months, over 7,000 farmers agreed to participate. Second, farmers who are clients of agricultural supply dealers might already be more likely to acquire inputs and, therefore, benefit from an information-based program. During the 2017 long rains season, IPA/PAD sent messages encouraging farmers to experiment with locally appropriate quantities of agricultural lime and chemical fertilizers on a small portion of their farm. In addition, a subset of farmers were also eligible to receive a phone

[29] The planting coupon was sent 10 days after the beginning of the experiment, after 7 recommendation messages, with a reminder 1 week later. The topdressing coupon was sent 1 month after the beginning of the experiment, after 18 messages,with a reminder after 10 days and another after 20 days

call that clarified the content of the text messages. All the messages were based on ward-level soil test data (additional information about recommendations is presented in Appendix B).[30] The messages focused on three types of recommendations: the use of agricultural lime in wards with median soil pH below 5.5, the use of planting fertilizer, and the use of top dressing fertilizers.

The SMS-based information service consisted of one welcome message followed by two sets of messages containing agronomic recommendations, each repeated twice. The complete list of messages is in Appendix B. Messages were sent in either English or in Swahili, depending on farmers' language preferences at the time of registration. At planting, farmers who lived in wards with pH measured to be lower than 5.5 received the following message: "The soil in your area is [very] acidic. To avoid low yields treat now. Apply [quantity] bottle tops of lime per planting hole. [quantity] kg for 1/4 acre". Farmers who lived in wards with pH higher than 5.5 received the following message: "The soil in your area is slightly acidic. According to soil analysis, farms in your area do not require lime." These messages were sent at the beginning of the planting season and re-sent 20 days later.

Farmers were also advised to use Diammonium phosphate (DAP) fertilizer at planting since it is widely available throughout the region. During the second round of fertilizer application (top dressing), farmers received two messages with information about the correct timing of top dressing fertilizer application, micro-dosing quantity for top dressing fertilizer, and how to choose which fertilizer to apply based on rainfall availability. Two different types of nitrogen-based top dressing fertilizer were recommended depending on the amount of rain experienced. If the rains were `good' the messages recommended the use of CAN, and if the rains were `poor' the messages recommended the use of urea.[31] The messages were repeated after 16 days.

Additionally, a random subset of farmers also received a phone call (or an SMS offer to receive a call) explaining the content of the text messages. This 15-minute phone call did not provide any additional information, but it allowed farmers to ask clarification questions to a PAD field officer and to hear the explanation multiple times. The purpose of the phone call was to strengthen the information provided via SMS.

To evaluate the program, a total of 102 agricultural supply dealers in 46 market centers recruited farmers into the experimental sample.[32] The registration period ran from early December 2016 to late January 2017. All registered farmers were then contacted over the phone by a member of the research team to obtain consent to participate in the study and baseline information about their farming practices and previous input use. A total of 5,890 farmers completed the phone baseline survey, met the eligibility criteria, and resided in eligible areas for which PAD had soil information.[33]

---

[30] The information was at the ward level. A ward is an administrative unit in Kenya. Wards were chosen because they are one of the smallest units that farmers can self-report and that soil tests could be mapped into. In western Kenya, the average size of a ward is 12 km^2

[31] Since it was not possible to have local rainfall patterns and make recommendations accordingly, farmers were provided with this information in order to decide which fertilizer was more appropriate based on their own observation of the rains.

[32] Initially 144 agricultural supply dealers across 60 market centers were asked to recruit their clients for a "maize farmer census". However, for logistical reasons the study area was restricted to 46 market centers.

[33] A total of 8,496 farmers were registered through the 144 agricultural supply dealers. Farmers who were reached but did not complete the baseline survey included 257 who did not consent to participate in the study, 53 who were

Farmers were randomized into four groups. The first three groups received PAD's SMS agricultural information services and the fourth group remained as a control. A third of treated farmers received information via SMS only, another third were also invited to express interest in receiving a phone call that would explain the messages, while the last third of treated farmers were contacted over the phone and offered an explanation of the messages.
Table 2, Panel C reports selected summary statistics and balance checks pooling all treatment arms. Appendix Table A4 shows balance checks and statistics for a range of different variables for each treatment. Baseline characteristics are balanced across treatment groups, with the exception of previous yields and land size, which are slightly higher for the control group. We control for these characteristics in the main specifications, but results are robust to their exclusion.

The research team collected two types of outcome data towards the end of the long rain season 2017. As in PAD/IPA1-K, all farmers participating in the experiment received two electronic coupons via SMS. Each coupon allowed farmers to obtain discounts on agricultural inputs from a local agricultural supply dealer. The first electronic coupon was redeemable for 15% on the first seven 10-kg bags of agricultural lime, and the second coupon provided a 15% discount on the first 1,000 Ksh (approximately $10 USD) spent on top dressing fertilizers (urea, CAN, or Mavuno).

To ensure that all farmers in treatment and control group were equally aware of the coupon, all farmers received a phone call a week before the program started, in which an enumerator explained how to use the coupon and at which agricultural supply dealers the coupons could be redeemed.[34] In addition, after the end of the agricultural season, farmers completed a phone survey that included questions about input use during the agricultural season in which the program took place and general agricultural knowledge. Around 84% of farmers completed the endline survey and we do not find evidence of differential attrition by treatment arm (Appendix Table A1, Panel C, column 1).

2.3.4 One Acre Fund's Programs

1AF is a non-profit social enterprise that provides training and agricultural inputs on credit to smallholder farmers across six countries in Eastern and Southern Africa. In 2017, they reported working with over 600,000 farmers (1AF, 2017). The 1AF model relies on training farmers on modern agricultural techniques and providing them with seeds and fertilizer on credit. To receive the 1AF input loan and training program, farmers must join a village group that is supported by a local 1AF field officer. Farmers sign contracts with an 1AF field officers well before the agricultural season starts and get inputs delivered right before the beginning of the planting season. Farmers repay their loans at any time during the growing season. 1AF clients form groups of eight to eleven farmers who participate in the program together through several shared activities, including signing a contract together and being jointly liable for their loans.

---

not planning to grow maize in 2017, and 40 who lived outside the four counties in which recruitment took place. Approximately 1,017 farmers lived in wards for which there was no soil test data available.

[34] During the IPA/PAD1-K experiment, farmer coupon redemption for fertilizer was lower for those in the treatment groups. One potential explanation is that the due to a higher number of messages in the treatment group, farmers did not read the electronic coupon. Therefore, all farmers including those in the control group received the phone call about the coupon. To avoid priming farmers about agricultural lime, they were just told that the coupon would provide them with a discount for an agricultural input. 93% of farmers were reached during this activity.

The standard bundle that 1AF offers includes hybrid seeds and chemical fertilizers. However, to address the problem of high soil acidity, 1AF started offering farmers agricultural lime as an optional add-on. Yet, across their many locations, demand for lime remained very low. Hypothesizing that this could reflect a lack of awareness, 1AF designed and evaluated several informational programs to increase lime take-up. Since 1AF field officers already follow detailed training protocols, a key objective was to test cheap programs that would not require additional field officer training and delivery. We describe their different strategies below.

2.3.4.1 Program 1 Kenya (1AF1-K)

Prior to 2016, less than 3% of 1AF clients in western Kenya purchased agricultural lime through the organization (1AF, 2015). To increase take-up, 1AF designed a phone-based extension pilot that consisted of six rounds of text messages targeting clients who had signed up for the 1AF package during the previous season in a selected district of western Kenya.

1AF tested two versions of the messages. One group of farmers received simple SMS messages encouraging lime use and providing them with a customer engagement toll-free line which they could call to receive more information. The message read "Hello [name], Your soil is acidic. Use lime to reduce acidity and increase yields. Call xxx-xxxx". A second randomly selected group of farmers received a more detailed message that mentioned the level of acidity measured in the farmer's area as well as the amount of lime recommended and expected return to its application: "Hello [name], Your soil is [highly/moderately] acidic. We recommend [amount] kg of lime per acre at [total cost] Ksh. Use lime to reduce acidity and increase yields by [percentage]%.Call xxx-xxxx".[35] Customized messages were based on soil tests that had been previously conducted in the region. We discuss how these recommendations were constructed in Appendix B. In total, 4,884 farmers participated, with 3,325 farmers randomly assigned to receive messages, and 1,559 farmers remaining as a control. The same SMS message was sent six times between August and September 2016, before the 1AF input contract signing period, when farmers had to decide whether to request inputs from 1AF for the following season.

To simplify our exposition we pool together the two treatment arms, but we discuss differences across treatment arms in section 5. Selected summary statistics are reported in Panel D of Table 2 and a full list of balance checks in Appendix Table A5. Since 1AF does not collect extensive demographic data we can only show a limited number of farmer characteristics at baseline. Running balance tests for twelve characteristics that 1AF had for the farmers, which mostly included the products that farmers had purchased in previous seasons, we only find small differences at baseline those in the treatment arms were less likely to plant onions, purchase CAN fertilizer, and receive a repayment incentive the previous year. We control for these variables in our main specifications, but the results are robust to their exclusion.

For this sample we can measure outcomes using two sources of data: 1AF administrative data and phone survey data collected by researchers. The administrative data contains information on loan enrollment and inputs purchased through the 1AF program. However, only 60% of farmers who received the text messages signed-up to receive loans in the 2017 long rain agricultural season. While we do not find evidence of a differential likelihood of placing an order by treatment arm (Table A1, panel D, column 2), we take a conservative approach in our main

[35] The percentage increase in yields depended on the local level of pH and the return estimated for that pH level based on 1AF farm trials.

specifications and define the outcome variable as lime purchased through 1AF. This outcome is an imperfect measure of the overall effects of the program on lime purchases if farmers acquired lime from other sources. To explore this possibility and obtain additional information from farmers, a follow-up phone survey led by IPA was conducted in May 2017 with a random sample of 30% of the farmers participating in the trial.
This survey asked respondents about their knowledge of lime and their input use during the 2017 long rains season. About 79% of selected farmers were surveyed, and we do not find differential treatment attrition for this sample (Table A1, panel D, column 1).

2.3.4.2 Program 2 Kenya (1AF2-K)

A second 1AF program was implemented with approximately 30,000 farmers in four Kenyan districts in September 2017. Former 1AF clients were randomized into a no message control group or a treatment group receiving SMS messages encouraging lime adoption (which did not depend on results from soil tests in the area). Additionally, a quarter of farmers were randomly assigned to receive additional messages encouraging the use of additional fertilizer (Extra CAN) for a second round of top dressing.

The messages randomly varied how the lime information was presented, number of repetitions (1 to 5 messages), and time between repetitions (every 2 to 8 days). Six main categories of messages were sent ranging from a basic messages that simply recommended to buy lime "[Name], 1AF recommends you register to buy Lime for your maize.", to messages encouraging experimentation "[Name], 1AF recommends you register to buy Lime for your maize. Try it on just a small part of your land to so that you and your neighbors can see the benefits.", or leveraging on social comparison "[Name], 1AF recommends you register to buy Lime for your maize. Farmers all over Western are getting bigger yields by using lime. Keep up with them!". For simplicity, we pool all the different treatment arms in the main tables, but all the information on different treatment arms can be found in Appendix B.[36]

Summary statistics and balance checks for treated and control farmers are reported in Panel E of table 2 and additional summary statistics and balance checks by treatment arm are in Appendix table A6.. Apart from a small differences in land reported, we do not find evidence of consistent statistically significant differences at baseline between treatment and control arms (we also reject the null of joint significance). Farmers were later matched to 1AF administrative data to measure their likelihood of demanding agricultural lime and other inputs for the following agricultural season. Only 76% of farmers who received text messages decided to acquire any inputs through 1AF, but we do not find differential likelihood of purchasing inputs by treatment status (Table A1, Panel E). Again, we define the primary outcome variable as the probability of purchasing agricultural lime from 1AF.

[36] In the same period, 1AF also conducted two other programs to encourage lime adoption via SMS. In Nambale district, where the 1AF1-K program took place, a randomly selected subset of farmers that did not purchase lime during the previous season were matched to lime users and encouraged to talk to them to learn more about the product.A version of the 1AF2-K program that did not involve topdressing fertilizer messages was implemented during the same period in Nambale. To simplify the exposition we exclude this district from the sample analyzed in this paper, however, its inclusion does not change the main results. Farmers outside the trial districts (excluding those in non-acidic areas) received different variations of SMS messages encouraging lime adoption. The content of the messages, number of repetitions, and frequency were randomly assigned. This component involved approximately 180,000 farmers. In this paper we focus exclusively on the results of the first program.

2.3.4.3 Program 3 Rwanda (1AF3-R)

In 2017 a modified version of the Kenya program was implemented in Rwanda. In Rwanda, 1AF (known as Tubura) partners with the government to provide goods, services, and training to rural farmers. Since 2016, 1AF and the government of Rwanda have engaged in a concerted effort to promote adoption of travertine, a type of agricultural lime. Activities involved marketing lime, widespread soil pH testing, and offering substantial price subsidies (75% off the price) in several districts. 1AF reported that in districts where the price subsidy was offered, lime demand went up from 7% to 21%. Since all these interventions were costly, 1AF also decided to test the effectiveness of text-messages as an inexpensive way to increase lime use.

In June 2017, during the enrollment period for the 2018 main agricultural season (September 2017 to January 2018), a large-scale program aimed at increasing lime adoption through the use SMS messages was implemented in all districts where 1AF operates. Since phone ownership is much lower in Rwanda than in Kenya (only 53% of the farmers registered in the program had a phone number reported in the database) one of the objective of this program was to measure spillovers among farmers in the same group, in particular to those who did not own a phone. Therefore, the randomization was done at the farmer group level, with some groups partially treated.

As in 1AF2-K, the messages varied content, framing, and number of repetitions.[37] In Rwanda, agricultural lime is known as travertine. Seven types of messages were sent, ranging from a general promotions "Many fields in Rwanda have acidic soil and need TRAVERTINE to increase yields. Order from 1AF now.", to messages explaining that "[acidity] blocks fertilizer uptake." and "Applying TRAVERTINE solves the problem, increasing crop yields". Another messages tried to create a sense of urgency urgency by using wording like `Order it immediately". All messages were either gain-framed or loss-frame with respect to yield increases generated by lime use. In addition to these messages, farmers in half of the treated groups received an additional message encouraging them to share the information with other farmers, especially those without phone. Additional information, including the complete list of messages, can be found in Appendix B.

From a total of 216,475 farmers registered in the 1AF program, only 114,569 had a phone registered in the database, and 85,160 had a unique phone number. Since the unit of randomization was the group and farmer (rather than phone number), some phone numbers shared among more than one farmer were sometimes assigned to multiple treatments. In our analysis we drop all farmers that did not have a phone registered in the database and consider the original treatment assignment, regardless of whether phones are shared or not. The main results are robust to excluding all farmers with shared phones from the analysis. Summary statistics and balance checks for treated and control farmers are reported in Panel F of table 2 and additional summary statistics and balance checks by treatment arm are in Appendix table A7. We detect some small differences in group characteristics and input purchases from 1AF in the 2017 main agricultural season (we also reject the null of joint significance). We measure whether farmers purchased lime from 1AF for the 2018 main agricultural season. Table A1 panel F shows that

[37] In Rwanda, agricultural lime is known as travertine. The messages read ”TRAVERTINE” not ”LIME”.We changed this for simplify of exposition.

only 65% of control farmers enrolled in the 1AF loan program, and that treated farmers were .9 percentage points more likely to purchase inputs from 1AF than those of the control group. Since we define the outcome as purchases from 1AF, this does not affect the interpretation from the coefficient.

2.4 Empirical Strategy

2.4.1 Estimating Impacts for Individual Studies

We evaluate the impact of the six programs on farmer agricultural knowledge and experimentation with the recommended inputs. Since all programs focused on agricultural lime, we focus on that variable. We also look at fertilizer acquisition for the programs that recommended it. In all cases we estimate intention-to-treat effects.[38] For ease of exposition we pool treatment arms for the IPA/PAD and 1AF programs, though disaggregated tables can be found in the Appendix. Therefore the general equation we estimate is:

$$y_i = \alpha + \sum_{j=1}^{z} \beta_j Treatment_{ij} + X_i \nu + \gamma_w + \varepsilon_i \quad (1)$$

y_i is the outcome measure for farmer i. $Treatment_{ij}$ denotes a dummy variable indicating each treatment group j. X_i is a vector of controls for farmer specific characteristics, γ_w controls for area fixed effects and ε_i is the error term. The coefficient β_j estimates the difference between treatment j and control.

For binary outcomes, we estimate equation (1) both with a logistic regression model and a linear probability model. For the former, we report the coefficient β_j both as percentage points increases and in terms of odds ratios (OR) for the probability of acquiring the input.

In order to improve precision and address some small baseline imbalances in our main specification, we include control variables. In particular, we control for the strata used in each randomization, demographic characteristics, farming practices, previous input use, and for the survey data we include enumerator fixed effects. We also control for variables that were found to be unbalanced at baseline.

The KALRO sample includes controls for gender, hearing about lime at baseline, index of baseline input use, grown legumes, land size, heard about soil tests. The controls for IPA/PAD1-K include: age, gender, primary education, sample of origin, preferred language, phone network, farm size, knowledge score at baseline, previous input use, and measures of interest in the program at baseline. The controls for IPA/PAD2-K include: age, gender, preferred language, farm size, previous lime use, and agricultural supply dealer (recruiter) dummies. The 1AF1-K sample includes controls for number of seasons in the program, repayment incentives received, ordering: size of maize package, bean seeds for intercropping, compost boost products, solar lamps, cook stoves, extra CAN, harvest sheets, storage bags, onion seeds, health insurance, and sanitary pads. The 1AF2-K sample includes controls for number of seasons in the program,

[38] Some of the treated farmers did not receive some or all of the messages (for example only farmers with a phone from the main network in the area could receive messages during the IPA/PAD1-K, for those in 1AF3-R some farmers did not receive messages because they did no own a phone and had listed someone else's) and some farmers assigned to the phone call treatment in IPA/PAD2-K were not reached. Moreover, since information on who actually reports receiving the messages is self-reported and could be contaminated by recall bias, we opt to not use treatment-on-treated effects.

group size, predicted pH level in the area, size of the maize package, and indicators for whether the farmer purchased solar lamps and extra CAN in the previous season. The 1AF3-R sample includes controls for number of seasons in the program, group size, a dummy indicating whether the farmer shares phones with others, number of farmers in group with a phone, and administrative information from the 2017 main agricultural season including: credit size, quantity of fertilizer (DAP and urea) purchased, and indicators of whether the farmer purchased lime and solar lamps. Finally, since the randomization was at the group for the 1AF3-R experiment the errors terms are clustered at that level.

2.4.2 Summarizing Impacts Across Studies

To formally synthesize the evidence across the various settings and present a weighted average of study estimates, we conduct a meta-analysis. We use both fixed effects and random effects models for this meta-analysis, where the fixed effects model is based on the assumption that there is a common effect across all the studies whereas the random effects model assumed that there is a distribution of true effects across settings, and that for each setting, we observe a treatment effect is drawn from a distribution centered at the true effect. Arguably, the random effects model might be better suited for this setting since the effects are likely to be heterogeneous across contexts, therefore most of our discussion focuses on this model.[39]

While the exact nature of the outcome measures varied across studies, we combine them (e.g. coupon redemption and lime purchases) to reflect a single common outcome representative of certain behavior. Since most of our dependent variables are binary, we present results in terms of odds ratios.

2.5 Results

2.5.1 How similar are the targeted populations?

Each program recruited farmers into the system in different ways. Farmers who participated in the KALRO program were targeted to be a representative sample of farmers in the population and who would benefit from the program. For IPA/PAD1-K researchers targeted farmers who had been engaged with existing organizations. The IPA/PAD2-K program worked with current clients of agricultural supply dealers. In all cases, 1AF targeted clients from previous seasons. Table 2 shows summary statistics on selected baseline characteristics for farmers from each experiment. Except for the IPA/PAD1-K sample, where the proportion is almost reversed, about two thirds of participants are females. On average, they have less than 2.5 acres of land (the land variable for the 1AF samples denotes the land size for which they reportedly purchased maize inputs from 1AF in the previous season, it does not necessarily correspond to the land size they own). Column 5 reports the probability that farmers report using (or purchased in the previous year in the 1AF3-R case) agricultural lime in the past. For both the 1AF3-R and KALRO samples, 6% of respondents report having used it (purchased it) in a previous season, whereas for IPA/PAD1-K and IPA/PAD2-K, about 12% and 9% report using it, respectively. Overall, this suggests that lime is not a widely used input. Differences in take up

[39] Observed study estimates are given by j,y|j,yN(j,y,j,y2) where j,yN(y,y2)is independently and identically distributed and the weights are given by 1-(2+2). Where j indexes study and y outcome.

across samples, could reflect differences targeted populations, but also, the fact that lime use has been strongly encouraged in Western Kenya during the time period of the experiments, so that later samples have higher lime adoption. In all samples, the majority of farmers had used chemical fertilizers in the previous agricultural season (column 5). Finally, in samples we have information for, over 50% of farmers report completing primary school (column 3).

2.5.2 Take-up and Message Reception

The programs were popular among farmers. None of the farmers who were invited to the KALRO program opted out. In the IPA/PAD1-K and IPA/PAD2-K, 95% and 99.5% of the farmers surveyed at baseline agreed to received the messages respectively. There was no opt-in process for the 1AF messages, since they were part of regular 1AF activities.

However, receiving a call from an agricultural field officer was less popular. Both the 1AF1-K and IPA/PAD2-K offered this service to farmers as an add-on. Only 8% of farmers in the IPA/PAD2-K sample who received both messages and an offer to receive a call, requested a phone call during planting season. In the 1AF case, farmers also had access to a toll-free number, but only about 1% of treated farmers (35 callers) used it to ask questions about lime (not shown).

Another dimension of take up has to do with whether the farmers received the messages. In the KALRO intervention only 55% of farmers in the e-extension group stated that they had received messages. Low message reception is correlated with whether the farmer received services from the main network provider in the area.

2.5.3 Awareness and Knowledge about Lime

We can only report on the four studies in which survey data was collected. There is some variation in the way the question was posed across different evaluations but in four all studies we collected measures of whether farmers had heard about lime and whether they knew lime was a remedy for soil acidity.

We first present the meta-analysis results for the average treatment effects with 95 percent confidence interval. Rows 1 and 2 in Table 3 present the random (column 1-4) and fixed effects (columns 5-8) meta-analytic results for our measures of awareness and knowledge. The odds ratio for awareness of lime is 1.21 (95% CI 0.93 to 1.57) and for the effect on knowledge of acidity is 1.57 (95% CI 1.40 1.75). Figure 2 displays these results graphically. Overall, we conclude that while in other contexts knowledge has shown not to be improved by extension services (Cole and Fernando, 2016), in these case farmers learned about the use of lime through the SMS messages.

We now describe the effects of each program. Column (1) in Table 4 estimates whether the treatments increased farmers' awareness about agricultural lime ("Have you heard of agricultural lime?"}). This question may only measure whether lime had been made salient to them.Column (2) shows estimates for whether farmers, unprompted, mention lime as a way to reduce soil acidity: for the KALRO sample ("Do you know strategies to reduce soil acidity? If yes, could you mention some of them?"}), for the IPA/PAD samples we asked ("What is the best way to deal with soil acidity?"}), and for 1AF1-K ("If you tested your soil and it was shown to be acidic, what would be the best way to reduce soil acidity?"}).

We do not find that the KALRO program increased knowledge about the existence of lime or its main use (Panel A). The coefficients are small and statistically insignificant. A likely explanation is that farmers only received one message on soil acidity. The effects on lime knowledge from IPA/PAD1-K are positive but statistically insignificant. Those who received the messages were 8 percentage points more likely to know that lime can reduce acidity (Panel B). Panel C shows effects for IPA/PAD2-K program. In this instance the program increased the likelihood that farmers reported knowing about agricultural lime by 5 percentage points and knowing its main use by 10 percentage points. Finally, Panel D shows results for the 1AF1-K sample. A large fraction of farmers in the control group (80%) had heard about agricultural lime, in line with 1AF efforts to promote these inputs to their clients, but we do not find differences in knowledge about it by treatment status. However, treated farmers are 10 percentage point more likely to know that agricultural lime can be used to reduce soil acidity. Differences in effects might be due to the fact that the second version of the IPA/PAD program and the 1AF1-K more focused on lime, and farmers received a higher number of repeated messages within those programs.

2.5.4 Purchase of Agricultural Lime

We now examine one of the key behaviors that the programs were expected to affect: acquisition of agricultural lime when it was recommended. We present evidence from two sources survey data and administrative records. In addition, we show results on whether purchases occurred concurrently with the program or during a subsequent agricultural season.

Table 3 shows the summary results. The odds ratio for following the lime recommendation, according to the survey data was 1.58 (95% CI 1.35 to 1.83). This includes the KALRO, IPA/PAD1-K, IPA/PAD2-K and 1AF1-K studies. Using administrative records for all studies (except for KALRO, for which we only have administrative records for the subsequent season) the odds ratio coefficient is 1.19 (95% CI 1.12 to 1.26). The effects for a second season (using administrative records and estimated from the KALRO, IPA/PAD1-K, IPA/PAD2-K and 1AF3-R studies). The average odds ratio effect is smaller 1.07 and we cannot reject the null of no effect (p-value=0.13). Figure 3 displays these results graphically. In all cases we cannot reject the null of homogeneous treatment effects across programs.

Table 5 shows the results from each experiment on whether farmers followed the lime recommendations provided by each system. Columns (1) to (4) show impacts of the messages for the first season of the program. Columns (5) and (6) measure effects on a subsequent season, for those programs that collected this information. Columns (2) and (5) show the impacts of the programs on whether the lime recommendations were followed, as measured by the administrative data (purchases or coupon redemption). Columns (3) and (6) restrict samples to those farmers in 1AF who decided to enroll in the program (we had previously shown that except for 1AF3-R, the messages had no impact on the likelihood of enrolling into the program). Columns (1) and (2) show these coefficient for the self-reported survey data. For the PAD/IPA programs, we consider the recommendation being followed if farmer used lime and lime was recommended or if the farmer did not use lime and lime was not recommended. For KALRO and 1AF positive amounts of lime were recommended to all farmers.

Panel A presents the results for the KALRO sample. We do not find evidence that the program increased self-reported use of lime during the season when it was implemented (Column 1), nor did it increase the redemption of lime coupons during the following season (column 5). Panel B shows the results for the first version of the IPA/PAD program. The coefficient on the redemption of lime coupons (choosing lime over a soap bar of similar price) shows that treated farmers are 2 percentage points more likely to follow lime recommendations (choose lime where recommended and not choose lime or not redeem when not recommended) but the standard errors are almost as large (column 2). However, treated farmers are 4 percentage points (19%) more likely to report following lime recommendations during the survey (column 1).

Panel C shows the corresponding results for IPA/PAD2-K, where farmers were recruited from agricultural supply dealers and received more focused messages. Using coupon redemption as our outcome variable, we show that the program increased the likelihood of following the recommendation (redeem if in area where recommended and do not redeem otherwise) by 3 percentage points. The point estimates from the self-reported data are larger. The effect of the program increase self-reported likelihood of following lime recommendations by 7 percentage points (a 24% increase).

Panel D displays impacts for the 1AF1-K sample. The program increased the probability of purchasing lime as measured by the administrative data by 3 percentage points (corresponding to a 30% increase). Farmers are also 4 percentage points more likely to report using lime in the phone-based survey (column 1). This suggests that there were not significant lime purchases from sources other than 1AF and that, at least for this group, there is little social desirability bias in self-reported data. We do not find that these effects extend to the following season.[40] Finally, column (3) estimates the effects for farmers who decided to enroll into the 1AF program following the intervention. This conditions on a post-treatment variable, so we do not put much weight on this specification, but as expected, the impacts are slightly higher (5 pp) for this group.

Panel E presents effects for the 1AF2-K program. The point estimates are similar to the first version of the program. We estimate that the program increased 1AF lime purchases by 3 percentage points (a 6% increase). We cannot compare to self-reported data since we do not have survey data for this sample. Panel F shows the results for the 1AF3-R program. We find that the treatment increase the likelihood of purchasing lime by .8 percentage points (a 20% increase).

For 1AF, the self-reported data lines up well with the administrative reports. However, for the IPA/PAD programs, particularly the second program, there is a significant difference between the size of the estimated coefficients using survey data and the coupon redemption data. One possibility is that for some farmers the survey data is affected by social desirability bias, and they might have over reported true lime use. This would overestimates true input use. A second possibility is that the coupon redemption underestimates true lime use, since farmers might have acquired lime from sources other than the shops where the coupon was valid. The true effect is likely lie between those point estimates.

To explore these possibilities we check whether farmers who might have other sources of lime are also more likely to report using lime but not using the coupon. Farmers who are clients of 1AF at baseline (35% of the IPA/PAD2-K sample) are more likely to report using lime but not redeeming the coupon: being an 1AF farmer is associated with a 4 percentage point higher likelihood of reporting using lime in survey but not redeeming the coupon (from 8 to 12%). Second, using data from a survey of agricultural supply dealers in the region about the products they stock, we find that only 36% of farmers who report using lime in the survey (but who did

[40] A subset of the farmers in this sample, both treated and untreated, SMS-based program the following season (a variation of 1AF2-K program). However, the treatment assignment of the new program was designed to be orthogonal to the original one so it should not prevent us from observing a second season treatment effect for 1AF1-K. The results are robust to controlling for treatment status in the second season (not reported).

not redeem coupon), said that they had acquired the lime from a shop that reported stocking lime during the intervention period.[41] This might be suggestive of misreporting in the survey data.

2.5.5 Purchase of Chemical Fertilizers

Next we examine the impact of these programs on the other type of input that was systematically recommended by four out of the six programs: chemical fertilizers. Since the 1AF1-K and 1AF3-R only focused on lime we do not include them in this section. In this sample the estimated odds ratio increase in the likelihood of purchasing fertilizer was was 1.02 (p-value= 0.80) and 1.31 (p-value < 0.00). The significant effects are dependent on the inclusion of 1AF2-K in this sample. Figure 4 shows the results and the test for homogeneity across programs (which we fail to reject).

Table 7 shows these results for each experiment. We see a positive effect from KALRO's program on coupon redemption, but the estimate is noisy (Panel A). We do not find evidence that the SMS-extension increase purchases of recommended fertilizers for the IPA/PAD samples, neither in self-reported data nor in redemption of coupons (Panels B and C). For 1AF2-K we estimate that receiving lime messages increased the likelihood of purchasing extra top dressing fertilizer by 1 percentage point. Within the group that was randomized to receive messages about lime and "Extra CAN" (a double amount type of top dressing fertilizer) the overall increase in 3 percentage points (Panel D).

2.5.6 Other Inputs and Practices

Both the KALRO and the IPA/PAD1-K programs provided a range of different agronomic recommendations in addition to the use of agricultural lime and fertilizers. In addition, it is possible that for the other programs, there might have been some substitution away from using other inputs. We construct an index for experimentation with all other possible inputs and practices for which we had information in each program. Table 3 shows these effects in terms of standard deviations. We do not find systematic evidence of an average decrease in purchase of inputs or adoption of other practices. We find a small but positive and statistically significant average impact of 0.01.

We look at these results for each project. Table 8 explores the impacts of the programs on use of other inputs and practices. Since there is a range of potential variables that could be affected, we standardize treatment effects following the construction of indices as per Kling et al. (2007).[42] The variables used to construct each index varied by sample, since different information was collected for each case (Table A12 in the appendix reports the list of variables used to construct each index).

[41] If we impute a zero for those farmers who reported using lime, but who who did not redeem coupon and reported obtaining the lime from a shop that did not stock it, the coefficient is 0.05 (0.011).

[42] We use the seemingly-unrelated regression framework to account to covariance across estimates.

2.6 Discussion

2.6.1 Are there information spillovers?

Randomization in these projects was done at the individual level. If farmers who participated in the program diffused the information to other farmers in the control group, we would underestimate impacts. While this is a possibility, other work in this area Duflo et al. (2008) and our own qualitative work supports the idea that there is overall low diffusion among farmers (Fabregas et al., 2017a). We also do not find significant changes in knowledge about the new inputs between baseline and endline estimates for the control group. However, it is likely that with more proximity among farmers (social and geographical) there would be more diffusion. Since 1AF clients work in groups, this would be a more important concern for individuals in those samples.

2.6.2 Did farmers understand the messages?

One question is whether farmers could read and correctly interpret the lime recommendations. If this is the case, a more intensive program where a person could explain the recommendations verbally would be more effective. We can test this hypothesis, by looking at the differential effects of each treatment arm of the PAD/IPA2-K experiment. In one arm farmers received text messages, and in other arm farmers received text messages and a phone call from an extension officer who explained the content of the text message (no new information was provided). We do not find statistically significant differences between treatment arms (p-value 0.369). These results are shown in in Appendix Table A9 (Panel B column 4).

2.6.3 Who is most responsive to the programs?

An important concern around these systems is whether some farmers might systematically be less or more likely to benefit from the programs. For instance, those with lower levels of education might find it difficult to use a phone to receive information. Similarly, traditionally excluded groups from from other information sources might benefit from these programs. We test these hypothesis, by estimating differential effects by gender, levels of education and farm size. Table 9 shows these results for lime coupon redemption (KALRO, IPA/PAD) or purchase (1AF). Overall, we do not find significant evidence that the effects varied by gender, education level or farm size.

2.7 Cost-Effectiveness

We conduct back-of-the-envelope calculations to provide some estimates of the cost-effectiveness of these interventions. To establish benefits, we combine information from the point estimates from the increased adoption of lime with existing agronomic data to estimate corresponding effects on yields from the increase of lime for an average farmers.[43] Agronomic trials performed in Western Kenya found that lime application on average, increases maize yields

[43] While these calculations rest on a number of assumptions, collecting maize outcome data would have been extremely costly.

by 2.47 kg per kg of lime applied (1AF, 2015). The cost of applying 1 additional kg of lime is estimated to be approximately \$0.15, which takes into account the local price of lime (\$0.10/kg) and transport and application cost (\$0.05/kg). The revenue obtained from one additional kg of maize is assumed \$0.35, which takes into account the market price of maize (\$0.40/kg), minus additional costs for harvesting and transport (\$0.05/kg).[44] Therefore, the profits from applying one kg of lime is estimated to be approximately \$0.71.

Ignoring other fixed costs related to setting up the programs, the cost of sending text messages at bulk rates is very low (and it would be lower if the social costs were used). We estimate that the cost of sending each message was under \$0.01. We find that the increased the quantity of lime used by 1-3.5 kg. Using these estimates the benefit-cost ratio is between 6-36.

We can compare these estimates to that of the in-person extension approach that was implemented by KALRO. In particular, we have experimental estimates of how Farmer Field Days, large meetings with farmers, affected the use of agricultural lime. Based on information reported by KALRO, we calculate that each FFD cost about US\$2,600 to implement. This includes all costs for staff, transport, compensation and materials required to set up the test plots, invite presenters, advertise the FFDs to farmers and carry out the events. Since each FFD hosted between 100 and 300 farmers, this amounts to a per farmer cost of at least US\$9.

2.8 Conclusion

The spread of cellphones in developing countries has opened new opportunities to reaching individuals: whereas in-person communication can be costly, text message delivery can be cheap. How effective are these approaches? This project focuses on the case of agricultural extension, and finds that simple text messages sent to farmers can improve learning and increase experimentation with some technologies in a cost effective manner.

A contribution of this study is to test these in different settings. We evaluate the effects of six programs that were designed to increase farmer knowledge and adoption of agricultural lime and fertilizers. Using a combination of survey and administrative data, we document that, on average, the programs increased purchases of agricultural lime. The results were mixed for fertilizer, a more commonly known input.

We also provide evidence to show that providing detailed content and other add-ons (e.g. additional phone calls by a field officer) did not significantly increase the impact of simple text messages.

[44] Maize prices and assumed costs are based on data collected by IPA in the study area during the 2017 main agricultural season.

2.9 Tables and Figures

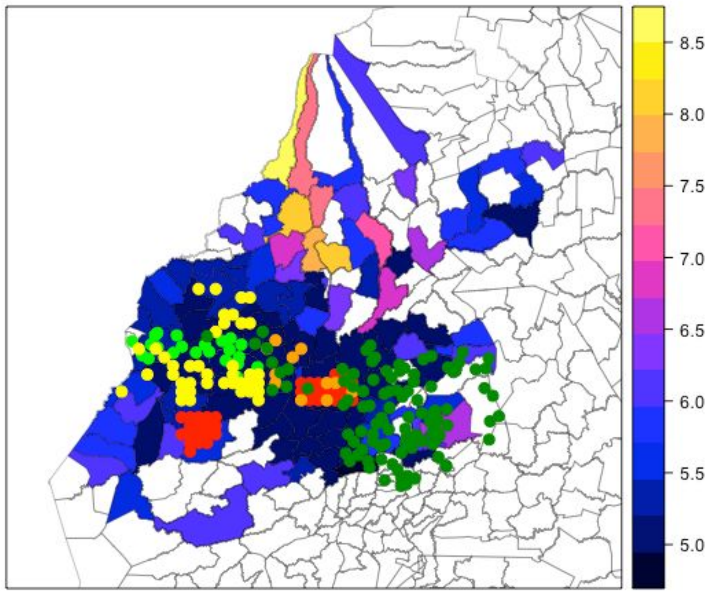
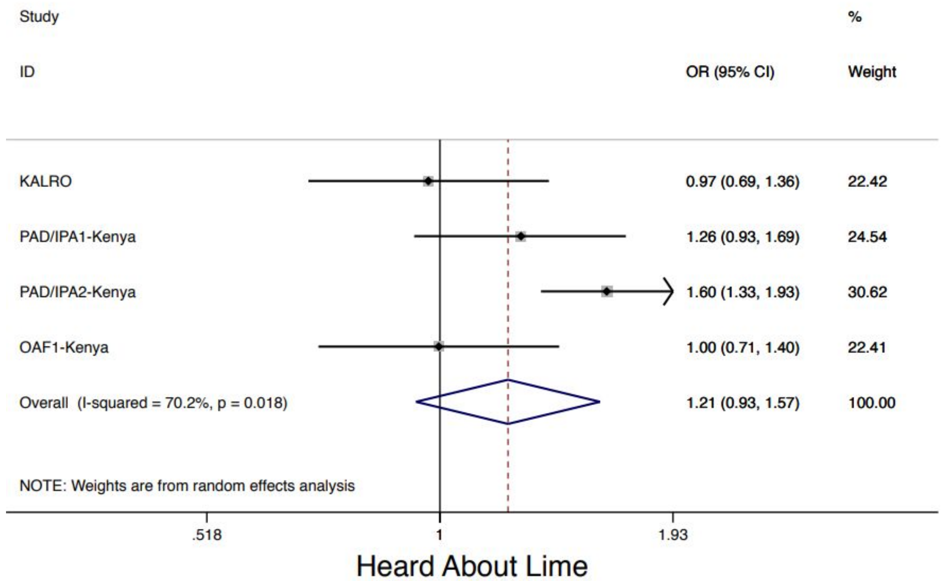
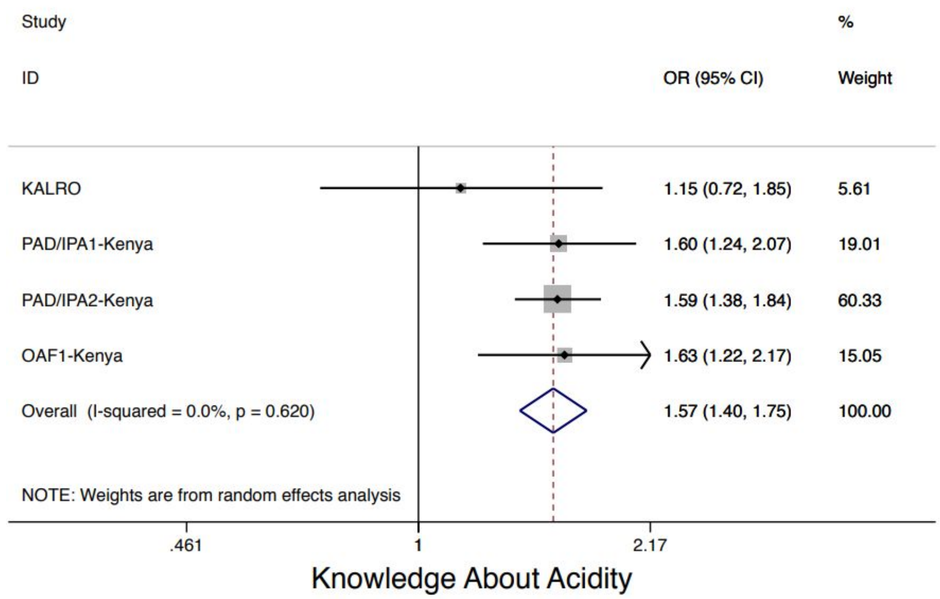


Figure 1: Project Map in Kenya[45]

[45] The map shows the median level of pH in all wards in which the IPA/PAD2-K program took place. Red dots indicate the location where the KALRO program took place (approximate farmers' locations). Green dots indicate the location of the 1AF experiment (1AF sites): light green for 1AF1-K and dark green for 1AF2-K. Yellow and orange dots indicate the location of IPA/PAD1-K program (yellow: schools where the IPA/PAD1-K sample farmers were recruited, orange: sublocation centroid for farmers from sugar cane company)



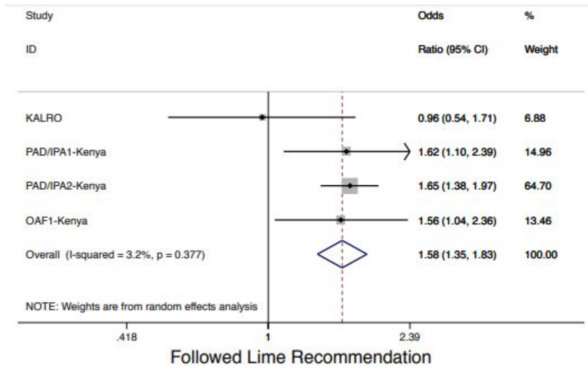
(a) "Have you heard about lime?"



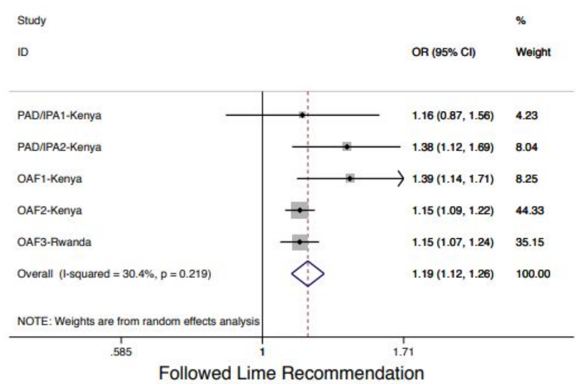
(b) Mentions Lime as a way to reduce acidity

Figure 2: Combined Effects on Knowledge About Lime[46]

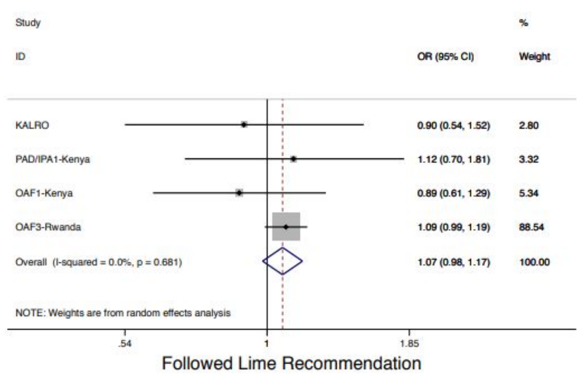
[46] The figure plots the meta-analysis results for specific outcomes. The effects are estimated using a random-effects meta-analysis model. Results are reported in odds ratios. The horizontal lines denote 95% confidence intervals.



(a) Survey Data (First Season)



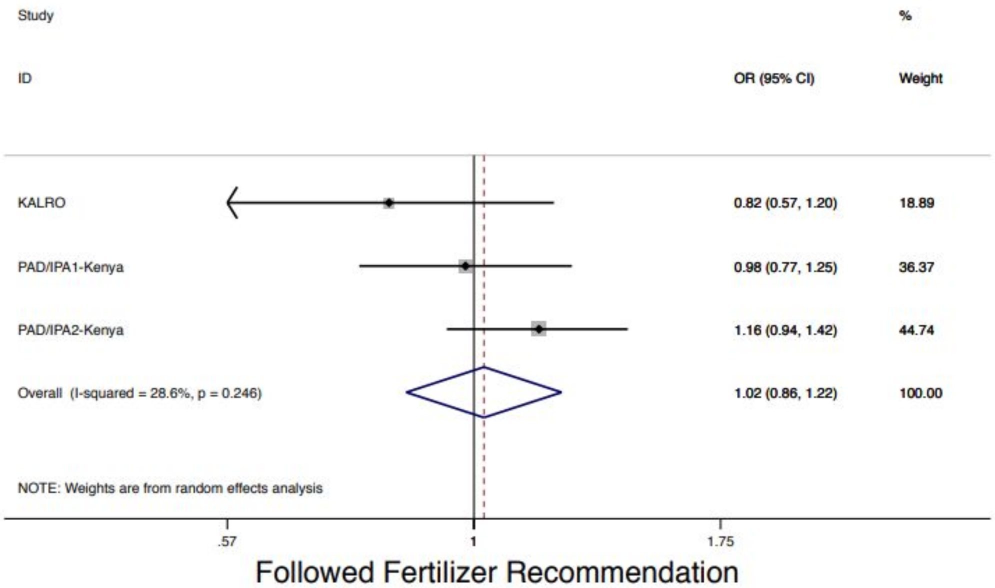
(b) Administrative Data (First Season)



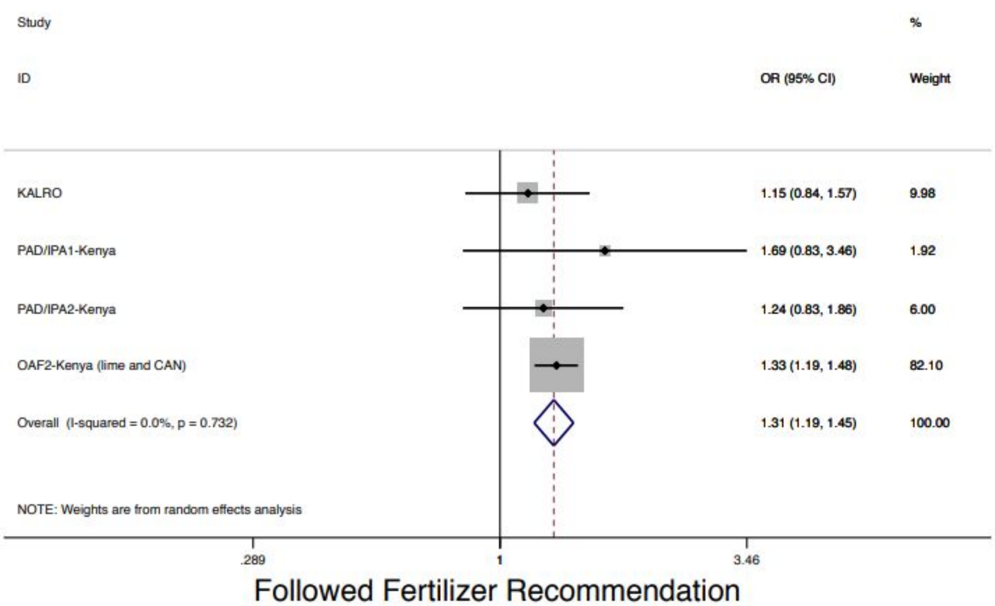
(c) Administrative Data (Second Season)

Figure 3: Combined Effects on Lime Purchases[47]

[47] The figure plots the meta-analysis results for specific outcomes. The effects are estimated using a random-effects meta-analysis model. Results are reported in odds ratios. The horizontal lines denote 95% confidence intervals.



(a) Survey Data



(b) Administrative Data

Figure 4: Combined Effects on Fertilizer Purchases[48]

[48] The figure plots the meta-analysis results for specific outcomes. The effects are estimated using a random-effects meta-analysis model. Results are reported in odds ratios. The horizontal lines denote 95% confidence intervals.

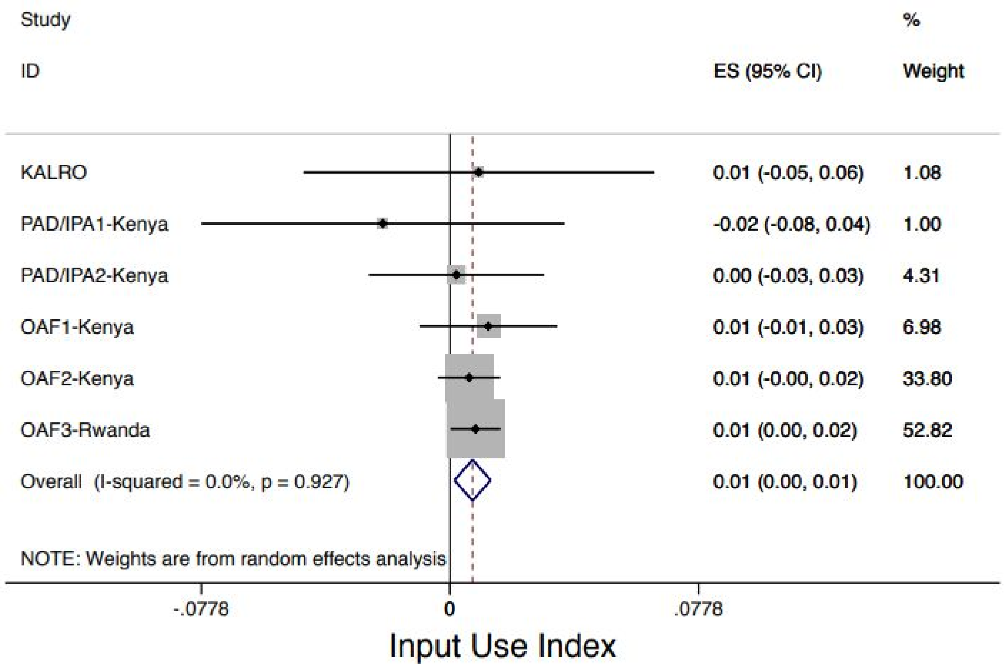


Figure 5: Combined Effects on Other Inputs (Index)[49]

[49] The figure plots the meta-analysis results for specific outcomes. The effects are estimated using a random-effects meta-analysis model. Results are reported in standard deviations. The horizontal lines denote 95% confidence intervals.

Table 1: Characteristics of the Programs

	KALRO	IPA/PAD1-K	IPA/PAD2-K	OAF1-K	OAF2-K	OAF3-K
Location	Kakamega and Siaya (Kenya)	Busia and Kakamega (Kenya)	Busia, Bungoma, Kakamega & Siaya (Kenya)	Busia and Kakamega (Kenya)	Bungoma, Busia, Kakamega and Vihiga (Kenya)	Rwanda
Agricultural Season	Messages concurrent with Short Rains 2015	Messages concurrent with Short Rains 2016 and Long Rains 2017	Messages concurrent with Long Rains 2017	Messages sent Short Rain 2016 for Long Rain 2017 input use	Messages sent Short Rain 2017 for Long Rain 2018 input use	Messages sent June 2017 August-September 2017 input use (2018 A season)
Number of Messages	20 total (2 acidity/lime; 5 fertilizer)	24-28 total (8 acidity/lime; 9 fertilizer)	13 total (6 acidity/lime; 4 fertilizer)	6 total (6 acidity/lime; 0 fertilizer)	1-10 total (1-5 acidity/lime; 1-5 fertilizer)	1-4 total (1-4 acidity/lime; 0 fertilizer)
Sample Type	Farmers drawn from village census.	Former NGO and contract farming participants.	Clients of agrodealers.	OAF clients in Long Rains 2016.	OAF clients in Long Rains 2018.	OAF clients in 2017 (seasons A and B)
Sample Size	834	1,897	5,890	4,884	32,572	110,400
Experiment	**(1) E-extension**: Treatment group received twenty SMS messages with recommendations about farming: use of lime, fertilizer, management practices. Two messages related to soil acidity and use of agricultural lime	**(1) General**: Twenty-four messages focusing on lime and fertilizer, use and field management. Eight messages focused on acidity or lime. Lime was broadly recommended but no information on local acidity was provided. **(2) Specific**: Twenty-eight messages focusing on lime and fertilizer, use and field management. Farmers informed about local soil tests results informed on specific quantities and price of inputs.	**(1) SMS**: Eleven messages focusing on lime and fertilizer use. Messages based on local soil tests. Farmers recommended specific quantities of lime. Those in areas with pH > 5.5 not recommended lime. **(2) SMS + Call**: Messages as in (1) and received call by field officer explaining information. **(3) SMS + Call offer**: Messages as in (1) and received offer for (2) if they texted or flashed.	**(1) Broad**: Six identical messages informing farmers that their soil was acidic, lime could reduce it and increase yields. **(2) Detailed**: Six identical messages informing farmers degree of soil acidity, the amount of lime to use, cost and predicted yield increase. All messages recommended positive amounts of lime.	**(1) Lime only**: 1-5 identical messages encouraging farmers to buy lime, sub-treatments varied by messages framing, repetitions and frequency. **(2) Lime+CAN**: in addition to 1-5 about lime, farmers received 1-5 extra messages encouraging to buy Extra CAN.	**(1) Same message**: 1-5 identical messages encouraging farmers to buy lime. Sub-treatments varied by messages framing, repetitions and frequency. All farmers in the same OAF group received the same message. **(2) Different messages**: 1-5 identical messages encouraging farmers to buy lime. Sub-treatments varied by messages framing, repetitions and frequency. Farmers in the same OAF group received different versions of the message.
Data Collection	Survey & Coupon (end of SR 2015)	Coupon (SR 2016) Survey (end LR 2017)	Coupon (LR 2017) Survey (end LR 2017 - SR 2017)	Purchases (SR 2016) Survey (end LR 2017)	Purchases (SR 2017)	Purchases (June 2017)
Results	No effect on lime knowledge or use. No evidence of effects on adoption of other inputs or practices.	Increased knowledge about acidity and use of lime.	Increased knowledge about acidity and use of lime in rec. areas.	Increased knowledge about acidity and use of lime.	Increased use of lime and fertilizer.	Increased use of lime.

	Age (1)	Female (2)	Primary (3)	Land size (4)	Used Lime (5)	Used Fert (6)	Seasons (7)	Credit Size (8)	Obs
Panel A. KALRO									
Control	41.29	0.65	0.53	2.22	0.06	0.84	.	.	417
	(13.43)	(0.48)	(0.50)	(5.39)	(0.24)	(0.37)	.	.	
Treated	39.79	0.65	0.54	1.92	0.07	0.84	.	.	415
	(13.17)	(0.48)	(0.50)	(2.03)	(0.26)	(0.37)	.	.	
P-value (T-Test)	0.10	0.87	0.67	0.30	0.57	0.98	.	.	
Panel B. IPA/PAD1-K									
Control	46.25	0.37	0.60	2.00	0.12	0.86	.	.	632
	(12.23)	(0.48)	(0.49)	(2.37)	(0.33)	(0.35)	.	.	
Treated	45.80	0.37	0.64	2.00	0.12	0.87	.	.	1,265
	(11.12)	(0.48)	(0.48)	(5.63)	(0.33)	(0.34)	.	.	
P-value (T-Test)	0.42	0.88	0.19	0.99	0.85	0.70	.	.	
Panel C. IPA/PAD2-K									
Control	42.10	0.66	0.72	2.02	0.09	0.84	.	.	1,470
	(12.41)	(0.47)	(0.45)	(2.38)	(0.28)	(0.37)	.	.	
Treated	41.44	0.66	0.70	1.99	0.09	0.85	.	.	4,420
	(11.95)	(0.47)	(0.46)	(2.61)	(0.29)	(0.36)	.	.	
P-value (T-Test)	0.07	0.98	0.25	0.71	0.55	0.59	.	.	
Panel D. OAF1-K									
Control	.	0.63	.	0.50	.	0.95	1.51	.	1,559
	.	(0.48)	.	(0.29)	.	(0.22)	(0.71)	.	
Treated	.	0.64	.	0.49	.	0.95	1.51	.	3,325
	.	(0.48)	.	(0.28)	.	(0.22)	(0.73)	.	
P-value (T-Test)	.	0.37	.	0.92	.	0.64	0.73	.	
Panel E. OAF2-K									
Control	48.40	0.69	.	0.51	.	0.93	2.23	9,504	8,142
	(13.56)	(0.46)	.	(0.33)	.	(0.26)	(1.52)	(4,497)	
Treated	48.33	0.69	.	0.52	.	0.93	2.23	9,498	24,430
	(13.52)	(0.46)	.	(0.33)	.	(0.25)	(1.52)	(4,445)	
P-value (T-Test)	0.70	0.57	.	0.49	.	0.13	1.00	0.91	
Panel F. OAF3-R									
Control	.	.	.	.	0.06	0.95	2.01	12,685	46,593
	.	.	.	.	(0.24)	(0.22)	(1.63)	(11,213)	
Treated	.	.	.	.	0.06	0.95	2.01	12,511	63,807
	.	.	.	.	(0.24)	(0.21)	(1.62)	(11,062)	
P-value (T-Test)	.	.	.	.	0.35	0.35	0.87	0.01	

Table 2: Summary Statistics & Balance[50]

[50] Standard deviations in parenthesis. Age refers to participants' age, in years, Female to the fraction of the sample that was female, Primary, to whether respondent completed primary school or more. Land size, to the self-reported size of farmers' land in acres, except that for 1AF1-K and 1AF2-K samples, where it corresponds to the size of the land for which farmers purchased 1AF inputs in the previous long rains season. Used Lime, on whether farmers report ever using lime in the past, except for the 1AF3-R sample, where it indicates whether the farmers purchased lime from 1AF in the previous year. Used Fertilizer, to whether they report using any chemical fertilizers (DAP, NPK, or CAN) in the previous agricultural season, except for the 1AF samples, where it indicates whether the farmers purchased fertilizer from 1AF in the previous year (DAP and CAN in Kenya; DAP, NPK, or urea in Rwanda). Seasons indicates the number of seasons the farmer was enrolled in the 1AF program, and Credit Size the amount of credit (in local currency) obtained from 1AF in the previous season: long rains 2017 for the 1AF2-K sample and 2017 season A for the 1AF3-R sample.

		Random Effects				Fixed Effects			
	N	Effect	95% CI		P-value	Effect	95% CI		P-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Odds Ratios									
Heard Lime	4	1.21	0.93	1.57	0.15	1.32	1.15	1.50	0.00
Knowledge Acidity	4	1.57	1.40	1.75	0.00	1.57	1.40	1.75	0.00
Lime Recommendation (survey, first season)	4	1.58	1.35	1.83	0.00	1.58	1.37	1.83	0.00
Lime Recommendation (admin, first season)	5	1.19	1.12	1.26	0.00	1.17	1.12	1.22	0.00
Lime Recommendation (admin, second season)	4	1.07	0.98	1.17	0.13	1.07	0.98	1.17	0.13
Fertilizer Recommendation (survey)	4	1.02	0.86	1.22	0.80	1.04	0.90	1.20	0.62
Fertilizer Recommendation (admin)	4	1.31	1.19	1.45	0.00	1.31	1.19	1.45	0.00
Standard Deviations									
Other Inputs (Index)	6	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02
Kg									
Kg Lime (admin)	6	1.18	0.14	2.22	0.03	1.18	0.14	2.22	0.03

Table 3: Meta-analysis Results[51]

[51] Meta-analysis results for each outcome reported in the rows. Column (1)-(4) reports results from a random-effects model; Column (5)-(8) reports results from a fixed-effects model. The coefficient represents the estimated summarized effects across studies, measured in odds ratios (except for the 'Other Input' variable which is a index variable and is measured in standard deviations.

	LPM		Odds ratios	
	Heard Lime	Knows Lime Use	Heard Lime	Knows Lime Use
	(1)	(2)	(3)	(4)
Panel A. KALRO				
Treated	-0.004	0.023	0.968	1.151
	(0.032)	(0.024)	(0.170)	(0.279)
Mean Control	0.58	0.16	0.58	0.16
Observations	773	773	773	773
Panel B. IPA/PAD1-K				
Treated	0.034	0.086***	1.257	1.598***
	(0.022)	(0.025)	(0.191)	(0.209)
Mean Control	0.78	0.33	0.77	0.33
Observations	1471	1471	1435	1471
Panel C. IPA/PAD2-K				
Treated	0.054***	0.099***	1.601***	1.591***
	(0.012)	(0.016)	(0.153)	(0.117)
Mean Control	0.81	0.43	0.81	0.43
Observations	4822	4822	4638	4771
Panel D. OAF1-K				
Treated	0.001	0.100***	0.997	1.629***
	(0.025)	(0.030)	(0.174)	(0.237)
Mean Control	0.80	0.32	0.80	0.32
Observations	1087	1087	1087	1087

Table 4: Awareness and Knowledge about Lime[52]

[52] Heard Lime is a dummy variable reporting whether farmers had heard about agricultural lime before. Knows Lime Use is coded as one if farmer mentions lime a strategy to deal with or reduce soil acidity. All regressions include controls. Columns (1) and (2) report marginal effects estimated using OLS, columns (3) and (4) report odds rations estimated using Logit * $p < .10$, ** $p < .05$, *** $p < .01$.

	Lime 1st season			Lime 2nd season		
	Survey	Admin (all)	Admin (enrol)	Survey	Admin (all)	Admin (enrol)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. KALRO						
Treated	-0.003				-0.006	
	(0.020)				(0.022)	
Mean Control	0.10				0.11	
Observations	773				773	
Panel B. IPA/PAD1-K						
Treated	0.041**	0.019		0.051***	0.006	
	(0.017)	(0.017)		(0.019)	(0.010)	
Mean Control	0.22	0.24		0.15	0.11	
Observations	1471	1897		1471	1897	
Panel C. IPA/PAD2-K						
Treated	0.075***	0.030***		0.010*		
	(0.013)	(0.009)		(0.006)		
Mean Control	0.31	0.30		0.22		
Observations	4822	5890		2566		
Panel D. OAF1-K						
Treated	0.041*	0.030***	0.054***		-0.004	-0.013
	(0.021)	(0.009)	(0.014)		(0.007)	(0.016)
Mean Control	0.12	0.10	0.17		0.06	0.15
Observations	1087	4884	2931		4884	1986
Panel E. OAF2-K						
Treated		0.025***	0.031***			
		(0.005)	(0.006)			
Mean Control		0.32	0.42			
Observations		32572	24825			
Panel F. OAF3-R						
Treated		0.006***	0.009***		0.002*	0.003*
		(0.001)	(0.002)		(0.001)	(0.002)
Mean Control		0.04	0.07		0.02	0.04
Observations		110400	67142		110400	54100

Table 5: Followed Lime Recommendations (LPM)[53]

[53] This table reports the effect of each program on whether farmers followed the lime recommendations expressed in terms of odds ratios. Columns (1)-(3) show results for the first growing season after each program was launched. Columns (4)-(6) show results for the second season. Columns (1) and (4) report survey result. Column (2) and (5) shows results for the administrative data (lime purchases or coupon redemption) for the entire sample of farmers participating int the experiment. Columns (3) and (6) show results for the administrative data for the subset of 1AF farmers registered in the program in that season. In Panels A and D-F the dependent variable takes value one if the farmer used or acquired agricultural lime. In Panels B and C, the dependent variable takes the value one if farmer used lime in an area where it was recommended, or did not use lime in an area where it was not recommended. All regressions include controls (but not location FEs). Robust standard errors shown in parenthesis. In panel F standard errors are clustered at the 1AF group level. *p < .10, ** p < .05, *** p < .01.

	Lime 1st season			Lime 2nd season		
	Survey	Admin (all)	Admin (enrol)	Survey	Admin (all)	Admin (enrol)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. KALRO						
Treated	0.963				0.904	
	(0.281)				(0.238)	
Mean Control	0.14				0.13	
Observations	561				664	
Panel B. IPA/PAD1-K						
Treated	1.617^{**}	1.164		1.570^{***}	1.121	
	(0.322)	(0.173)		(0.273)	(0.273)	
Mean Control	0.22	0.25		0.16	0.07	
Observations	1378	1854		1409	1531	
Panel C. IPA/PAD2-K						
treated	1.653^{***}	1.379^{***}		1.374		
	(0.150)	(0.145)		(0.267)		
Mean Control	0.31	0.28		0.19		
Observations	4641	5476		1745		
Panel D. OAF1-K						
Treated	1.563^{**}	1.394^{***}	1.496^{***}		0.925	0.888
	(0.328)	(0.145)	(0.165)		(0.129)	(0.128)
Mean Control	0.12	0.10	0.17		0.06	0.15
Observations	1087	4884	2931		4884	1986
Panel E. OAF2-K						
Treated		1.153^{***}	1.201^{***}			
		(0.035)	(0.043)			
Mean Control		0.32	0.42			
Observations		32572	24623			
Panel F. OAF3-R						
Treated		1.161^{***}	1.162^{***}		1.089^{*}	1.084^{*}
		(0.039)	(0.040)		(0.051)	(0.052)
Mean Control		0.05	0.09		0.03	0.06
Observations		87928	55068		75682	37970

Table 6: Followed Lime Recommendations (Odds ratios)[54]

[54] This table reports the effect of each program on whether farmers followed the lime recommendations measured in offs ratios (estimated through a logistic regression). Columns (1)-(3) show results for the first growing season after each program was launched. Columns (4)-(6) show results for the second season. Columns (1) and (4) report survey result. Column (2) and (5) shows results for the administrative data (lime purchases or coupon redemption) for the entire sample of farmers participating int the experiment. Columns (3) and (6) show results for the administrative data for the subset of 1AF farmers registered in the program in that season. In Panels A and D-F the dependent variable takes value one if the farmer used or acquired agricultural lime. In Panels B and C, the dependent variable takes the value one if farmer used lime in an area where it was recommended, or did not use lime in an area where it was not recommended. All regressions include controls (in panel C, column 4 location and enumerator FEs have been removed to ensure convergence). Robust standard errors shown in parenthesis. In panel F standard errors are clustered at the 1AF group level. $^{*}p < .10$, $^{**} p < .05$, $^{***} p < .01$.

	LPM				Odds ratios			
	1st season		2nd season		1st season		2nd season	
	Survey	Admin	Survey	Admin	Survey	Admin	Survey	Admin
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. KALRO								
Treated	-0.029			0.030	0.824			1.151
	(0.029)			(0.035)	(0.158)			(0.184)
Mean Control	0.81			0.41	0.81			0.41
Observations	773			773	773			773
Panel B. IPA/PAD1-K								
Treated	-0.004	0.011	-0.006		0.981	1.695	0.945	
	(0.026)	(0.007)	(0.019)		(0.121)	(0.617)	(0.161)	
Mean Control	0.68	0.02	0.87		0.68	0.03	0.86	
Observations	1471	1897	1471		1471	1278	1363	
Panel C. IPA/PAD2-K								
Treated	0.016	0.005	0.021		1.160	1.244	1.101	
	(0.011)	(0.005)	(0.022)		(0.122)	(0.256)	(0.120)	
Mean Control	0.86	0.02	0.68		0.86	0.04	0.67	
Observations	4822	5890	2566		4598	3471	2453	
Panel D. OAF2-K								
Lime only		0.009**				1.102**		
		(0.004)				(0.048)		
Lime+CAN		0.030***				1.345***		
		(0.006)				(0.078)		
Mean Control		0.14				0.14		
Observations		32572				32572		
p-value Lime only=Lime+CAN		0.000				0.000		

Table 7: Use of Recommended Fertilizers[55]

[55] This table reports the effect of each program on use chemical fertilizers. In panel A, the dependent variable takes value one if the farmer used any recommended chemical fertilizer (DAP, NPK, CAN, or Mavuno). In paneld B and C, the dependent variable in columns (1), (3), (5), and (7) indicates whether the farmer used any recommended fertilizer (DAP or urea), while the dependent variable in columns (2) and (6) indicate whether the used the electronic coupon to purchase the recommended topdressing fertilizer(urea). In panel D, the dependent variable indicates whether the farmer purchased "Extra CAN" from 1AF, since only a subset of treated farmers were also recommended Extra CAN, we show the results for both treatments. The outcomes reported in odd columns are measured using using survey data, while the outcomes reported in even columns are measured administrative data from coupon redemption or purchases from 1AF. All regressions include controls. Robust standard errors in parentheses. Columns (1)-(4) report marginal effects measured using OLS, columns (5)-(8) report odds ratios measured using Logit. * p < .10, ** p < .05, *** p < .01.

	Other Input Use (Index)
	(1)
Panel A. KALRO	
Treated	0.009
	(0.028)
Observations	773
Panel B. IPA/PAD1-K	
Treated	-0.021
	(0.029)
Observations	1471
Panel C. IPA/PAD2-K	
Treated	0.002
	(0.014)
Observations	4822
Panel D. OAF1-K	
Treated	0.012
	(0.011)
Observations	4884
Panel E. OAF2-K	
Treated	0.006
	(0.005)
Observations	32572
Panel F. OAF3-R	
Treated	0.005
	(0.004)
Observations	110400

Table 8: Use of Other Inputs and Management Practices[56]

[56] This table present results of indexes of knowledge and use of other inputs not listed. Each index is composed of different variables depending on the project. For a full list of variables see Appendix Table A12. The coefficients are average effect sizes. * p < .10, ** p < .05, *** p < .01.

[X]	LPM Female (1)	Primary School (2)	Farm Size (3)	Odds ratios Female (4)	Primary School (5)	Farm Size (6)
Panel A. KALRO						
Treated	-0.045	0.018	-0.012	0.616	1.336	0.823
	(0.041)	(0.027)	(0.029)	(0.255)	(0.560)	(0.273)
[X]	-0.042	0.064*	-0.001	0.663	2.367**	0.976
	(0.037)	(0.034)	(0.001)	(0.259)	(1.002)	(0.047)
[X] *Treated	0.060	-0.049	0.004	1.924	0.527	1.055
	(0.048)	(0.044)	(0.013)	(1.045)	(0.296)	(0.141)
Mean Control	0.11	0.11	0.11	0.13	0.13	0.13
Observations	773	773	773	664	664	664
Panel B. IPA/PAD1-K						
Treated	0.027	-0.001	0.004	1.222	0.993	1.049
	(0.022)	(0.026)	(0.020)	(0.226)	(0.249)	(0.194)
[X]	-0.013	-0.023	-0.008*	0.872	0.859	0.941
	(0.029)	(0.029)	(0.004)	(0.234)	(0.231)	(0.039)
[X] *Treated	-0.022	0.032	0.007*	0.870	1.277	1.054
	(0.034)	(0.034)	(0.004)	(0.270)	(0.403)	(0.056)
Mean Control	0.24	0.24	0.24	0.25	0.25	0.25
Observations	1897	1897	1897	1854	1854	1854
Panel C. IPA/PAD2-K						
Treated	0.012	0.034*	0.042***	1.110	1.355*	1.402***
	(0.018)	(0.019)	(0.014)	(0.168)	(0.237)	(0.164)
[X]	-0.032*	0.011	0.004	0.760*	1.114	1.027
	(0.019)	(0.020)	(0.004)	(0.124)	(0.206)	(0.026)
[X] *Treated	0.031	-0.002	-0.005	1.302	0.966	0.970
	(0.022)	(0.023)	(0.005)	(0.243)	(0.197)	(0.035)
Mean Control	0.30	0.30	0.30	0.30	0.30	0.30
Observations	5890	5890	5890	5732	5732	5732
Panel D. OAF1-K						
Treated	0.050		0.054	0.417		1.782
	(0.035)		(0.037)	(0.303)		(0.649)
[X]	-0.049		0.044	-0.613*		1.520
	(0.033)		(0.057)	(0.353)		(0.768)
[X] *Treated	-0.009		-0.020	0.109		0.819
	(0.042)		(0.067)	(0.406)		(0.474)
Mean Control	0.11		0.11	0.11		0.11
Observations	1151		1151	1151		1151
Panel E. OAF2-K						
Treated	0.026***		0.016*	0.135***		1.119**
	(0.009)		(0.010)	(0.051)		(0.056)
[X]	0.080***		0.070***	0.391***		1.458***
	(0.010)		(0.014)	(0.053)		(0.102)
[X]*Treated	-0.002		0.017	-0.025		1.006
	(0.011)		(0.016)	(0.060)		(0.080)
Mean Control	0.32		0.32	0.32		0.32
Observations	32572		32572	32572		32572

Table 9: Heterogeneity in Lime Redemption or Purchases[57]

[57] This table shows results of heterogeneity analysis by sample. The dependent variable is whether the farmer followed the recommendations. We show results for gender, whether respondent completed primary school, and land size. The coefficients are average effect sizes. All regressions include controls (Panel E columns (4) and (6) do not include site FEs). Columns (1)-(3) report marginal effects measured using OLS, columns (4)-(6) report odds ratios measured using Logit. * $p < .10$, ** $p < .05$, *** $p < .01$.

2.10 Appendix A: Additional Tables

	LPM		Odd ratios	
	Survey	Enrolled	Survey	Enrolled
	(1)	(2)	(3)	(4)
Panel A. KALRO				
Treated	0.019		1.325	
	(0.018)		(0.358)	
Mean Control	0.919		0.919	
Observations	833		833	
Panel B. IPA/PAD1-K				
Treated	0.014		1.086	
	(0.020)		(0.126)	
Mean Control	0.79		0.79	
Observations	1897		1897	
Panel C. IPA/PAD2-K				
Treated	-0.002		0.985	
	(0.012)		(0.077)	
Mean Control	0.82		0.82	
Observations	5890		5890	
Panel D. OAF1-K				
Treated	-0.018	-0.002	0.904	0.991
	(0.013)	(0.015)	(0.066)	(0.062)
Mean Control	0.25	0.60	0.25	0.60
Observations	4884	4884	4884	4884
Panel E. OAF2-K				
Treated		0.002		1.009
		(0.005)		(0.030)
Mean Control		0.76		0.76
Observations		32572		32572
Panel F. OAF3-R				
Treated		0.008***		1.036***
		(0.003)		(0.013)
Mean Control		0.65		0.65
Observations		110400		110400

Table A1: Probability of Collecting Data by Sample and Treatment Group[58]

[58] The dependent variable in Panel A takes the value of one if the farmer completed the in person endline survey. In panels B and C the dependent variable indicates whether the farmer completed the phone-based endline survey. In panel D the dependent variable in columns (1) and (3) is a dummy variable indicating whether the farmer completed a phone-based survey (conducted with 30% of original sample). In panels D-F, columns (2) and (4) the dependent variable indicates whether the farmer enrolled in the 1AF program (i.e. placed an input order). Columns (1) and (2) report marginal effects estimated using OLS, columns (3) and (4) report odds ratios estimated using Logit. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Control	Treated	(1) vs. (2)
	(1)	(2)	(3)
Age	41.29	39.79	1.50
	(0.66)	(0.65)	(0.92)
Female	0.65	0.65	-0.01
	(0.02)	(0.02)	(0.03)
Primary school	0.53	0.54	-0.01
	(0.02)	(0.02)	(0.03)
Secondary school	0.03	0.04	-0.01
	(0.01)	(0.01)	(0.01)
Footwear	0.61	0.56	0.05
	(0.02)	(0.02)	(0.03)
Mumias	0.56	0.57	-0.01
	(0.02)	(0.02)	(0.03)
Acres (owned and rented)	2.22	1.92	0.29
	(0.26)	(0.10)	(0.28)
Reads Swahili	0.91	0.91	0.00
	(0.01)	(0.01)	(0.02)
Had soil test	0.12	0.10	0.02
	(0.02)	(0.01)	(0.02)
Mentions Lime	0.03	0.05	-0.02
	(0.01)	(0.01)	(0.01)
Used Lime	0.06	0.07	-0.01
	(0.01)	(0.01)	(0.02)
Used fertilizer last LR season	0.84	0.84	0.00
	(0.02)	(0.02)	(0.03)
Grows legumes	0.81	0.83	-0.02
	(0.02)	(0.02)	(0.03)
Heard Lime	0.40	0.40	0.00
	(0.02)	(0.02)	(0.03)
Heard soil test	0.80	0.87	-0.07**
	(0.02)	(0.02)	(0.03)
Ever used DAP	0.94	0.94	0.00
	(0.01)	(0.01)	(0.02)
Ever used CAN	0.61	0.63	-0.02
	(0.02)	(0.02)	(0.03)
Ever used NPK	0.12	0.14	-0.02
	(0.02)	(0.02)	(0.02)
N	417	415	832
Joint F-Stat			1.06
P-value			0.386

Table A2: KALRO: Additional Summary Statistics & Balance[59]

[59] The table shows summary statistics and balance tests using covariate variables from a baseline survey. Columns (1)–(2) display the mean and standard error of each characteristic for each treatment group. Column (3) displays the differences across columns and corresponding standard error. Mumias denotes share of farmers from Kakamega county (Mumias area), Had soil test denotes ever having a soil test, Mentions Lime is a dummy variable with value one if respondent mentioned lime as a strategy to reduce soil acidity. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Control (1)	General (2)	Specific (3)	(1) vs. (2) (4)	(1) vs. (3) (5)	(2) vs. (3) (6)
Age	46.25	46.01	45.59	0.25	0.66	0.42
	(0.49)	(0.45)	(0.43)	(0.66)	(0.65)	(0.63)
Female	0.37	0.37	0.37	-0.01	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Primary school	0.60	0.61	0.66	-0.01	-0.05*	-0.04
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Secondary school	0.10	0.10	0.10	0.00	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
MSC sample	0.53	0.53	0.53	0.00	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
pH prediction	5.42	5.40	5.40	0.02	0.02	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Prefers English	0.30	0.27	0.30	0.03	0.00	-0.03
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Mentions lime	0.16	0.17	0.17	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Land size (acres)	2.00	1.86	2.14	0.14	-0.14	-0.28
	(0.09)	(0.08)	(0.31)	(0.12)	(0.32)	(0.32)
Has ever used lime	0.12	0.13	0.12	-0.01	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Used DAP last LR season	0.78	0.78	0.80	0.00	-0.02	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Used NPK last LR season	0.04	0.05	0.04	-0.01	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Used CAN last LR season	0.62	0.62	0.59	0.00	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Used Urea last LR season	0.18	0.18	0.18	0.00	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Used Mavuno last LR season	0.15	0.13	0.16	0.02	-0.01	-0.03
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Main network	0.95	0.94	0.94	0.01	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	632	633	632	1265	1264	1265
Joint F-Stat				0.57	0.78	0.75
P-value				0.909	0.714	0.746

Table A3: IPA/PAD1-K: Additional Summary Statistics & Balance[60]

[60] The table shows summary statistics and balance tests using covariate variables from a baseline survey. Columns (1)–(3) display the mean and standard error of each characteristic for each treatment group. Columns (4)-(6) display the difference across columns and the corresponding standard error. MSC Sample denotes share of farmers from the Mumias Sugar Company sample. pH prediction represents the median pH level measured in the farmer's catchment area. Mentions Lime is a dummy variable with value one if respondent mentioned lime as a strategy to reduce soil acidity. Fertilizer use variables refer to input use during the 2016 long rains season. Main network indicates whether the farmer's phone service provider is the main network in area. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Control	SMS	SMS+Call	SMS+Call Offer	(1) vs. (2)	(1) vs. (3)	(1) vs. (4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Age	42.10	41.40	41.48	41.44	0.70	0.61	0.66
	(0.32)	(0.31)	(0.32)	(0.31)	(0.45)	(0.46)	(0.45)
Female	0.66	0.66	0.66	0.66	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Primary school	0.72	0.70	0.69	0.71	0.01	0.02	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Secondary school	0.13	0.13	0.12	0.13	-0.01	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
pH prediction	5.37	5.37	5.37	5.37	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Prefers English	0.36	0.35	0.34	0.35	0.01	0.02	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Mentions lime	0.26	0.26	0.24	0.25	0.00	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Land size (acres)	2.02	1.85	2.09	2.03	0.17**	-0.07	-0.02
	(0.06)	(0.05)	(0.09)	(0.06)	(0.08)	(0.11)	(0.08)
Maize yield (t/ha)	1.51	1.46	1.37	1.49	0.05	0.15***	0.02
	(0.04)	(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)
Has ever used lime	0.09	0.09	0.09	0.10	-0.01	0.00	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
OAF Participant	0.35	0.34	0.35	0.37	0.00	-0.01	-0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Used CAN last LR season	0.64	0.62	0.65	0.62	0.02	0.00	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Used Urea last LR season	0.18	0.20	0.20	0.18	-0.02	-0.02	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Used Mavuno last LR season	0.08	0.08	0.07	0.09	0.00	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Recommended Lime	0.77	0.76	0.77	0.76	0.01	0.00	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
N	1470	1475	1473	1472	2945	2943	2942
Joint F-Stat					0.93	0.64	0.51
P-value					0.52	0.84	0.93

Table A4: IPA/PAD2-K: Additional Summary Statistics & Balance[61]

[61] The table shows summary statistics and balance tests using covariate variables from a baseline survey. Columns (1)–(4) display the mean and standard error of each characteristic for each treatment group. Columns (5)-(7) display the difference across columns and the corresponding standard error. pH prediction represents the median pH level measured in the farmer's ward used to provide lime recommendations. 1AF Participant is dummy variable indicating whether the farmer has ever been enrolled in the 1AF program. Mentions Lime is a dummy variable with value one if respondent mentioned lime as a strategy to reduce soil acidity. Fertilizer use variables refer to input use during the 2016 long rains season. Recommended lime indicates whether the farmer resided in a ward where lime was recommended. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Control	Broad	Detailed	(1) vs. (2)	(1) vs. (3)	(2) vs. (3)
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.63	0.63	0.65	0.00	-0.02	-0.02
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Group size	9.08	9.24	9.07	-0.16	0.01	0.17*
	(0.07)	(0.07)	(0.07)	(0.10)	(0.10)	(0.10)
OAF Seasons	1.51	1.50	1.52	0.00	-0.02	-0.02
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Maize inputs (acres)	0.50	0.49	0.50	0.01	-0.01	-0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Repayment Incentive (hoe)	0.06	0.07	0.08	-0.01	-0.02***	-0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
pH prediction	5.48	5.48	5.48	0.00	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Intercropped (acres)	0.01	0.01	0.01	0.00*	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Extra CAN purchased	0.04	0.05	0.05	-0.01**	-0.02***	0.00
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Onions	0.09	0.13	0.12	-0.04***	-0.03**	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Storage Bags	0.23	0.31	0.24	-0.07*	-0.01	0.06
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Solar Lamp	0.44	0.45	0.46	-0.01	-0.02	-0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Health Insurance	0.22	0.23	0.21	-0.01	0.01	0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	1559	1684	1641	3243	3200	3325
Joint F-Stat				1.77	1.96	1.84
P-value				0.048	0.024	0.037

Table A5: 1AF1-K: Additional Summary Statistics & Balance[62]

[62] The table shows summary statistics and balance tests using covariate variables from 1AF long rain 2016 administrative records (before the trial took place). Columns (1)-(3) display mean and standard errors of each variable, by treatment group. Columns (4)-(6) display the difference across columns and the corresponding standard error. Group size denotes number of farmers in the participant's 1AF group, 1AF seasons denotes the number of seasons of enrollment in the 1AF program, Maize inputs (acres) represents the size of maize inputs package purchased, Repayment Incentive is a dummy variable with value one if the farmer obtained a hoe as bonus for early repayment, pH prediction is the variable obtained using kriging interpolation that was used to produce detailed recommendations. Intercropped indicates the size of beans input package, for maize-beans intercropping, Extra CAN, Onions, Solar Lamps, and Health Insurance are dummy variables equal to one if the farmer purchased those additional products. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Control	Lime only	Lime + CAN	(1) vs. (2)	(1) vs. (3)	(2) vs. (3)
	(1)	(2)	(3)	(4)	(5)	(6)
Age (years)	48.40	48.30	48.44	0.10	-0.04	-0.14
	(0.15)	(0.10)	(0.20)	(0.18)	(0.25)	(0.22)
Female	0.69	0.69	0.68	0.00	0.01	0.01
	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Group size	9.87	9.82	9.92	0.04	-0.05	-0.10**
	(0.03)	(0.02)	(0.04)	(0.04)	(0.05)	(0.05)
OAF Seasons	2.23	2.23	2.22	0.00	0.01	0.01
	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)
Maize inputs (acres)	0.51	0.51	0.53	0.00	-0.01**	-0.01**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
pH prediction	5.33	5.33	5.33	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Intercropped (acres)	0.01	0.01	0.01	0.00*	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Extra CAN purchased	0.15	0.15	0.14	0.00	0.01	0.01
	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)
Onions	0.04	0.04	0.04	0.00	-0.01	-0.01*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Storage Bags	0.40	0.42	0.40	-0.02	0.00	0.02
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Solar Lamp	0.42	0.42	0.43	0.00	0.00	-0.01
	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Credit size	9504.07	9467.78	9616.95	36.29	-112.88	-149.17**
	(49.84)	(31.67)	(64.63)	(58.68)	(81.55)	(71.17)
N	8142	19558	4872	27700	13014	24430
Joint F-Stat				0.63	0.86	1.78
P-value				0.816	0.583	0.045

Table A6: 1AF2-K: Additional Summary Statistics & Balance[63]

[63] The table shows summary statistics and balance tests using covariate variables from 1AF long rain 2017 administrative records (before the trial took place). Columns (1)-(3) display mean and standard errors of each variable, by treatment group. Columns (4)-(6) display the difference across columns and the corresponding standard error. Group size denotes number of farmers in the participant's 1AF group, 1AF seasons denotes the number of seasons of enrollment in the 1AF program, Maize inputs (acres) represents the size of maize inputs package purchased, pH prediction was obtained using kriging interpolation. Intercropped indicates the size of beans input package, for maizebeans intercropping, Extra CAN, Onions, Solar Lamps, are dummy variables equal to one if the farmer purchased those additional products. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table A7: OAF3-R: Additional Summary Statistics & Balance

	G0: No SMS	G1: Same SMS	G2: Diff SMS	G3-Control	G3-Treated	(1) vs. (2)	(1) vs. (3)	(1) vs. (4)	(1) vs. (5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Group size	10.71	10.69	10.80	10.72	10.71	0.01	-0.09***	-0.01	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
OAF Seasons	2.01	2.02	2.01	2.01	2.01	-0.01	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Enrolled(A)	0.89	0.89	0.88	0.89	0.89	0.00	0.01**	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Enrolled(B)	0.78	0.78	0.80	0.79	0.79	0.00	-0.02***	-0.01***	-0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lime(A)	0.05	0.05	0.05	0.04	0.04	0.00	0.01**	0.01***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lime(B)	0.02	0.02	0.02	0.02	0.02	0.00**	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
DAP kg(A)	9.58	9.53	9.31	9.40	9.41	0.05	0.27**	0.18	0.17
	(0.09)	(0.09)	(0.09)	(0.07)	(0.07)	(0.13)	(0.13)	(0.12)	(0.12)
DAP kg(B)	5.81	5.87	6.14	5.96	6.01	-0.05	-0.33***	-0.15	-0.20**
	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)	(0.10)	(0.11)	(0.10)	(0.10)
Urea kg(A)	5.45	5.46	5.28	5.30	5.33	-0.02	0.16**	0.15**	0.12*
	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.08)	(0.08)	(0.07)	(0.07)
Urea kg(B)	1.92	1.95	2.05	1.96	1.96	-0.03	-0.13***	-0.04	-0.04
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.05)	(0.05)	(0.04)	(0.04)
Solar Lamp(A)	0.28	0.26	0.27	0.29	0.28	0.01***	0.00	-0.01**	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Credit(A)	12703.73	12504.15	12376.07	12672.36	12603.65	199.57*	327.66***	31.36	100.08
	(81.95)	(81.77)	(81.47)	(67.15)	(67.22)	(115.83)	(115.66)	(105.65)	(105.70)
Credit(B)	9807.56	9625.64	10101.65	10050.32	10065.20	181.92*	-294.09***	-242.76**	-257.65**
	(75.23)	(76.45)	(78.95)	(64.31)	(65.33)	(107.27)	(108.99)	(99.45)	(100.39)
N	19066	18254	18082	27527	27471	37320	37148	46593	46537
Joint F-Stat						1.46	3.48	2.31	2.31
P-value						0.12	0.00	0.00	0.00

Notes: The table shows summary statistics and balance tests using covariate variables from OAF 2017 main agricultural season administrative records (before the trial took place). Columns (1) - (5) display mean and standard errors of each variable, by treatment group. Columns (6)-(9) displays the difference across columns and the corresponding standard error. *Group size* denotes number of farmers in the participant's OAF group, *OAF seasons* denotes the number of seasons of enrollment in the OAF program, *Enrolled* indicates whether the farmer was enrolled in the OAF program in the 2017 main agricultural season, *Shared phone* is a dummy indicating whether the farmer shared phone number with other farmers in the database, *N farmers in group* indicates the number of farmers in the group with a phone number in the OAF database, *Lime* is a dummy indicating whether the farmer purchased lime. *DAP* and *Urea* indicates the quantities of those fertilizers purchased. *Solar Lamps* is a dummy variables equal to one if the farmer purchased any solar lamps. *Credit* reports the size of the OAF loan. All variables indicating inputs and credit refer to the 2017 main agricultural season and are coded as zero if the participant did not enroll in the OAF program for that season. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Kg Lime (survey)		Kg Lime (admin)	
	Lime Rec	Lime not Rec	Lime Rec	Lime not Rec
	(1)	(2)	(3)	(4)
Panel A. KALRO				
Treated			-1.787	
			(3.738)	
Mean Control			16.93	
Observations			773	
Panel B. IPA/PAD1-K				
Treated	0.872	1.008	0.119	1.379
	(0.538)	(0.999)	(0.618)	(1.233)
Mean Control	1.85	1.70	2.85	3.32
Observations	1197	274	1552	345
Panel C. IPA/PAD2-K				
Treated	2.896***	-0.692	0.966**	-1.495*
	(0.812)	(1.228)	(0.444)	(0.768)
Mean Control	6.71	5.97	3.52	3.56
Observations	3729	1093	4512	1378
Panel D. OAF1-K				
Treated	2.261*		3.209***	
	(1.300)		(0.793)	
Mean Control	5.18		5.82	
Observations	1087		4884	
Panel E. OAF2-K				
Treated			2.237***	
			(0.449)	
Mean Control			17.90	
Observations			32572	
Panel F. OAF3-R				
Treated			0.175**	
			(0.088)	
Mean Control			2.05	
Observations			110400	

Table A8: Amount of Agricultural Lime by Type of Recommendation[64]

[64] The table reports the effects of the programs on quantity of lime purchased, expressed in kgs. Columns (1) and (2) show results for the survey data, while columns (3) and (4) show the results for administrative data. For the 1AF and IPA/PAD samples, the quantity of lime measured through survey data was winsorized at 99th percentile to remove outliers and the farmers who did not remember the amount of lime used were assigned quantity used by the median lime user. All regressions include controls. Robust standard errors in parentheses. In panel F the standard errors are clustered at the 1AF group level. * $p < .10$, ** $p < .05$, *** $p < .01$.

	LPM				Odds ratios			
	Heard Lime	Knows Lime	Followed Lime Rec		Heard Lime	Knows Lime	Followed Lime Rec	
			Survey	Admin			Survey	Admin
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. IPA/PAD1-K								
General	0.037	0.061**	0.032	0.020	1.282	1.416**	1.455	0.173
	(0.025)	(0.029)	(0.020)	(0.020)	(0.226)	(0.213)	(0.337)	(0.173)
Specific	0.030	0.111***	0.050**	0.017	1.232	1.793***	1.785***	0.132
	(0.025)	(0.030)	(0.020)	(0.019)	(0.218)	(0.268)	(0.391)	(0.167)
Mean Control	0.78	0.33	0.22	0.24	0.77	0.33	0.22	0.25
Observations	1471	1471	1471	1897	1435	1471	1378	1854
p-value General=Specific	0.751	0.095	0.389	0.863	0.826	0.105	0.330	0.807
Panel B. IPA/PAD2-K								
SMS	0.042***	0.092***	0.067***	0.030***	1.414***	1.533***	1.569***	1.390***
	(0.015)	(0.020)	(0.016)	(0.012)	(0.168)	(0.137)	(0.168)	(0.172)
SMS + Call	0.070***	0.116***	0.095***	0.020*	1.881***	1.730***	1.870***	1.247*
	(0.014)	(0.019)	(0.017)	(0.012)	(0.231)	(0.156)	(0.205)	(0.162)
SMS + Call Offer	0.051***	0.089***	0.064***	0.039***	1.568***	1.520***	1.539***	1.507***
	(0.015)	(0.020)	(0.016)	(0.012)	(0.192)	(0.139)	(0.168)	(0.187)
Mean Control	0.81	0.43	0.31	0.30	0.81	0.43	0.31	0.28
Observations	4822	4822	4822	5890	4638	4771	4641	5476
p-value SMS=SMS+Call	0.041	0.206	0.109	0.369	0.026	0.178	0.092	0.373
p-value SMS=SMS+Call Offer	0.553	0.882	0.870	0.474	0.416	0.927	0.854	0.484
p-value SMS+Call=SMS+Call Offer	0.155	0.162	0.077	0.107	0.167	0.159	0.065	0.117
Panel C. OAF1-K								
main								
Broad	0.006	0.091**	0.038	0.026**	1.019	1.572***	1.529*	1.340**
	(0.029)	(0.035)	(0.025)	(0.011)	(0.206)	(0.264)	(0.370)	(0.158)
Detailed	-0.003	0.109***	0.045*	0.034***	0.975	1.689***	1.597**	1.451***
	(0.030)	(0.035)	(0.025)	(0.011)	(0.198)	(0.279)	(0.372)	(0.171)
Mean Control	0.80	0.32	0.12	0.10	0.80	0.32	0.12	0.10
Observations	1087	1087	1087	4884	1087	1087	1087	4884
p-value Broad=Detailed	0.764	0.631	0.799	0.484	0.833	0.661	0.846	0.474
Panel D. OAF2-K								
Lime only				0.023***				1.141***
				(0.005)				(0.036)
Lime+CAN				0.032***				1.202***
				(0.008)				(0.052)
Mean Control				0.32				0.32
Observations				32572				32572
p-value Lime only=Lime+CAN				0.174				0.178
Panel E. OAF3-R								
G1: Same SMS				0.007***				1.201***
				(0.002)				(0.075)
G2: Diff SMS				0.010***				1.283***
				(0.002)				(0.081)
G3-Control				0.004*				1.114*
				(0.002)				(0.063)
G3-Control				0.008***				1.236***
				(0.002)				(0.069)
Mean Control				0.04				0.05
Observations				110400				87928

Table A9: Knowledge and Adoption by Treatment[65]

[65] The table shows the effect of each of the main treatments on knowledge of lime and probability to follow recommendations. The dependent variable in column (1) is a dummy variable reporting whether farmers had heard about agricultural lime before. The dependent variable in column (2) is coded as one if farmer mentions lime a strategy to deal with or reduce soil acidity. The dependent variable in columns (3) and (4) indicates whether farmers followed lime recommendations. In panels A and B, it takes value one if the farmer used lime and lime was recommended or if farmer did not use lime and lime was not recommended, zero otherwise. In panels C-E takes value one if the farmer used lime, zero otherwise. All regressions include controls. Robust standard errors in parenthesis. In panel E the standard errors are clustered at the 1AF group level. Columns (1) - (4) report marginal effects estimated using OLS, columns (5) - (8) report odds ratios estimated using Logit. * $p < .10$, ** $p < .05$, *** $p < .01$.

	Purchased lime					
	LPM			Odds ratios		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. OAF2-K						
main						
N Lime SMS	0.006***			1.035***		
	(0.001)			(0.008)		
N Lime SMS$\geq$ 1		-0.003			0.983	
		(0.012)			(0.068)	
N Lime SMS $\geq$ 2		0.025**	0.025**		1.159**	1.157**
		(0.012)	(0.013)		(0.083)	(0.083)
N Lime SMS $\geq$ 3		0.004	0.004		1.023	1.023
		(0.008)	(0.008)		(0.046)	(0.045)
N Lime SMS $\geq$ 4		0.004	0.004		1.023	1.022
		(0.008)	(0.008)		(0.045)	(0.045)
N Lime SMS = 5		-0.005	-0.005		0.973	0.974
		(0.008)	(0.008)		(0.044)	(0.043)
Mean Control	0.32	0.32		0.32	0.32	
Observations	32572	32572	24430	32572	32572	24430
Includes Control Group	Yes	Yes	No	Yes	Yes	No
Panel B. OAF3-R						
N Lime SMS	0.002***			1.057***		
	(0.000)			(0.011)		
N Lime SMS$\geq$ 1		0.002			1.043	
		(0.002)			(0.051)	
N Lime SMS $\geq$ 2		0.005**	0.005**		1.131**	1.129**
		(0.002)	(0.002)		(0.063)	(0.063)
N Lime SMS $\geq$ 3		0.001	0.001		1.015	1.012
		(0.002)	(0.002)		(0.055)	(0.055)
N Lime SMS = 4		0.001	0.001		1.022	1.027
		(0.002)	(0.002)		(0.056)	(0.056)
Mean Control	0.04	0.04		0.05	0.05	
Observations	110400	110400	63807	87928	87928	47625
Includes Control Group	Yes	Yes	No	Yes	Yes	No

Table A10: Number of Messages[66]

[66] The table shows the effect number of messages on lime purchases. The dependent variable in column indicates whether farmers purchased lime from 1AF. All regressions include controls. Robust standard errors in parenthesis. In panel B the standard errors are clustered at the 1AF group level.Columns (1) - (4) report marginal effects estimated using OLS, columns (5) - (8) report odds ratios estimated using Logit. * p < .10, ** p < .05, *** p < .01.

	Purchased lime					
	LPM			Odds ratios		
	(1)	(2)	(3)	(5)	(6)	(7)
Panel A. OAF2-K						
Basic	0.017**			1.108**		
	(0.008)			(0.051)		
Yield Increase	0.034***	0.017*		1.217***	1.097*	
	(0.008)	(0.009)		(0.056)	(0.057)	
Experimentation (self)	0.027***	0.010		1.171***	1.056	
	(0.008)	(0.009)		(0.054)	(0.055)	
Experimentation (neighbors)	0.013*	-0.004		1.079	0.973	
	(0.008)	(0.009)		(0.050)	(0.051)	
Social Compasison	0.028***	0.010		1.174***	1.058	
	(0.008)	(0.009)		(0.054)	(0.056)	
Self-efficacy	0.028***	0.010		1.175***	1.059	
	(0.008)	(0.009)		(0.054)	(0.056)	
Family framed SMS			-0.016***			0.912***
			(0.005)			(0.028)
Mean Control	0.32			0.32		
Observations	32572	24430	24430	32572	24430	24430
Includes Control Group	Yes	No	No	Yes	No	No
Panel B. OAF3-R						
General promotion	0.008***			1.232***		
	(0.002)			(0.078)		
Specific + yield impact	0.008***	-0.001		1.229***	0.968	
	(0.002)	(0.003)		(0.078)	(0.072)	
Self-diagnosis	0.007***	-0.002		1.216***	0.969	
	(0.002)	(0.003)		(0.077)	(0.072)	
Soil test	0.005**	-0.004		1.151**	0.917	
	(0.002)	(0.003)		(0.075)	(0.069)	
How travertine works	0.006***	-0.003		1.190***	0.930	
	(0.002)	(0.003)		(0.077)	(0.071)	
Order immediately	0.008***	-0.001		1.229***	0.996	
	(0.003)	(0.003)		(0.081)	(0.074)	
Your cell is acidic + yield impact	0.006**	-0.003		1.172**	0.928	
	(0.002)	(0.003)		(0.075)	(0.069)	
SMS framed as gain			0.001			1.014
			(0.002)			(0.042)
Mean Control	0.04			0.05		
Observations	110400	63807	63807	87928	47625	47625
Includes Control Group	Yes	No	No	Yes	No	No

Table A11: Message Framing[67]

[67] The table shows the effect different messages framing of messages on lime purchases. The dependent variable in column indicates whether farmers purchased lime from 1AF. All regressions include controls. Robust standard errors in parenthesis. In panel B the standard errors are clustered at the 1AF group level. * p < .10, ** p < .05, *** p < .01.

Sample	Use
KALRO	Used: composting, manure, intercropping, crop rotation, rhizobia, improved seeds, improved legumes, striga control, storage bags, soil acidity testing
IPA/PAD1-K	Used: DAP, NPK, Mavuno
IPA/PAD2-K	Used: DAP, NPK, Mavuno, hybrid seeds, pesticides
OAF1-K	Purchased: actellic, compost, cookstove, extra CAN, drying sheets, storage bags, onion seeds, reusable pads, storage bags, solar lamp, machete, hoe
OAF2-K	Purchased: actellic, compost, cookstove, drying sheets, storage bags, red onions, reusable pads, solar lamp
OAF3-R	Purchased: DAP, NPK, urea , storage bags

Table A12: Variables Used to Construct Indexes in Table 8

2.11 Appendix B: Additional Program and Evaluation Details

2.11.1 B.1 KALRO

KALRO's e-extension program consisted in 20 SMS messages send in the period corresponding to the 2015 short rains season: June-November 2015. The messages contained recommendations on several farming practices including land preparation, use of fertilizer and agricultural lime, inter-cropping, and storing the maize. While the fertilizer messages recommended specific quantities and application method: one soda bottle top of either DAP or NPK per planting hole at planting and at one bottletop of either CAN or Mavuno per plan at topdressing, the lime messages did not specify what the recommended rate was. In fact, farmers were encouraged to test their soil and, if they found that the pH level was below 5.5, seek further advice on lime application. The first set of messages were in English. Mid-intervention the messages were switched to Swahili.[68] We list all messages sent by KALRO below:

- We at KALRO- Kakamega shall be sending you 20 SMS tips on how to increase your maize and legume (beans, groundnuts, soybeans) yield
- Keep all the records of your farming activities including inputs and outputs to help you know whether your farming is profitable
- Test your soil after every 4 years. Enquiries: KALRO Tel:[phone] or Soil Cares Ltd: [phone]
- If soil is acidic (pH less than 5.5), apply recommended rate of agricultural lime at least 30 days before planting. Enquiries: Tel.[phone]
- Construct raised bands and trenches to control soil erosion, reduce nutrient loss and keep rain water in the soil
- Add and/or leave all organic matter (manure, crop/weed residues and compost) to your field. Do not burn your fields. Burning destroys useful micro-organisms.
- Prepare land early, at least one plough and one harrow, ready for planting before onset of rains
- Plant before or at the onset of rains. Plant on well drained, fertile soils
- Use certified maize and legume seed recommended for your area, bought from an approved agro-dealer. Use 10 kg maize seed and 40kg of legume seed per acre. Enquiries: [phone]
- Maize and legumes planted in rows are easier to weed apply fertilizer. You may plant maize alone/pure or together with legumes as follows:
- For pure maize make rows 2.5 feet (75cm) apart and holes 1 foot (30cm) apart along the row. Place 2 and 1 maize seeds in alternate holes.
- For maize and legume intercrop, plant maize as for pure stand and one row of legume (beans, soybean or groundnut) between two maize rows at spacing of 10cm from one hole to another.
- For better maize and legume harvests, inoculate legumes, rotate or intercrop, use fertilizer and manage your crop and soils appropriately.

[68] While 75% of farmers report speaking English at baseline, there is a risk that some farmers might have not understood the initial messages. We do not find heterogeneous treatment effects by language spoken.

- Use fertilizer to increases yields. Apply 1 heaped Fanta top of NPK or DAP in each hole for maize, cover with little soil, add seed and cover seed with soil. Fertilizer MUST not touch the seed
- Weeds compete with your crops for nutrients and so reduce yields. Keep fields free of weeds and pests. Thin maize seedlings to 1 plant per hole as you weed.
- Topdress your maize with a level Fanta bottle top of CAN or Mavuno top dress fertilizers 6 weeks after planting. Apply around each plant-5cm away and cover with soil. Apply when soil is moist.
- Harvest as soon as the crops are mature. For maize look for the black eye; for legumes when 90-100% of pods are brown. In late harvests, termites, rodents, insects, diseases birds reduce yield.
- Remove husk from maize cobs in the field to avoid transporting weevils from the field to the store. The husks will improve the organic matter in the soil.
- Dry your harvest in open sun, but protect it from rain. Thresh/shell and re-dry to moisture content of 11-12%.

In addition to the e-extension intervention, KALRO also evaluated the impact of the Farmer Field Days (FFDs), a one-day educational events where farmers could observe results from demonstration plots (hosted by a farmer in the area) and learn about various technologies and management practices from extension workers. As part of a broader program to increase smallholder farmer productivity, KALRO organized several FFDs in western Kenya. All demonstration plots organized by KALRO showcased different types of fertilizers (including DAP, Mavuno, NPK and CAN), intercropping of maize with legumes and agricultural lime. FFDs were held on pre-specified days and they generally lasted the entire morning. Host farmers were selected by KALRO at the onset of the planting season and they received all the inputs and technical support to set up the demonstration plots. To promote ownership of the demonstrations, KALRO requested farmers to provide most of the labor to maintain the plots. Therefore, these plots were a fair representation of how the inputs and practices would work outside of controlled environments, such as agricultural experiment stations. One of the key messages highlighted by extension workers during FFDs was the recommendation to conduct soil analyses and apply lime if the soil was acidic (pH less than 5.5), intercrop their maize with legumes and use chemical fertilizers, in particular CAN, DAP and Mavuno.

2.11.2 B.2 IPA/PAD1-K

The first program implemented by IPA and PAD consisted in a series of 24-28 messages sent during the 2016 short rains season: August-December 2016. The message provided recommendations on several aspects of maize farming: land preparation, fertilizer and lime application, weeding, and pest and disease management, and harvesting practices. Two versions of the service were tested. The first, denoted as "General" provided blanket recommendations on maize farming in western Kenya. The second, denoted as "Specific", included customized recommendations for planting fertilizer and agricultural lime based on local soil characteristics. Farmers participating in this programs were recruited from two sources: a database of farmers who had previously participated in IPA activities (IPA farmers), and administrative record of Mumias Sugar Company, a company that works with contract farmers in the area (MSC farmers). In order to construct customized recommendations for the specific messages, farmers were linked to a local landmark that could then be matched with soil data. IPA farmers were matched to the primary school in which they were originally (usually the closest one to their farm) and provided recommendation based on median soil characteristics (exchangeable acidity and phosphorous) obtained from soil tests performed in the 2 km area around the school.[69] The soil data were collected for previous projects by IPA (Fabregas et al., 2017b) and analyzed by the Kenya Agricultural Research Institute (KARI) using wet chemistry in 2011 and 2014. MSC farmers were matched to their "field", a set of plots cultivated by multiple farmers and aggregated by the company for organizing their activity, including soil testing. The recommendations provided to them were based on median soil characteristics (pH and phosphorus) of the sample collected from that field and analyzed by MSC in the period 2009-2016. Since the topdressing fertilizer recommendation were not specific to the farmers' catchment area, but based on the quantity of nitrogen required to achieve a certain expected yield, specific application rates were provided to all treated farmers. Messages were sent either in English or in Swahili, depending on farmers' preferences indicated during the baseline phone survey. We report all the messages below: [G] indicates that the message was received by the General treatment group, and [S] denotes it was received by the Specific treatment group.

- [G/S]: Welcome to PAD's SMS information service. We will give you tips on agricultural inputs to apply on 1/8 of an acre so you can experiment during this short rains season. Receiving SMS messages is free.
- [G]: High soil acidity levels reduce nutrients available to plants, such as phosphorus, which causes symptoms of stunted growth and purple colouration of maize.
- [S]: Previous soil tests of shambas around [landmark] showed [degree] soil acidity levels. High acidity levels reduce nutrients available to plants, such as phosphorus, which causes symptoms of stunted growth and purple colouration of maize.
- [G]: Lime reduces soil acidity and makes nutrients such as phosphorus available for your maize.
- [S]: Based on soil tests of shambas around [landmark], we recommend you buy [quantity] kg of lime, [quantity] kg of DAP, and 6 kg of urea for microdosing 1/8 acre of your maize. Lime reduces soil acidity and makes phosphorus available for your maize.
- [S]: We would like you to try our recommendations in 1/8 of an acre. To measure 1/8 of an acre, walk around your farm and draw a square with each side 33 steps long. Walk normally, don't make long strides. If you land is a rectangle, the sum of 2 sides should measure in total 66 steps. Start from a corner, walk along the short side, count your steps until you reach the end. Turn around and keep walking along the long side until you finish counting 66 steps.
- [S]: When planting this season try adding a layer of lime [quantity] bottletop, then cover with soil and add a second layer of DAP ([quantity] bottletop) per hole on 1/8 acre to correct soil acidity and make more nutrients available for your plants. Apply 1 bottletop of urea per hole at top dressing.
- [G]: Use a ruler or measured rope to plant maize in rows using correct spacing of 75 cm x 25 cm. This offers maximum yield while limiting competition for nutrients, light and water.

[69] This is a context in which there are no addresses and a lot of variation on how village names are reported. Therefore it was difficult for farmers to report their exact location. Primary schools are often used as landmarks.

- [S]: Use a ruler or measured rope to plant maize in rows using correct spacing of 75 cm x 25 cm. This offers maximum yield while limiting competition for nutrients, light, and water. You should be able to fit 2580 planting holes in 1/8 of an acre. Use sisal twine to encircle this area so you can compare the results at harvest.
- [S]: Have you bought lime and DAP yet? If not, buy a total of [quantity] kg of lime and use with [quantity] kg DAP for microdosing on 1/8 of your acre. DAP is the most cost efficient source of phosphorous. When lime is combined with DAP, it reduces soil acidity and makes nutrients available for your maize.
- [G]: Calcium lime is safer for your health and the plant. This lime could be either brown or grey.
- [S]: [agrovet] will be stocked with lime (calcium lime) and DAP during this short rain season. This lime is brown and it is safer for your health and the plant. It is also heavier than the white lime so you only need to apply [quantity] bottletop per plant. The price of lime today is Ksh 7 per kg. The price of DAP today is Ksh [price] per kg.
- [G/S]: Plant maize seed when there is enough moisture after 2-3 rains, to enable absorption of water by seed and fertilizer. Delayed planting leads to reduced yields. To stop receiving these SMS messages reply ”STOP”.
- [G/S]: Plant two maize seeds per hole to ensure one survives. Do not use broken or damaged seeds because they will not germinate. Use certified seeds, they grow faster and are high yielding.
- [G]: Are you ready to plant your maize? We recommend you apply both lime and fertilizer in micro-doses at planting. 5 weeks later we recommend you apply top dressing fertilizer in micro-doses
- [S]: Do you know the 5 Golden Rules for successful micro-dosing? Based on soil tests performed around [landmark], we recommend you to: Apply [quantity] bottletop of lime and cover with soil and then add [quantity] bottletop of DAP. Cover with 2 inches of soil. Use 2 seeds per planting hole.Cover the seeds with 2 inches of loose soil. Apply 1 bottletop of urea as top dressing fertilizer 5 weeks later when the plant is knee high.
- [G/S]: Remember, lime should only be used during planting and not at top dressing. Lime is not a fertilizer and could burn the plant if applied at top dressing.
- [G/S]: At planting, if you are applying lime in micro-doses, remember to cover it with soil before applying fertilizer and planting seeds. Lime should not be in direct contact with the seeds as it may burn them. When you apply lime, wear protective clothing such as long sleeves and gloves. Cover your mouth and nose with a scarf and wear goggles.
- [G/S]: Gap your maize immediately after emergence. Gapping is done by replanting maize seeds in places that have not germinated. This gives you optimum plant population that leads to optimum yields.
- [G/S]: During first weeding, thin to one maize plant per hole. You should remove striga immediately to reduce competition for nutrients and water, and to prevent stunted growth!
- [G]: Have you already planted your maize this season? If not, we recommend applying lime at planting. Lime reduces soil acidity and makes nutrients such as phosphorus available for your maize.
- [S]: Have you already planted your maize this season? If not, we recommend applying lime at planting. We recommend you apply [quantity] bottletop per planting hole. Buy

[quantity] kg of lime to experiment on 1/8 of an acre. Lime reduces soil acidity and makes nutrients such as phosphorus available for your maize.

- [G]: If you applied lime on your maize at planting, we recommend using urea at top dressing because it is a less expensive source of nitrogen.
- [S]: If you applied lime on your maize at planting, we recommend using urea for top dressing because it is a less expensive source of nitrogen. Buy 6 kg of urea for use on 1/8 of an acre.
- [S]: [agrovet] will be stocked with urea during this short rain season. The price of urea is Ksh [agrovet] per kg.
- [G]: When the maize reaches knee high (5 weeks after planting), apply top dressing fertilizer.
- [S]: When the maize reaches knee high(5 weeks after planting), based on soil tests around [landmark], we recommend you apply 1/2 bottletop of urea per plant, making a 15 cm circle around the maize plant.
- [G/S]:Conduct second weeding 6 or 7 weeks after planting. Uproot all striga before it produces seeds because it reduces maize yields if not removed
- [G/S]: We invite you to participate in an SMS poll to help you recognize potential maize diseases and provide advice. Reply OK to start. Messages are free.
    - Do you see straight lines of holes on newly formed maize leaves? [if yes] This could be stalk borers. Apply insecticide e.g. bulldock or tremor, into the funnel or spay the maize plant with pentagon at top dressing.We hope this information was helpful. We will be sending another poll question tomorrow. Thank you! [if no] This is good news! Thank you for answering our question. We will send another question tomorrow.
    - Do you notice yellow or white streaks or discoloration on the leaves of your stunted maize plants? [if yes] It could be Maize Streak Virus. Eradicate grass weeds and use malathion or dimethoate to control as soon as possible. We hope this information was helpful. We will be sending another poll question tomorrow. Thank you! [if no] This is good news! Thank you for answering our question. We will send another question tomorrow.
    - Do you see striga weed in your maize plot? Striga has thin leaves and pink or purple flowers and attaches onto the maize roots. [if yes] Uproot all striga that has emerged. Striga competes with your maize for nutrients, water, and light and leads to reduced maize yields. We hope this information was helpful. We will be 58 sending another poll question tomorrow. Thank you! [if no] This is good news! Thank you for answering our question. We will send another question tomorrow.
    - Do you see ants that cut maize stalks and feed on fallen maize cobs? [if yes] It could be termites. Dig out all anthills around your maize farm and ensure that you destroy the queen. Alternatively, you can dig a deep hole at the center of the anthill and use insecticide to kill the ants. We hope this information was helpful. This is the last poll question. We will NOT send another question tomorrow. Thank you for your participation! [if no] This is good news! This is the last poll question. We will NOT send another question tomorrow. Thank you for your participation!

- [G/S]: WEEDING REMINDER! Conduct second weeding 6 or 7 weeks after planting. Weeds compete with your maize for nutrients, water, and light, which reduces yields.
- [G]: Have you already applied top dressing fertilizer on your maize? If not, we recommend using urea at top dressing because it is a less expensive source of nitrogen.
- [S]:Have you already applied top dressing fertilizer on your maize? If not, we recommend using urea at top dressing because it is a less expensive source of nitrogen. Buy 6 kg of urea for use on 1/8 of an acre and apply 1/2 bottletop of urea per plant. Apply urea when there is enough moisture in the soil to avoid loss through evaporation.
- [G/S]: Harvest maize at physiological maturity when cobs droop and leaves dry. Dry maize in the sun even after shelling to avoid mold and attack by weevils. Maize grain must remain dry and clean during storage to avoid reduction in quantity and quality.
- [G/S]: We hope you enjoyed these messages from Precision Agriculture for Development. Our team will follow up with a phone call in the coming weeks to hear more about how your planting season went.

During the 2017 long rains season, all treated farmers received 5 identical SMS messages about agricultural lime. These messages were based on the local recommendations constructed for the IPA/PAD2-K program. We report the text these messages below:

- [If pH≤5.5]: The soil in your area is [level] acidic. To avoid low yields, treat now. Apply [quantity] bottletop of lime per planting hole. [quantity] lime per 1/4 acre.
- [If pH>5.5]: The soil in your area is slightly acidic. According to our analysis, farms in your area do not need lime.

2.11.3 IPA/PAD2-K

The second program implemented by IPA and PAD consisted in 3 message about planting inputs for maize farmers (lime and fertilizer), repeated twice, plus 2 additional messages on topdressing fertilizer, also repeated twice. Planting recommendations were based on local soil data: ward level median level of pH and phosphorous, and target yield of 2 t/ha, while topdressing recommendations were only based on target yield.[70]

Recommendations for lime and DAP were provided based on median soil characteristics in the farmers' ward.[71] The soil data used to generate these recommendations was obtained by pooling data collected by 4 different organizations: IPA, 1AF, Mumias Sugar Company, and the German Agro Action (Welthungerhilfe).[72] These sources provided over 30,000 soil tests for

[70] The target yield of 2 t/ha aimed at generating an improvement over the baseline average of 1.42 t/ha, while keeping the cost of the input package affordable for farmers. The government's recommended application of phosphorus for Western Kenyan soils, for a target yield of 3.9 t/ha in soils with P below 10 mg/kg, is 26kg P/ha, corresponding to 130 kg DAP/ha, (FURP, 1995; Wasonga et al., 2008). With a target yield of 2 t/ha, the recommendations provided as part of this program involved applying 21kg P/ha, corresponding to 107 kg DAP/h.

[71] Recommendations were provided based at the ward level because that is the most precise information collected about farmers' location. The data was aggregated into medians because the majority of the soil data available was not geocoded and only provides information on the administrative unit in which the sample was collected.

[72] The IPA dataset was assembled in 2011 and 2014 in Busia county for previous projects (Fabregas et al., 2017a) and extended in 2016 as part of test plot activities in the same area. The 1AF data was collected in 2016 across the entire study area. Mumias Sugar Company shared the data they collected for their operations in Busia and

program area. However, in order to base the recommendations on the most recent data, data was dropped for soil tests performed before 2014, when possible. The final dataset used included about 7,085 observations for 108 wards.[73]

Messages were sent either in English or in Swahili, depending on farmers' preferences indicated during the baseline phone survey. We list the messages below:

- Welcome to PAD, IPA's free advice service for maize growers. You will receive advice for your needs based on more than 10,000 soil tests from Western Kenya.
- The soil in your area is [level] acidic. To avoid low yields treat now. Apply [quantity] bottle top of lime per planting hole. [quantity] kgs for 1/4 acre. OR The soil in your area is slightly acidic. According to our soil analysis, farms in your area do not need lime.
- Soil acidity causes stunted growth.Lime reduces soil acidity and makes nutrients of DAP more available for your maize.
- When planting, apply [quantity] bottle top of lime. Cover with a handful of soil. Add [quantity] bottletop of DAP, cover with enough soil to avoid direct contact of inputs. OR When planting, apply [quantity] bottle top of DAP, cover with enough soil to avoid direct contact of inputs.
- Check your phone! We sent you 3 planting recommendations last week [If you flash [number] before Friday this week, we will you callback soon to explain them/We will call you soon to explain them]
- Top-dress when your maize has more than 4 leaves up to knee high. If rains are good.apply 3/4 bottle top of UREA. If rains are low, apply 3/4 bottle top of CAN.
- UREA can increase your maize yields as much as CAN if rains are good. Try 11 kg of urea in 1/4 acre and see the results
- Check your phone! We sent you 2 topdressing messages this week [If you reply YES or flash [phone] by Tuesday, we will call you back soon to explain them/We will call you soon to explain them.]

2.11.4 B.4 1AF1-K

In September 2016, during the period in which 1AF farmers were placing their orders for the 2017 long rains season, 1AF sent SMS messages about soil acidity and agricultural lime. Two types of messages were sent: the first, denoted as "Broad", simply encouraged farmers to use lime to reduce soil acidity and increase yields, while the second, denoted as "Detailed" provided recommendations on lime application rates and expected yield increase customized to the farmers' site.

Kakamega counties between 2009 and 2016. The German Agro Action data was collected in Kakamega and Siaya counties in 2015.

[73] Data collected before 2014 was dropped if at least 30 more recent observations in the ward were available. Since the data displays clear trends of decreasing pH and phosphorus levels over time, they were adjusted using coefficients based on the Mumias Sugar Company soil data: a coefficient of −0.027 per year was applied for pH and −0.504 per year for phosphorus. These coefficients were obtained by regressing pH and phosphorus data on a time trend and constant, controlling for field fixed effects, these regressions are based on a sample of over 60,000 observations.

Although the standard application rate recommended by 1AF and reflected in field materials was 200kg/acre across the entire program, the detailed message encourages farmers to use different application rates based on the pH level predicted for the farmers' site.[74] To obtain these predictions, 1AF used their own own soil tests, performed using soil spectroscopy, and soil data collected for a previous project by IPA (Fabregas et al., 2017b) and analyzed by the Kenya Agricultural Research Institute (KARI) using wet chemistry in 2011 and 2014. These soil chemistry results were then interpolated across areas through Kriging to create a continuous field of soil chemistry predictions. Optimal lime application rates, for each level of pH, were based on 1AF on-farm agronomic trials conducted in 2015 (1AF, 2015). During that trial three different lime application rates were tested: 50kg/acre, 100kg/acre and 200kg/acre. The sample was divided according to pH quintiles and, for each quintile, the lime application rates that resulted in the most precisely estimated effect on yield was chosen. Two different lime application rates were recommended, based on the local predicted level of pH: 200kg/acre and 50kg/acre. [75]

Farmers in both treatment groups received 6 identical messages, all messages were sent in Swahili. We report the messages below:

- [Broad]: Hello [name],Your soil is acidic. Use lime to reduce acidity and increase yields.Call [phone]
- [Detailed]: Hello [name],Your soil is [level] acidic. We recommend [amount] kg of LIME per acre at [total cost] Ksh. Use lime to reduce acidity and increase yields [percentage increase]%.Call [phone]

2.11.5 B.5 1AF2-K

In September 2017, when 1AF farmers were enrolling for the 2018 long rains season, 1AF implemented a second program aimed at encouraging lime adoption. The purpose of this program was to understand how to optimize message content, framing, number of repetitions and framing. In addition subset of farmers was randomly assigned to receiving additional messages encouraging use of an extra amount of topdressing fertilizer (Extra CAN).

Six different types of messages were sent: a "Basic" message simply recommended to purchase lime, a message, "Yield increase", also mentioned that lime would increase yields, two encouraged experimentation, "Experimentation (selfish)" and "Experimentation (neighbors)", and two leveraged on behavioral nudges "Social comparison" and "Self-efficacy". Half of the treated farmers were randomly assigned to receive messages addressing the whole family instead of the individual (by replacing the word "you" with "your family"). The messages encouraging use of additional quantities were identical to those encouraging use of lime (the word "Lime" was replaced by "Extra CAN"). Farmers assigned to receive both lime and fertilizer message were randomly assigned to receive one of the two first and the other on the next day for all repetitions. The number of repetitions (from 1 to 5) and the frequency of the messages (every 2, 4, 6, or 8 days) were cross-randomized. We report all messages below:

- [M1: Basic] [Name], 1AF recommends [you/your family] register to buy [Lime/Extra CAN] for your maize.

[74] Since 1AF does not collect the coordinates of farmers' plots, farmers were assigned to the GPS coordinates of the site to which inputs are delivered by 1AF.

[75] Robert On, Matthew Lowes, and David Guerena produced these recommendations.

- [M2: Yield increase] [Name], 1AF recommends [you/your family] register to buy [Lime/Extra CAN] for your maize. You'll get higher yields by using [Lime/Extra CAN].
- [M3: Experimentation (selfish)] [Name], 1AF recommends [you/your family] register to buy [Lime/Extra CAN] for your maize. Try it on just a small part of your land to see the benefits.
- [M4: Experimentation (neighbors)] [Name], 1AF recommends [you/your family] register to buy [Lime/Extra CAN] for your maize. Try it on just a small part of your land to so that you and your neighbors can see the benefits.
- [M5: Social Comparison] [Name], 1AF recommends [you/your family] register to buy [Lime/Extra CAN] for your maize. Farmers all over Western are getting bigger yields by using [Lime/Extra CAN]. Keep up with them!
- [M6: Self-efficacy] [Name], 1AF recommends [you/your family] register to buy [Lime/Extra CAN] for your maize. You have the ability to achieve higher yields by using [Lime/Extra CAN]!

2.11.6 B.6 1AF3-R

1AF-Rwanda, known as Tubura, implemented an SMS-based program aimed at encouraging experimentation with a type of agricultural lime, travertine. The messages were sent in June 2017, when farmers were enrolling for the 2018 main rains season (August 2017 to January 2018). As for 1AF2-K, The purpose of this program was to understand how to optimize message content, framing, number of repetitions and framing. In addition, given the relatively low mobile phone penetration in the country, 1AF wanted 61 to explore ways to increase spillovers within farmers' group in order to reach all farmers. For this reason, the first stage of randomization took place at the group level, assigning farmers groups to four group-level treatments: a pure control group where no farmers in the group (GO: No SMS), a treatment in which all farmers received identical messages (G1: Same SMS), a treatment in which all farmers in the group received messages, but content and framing were randomly assigned at the individual level (G2: Diff SMS), and a treatment in which farmers received messages with probability 0.5 and content and framing were randomized at the individual level (G3). In this paper we focus on the direct effect of the receiving messages on individual farmer rather than on group level outcomes. Therefore, we divide farmers in G3 into messages receiving (G3 - Treated) and non message receiving (G3-Control). We exclude from our analysis all farmers who did not have a phone number registered in 1AF's database.

Seven types of messages were sent: a basic message encouraged to purchase lime "General Promotion", the second indicated the application rate and expected impact "Specific + yield impact", the third helped farmers assess their need for lime "Self-diagnosis", the fourth encouraged farmers to have their soil tested "Soil test", the firth explained that lime can be used to increase fertilizer efficiency "How travertine works", the sixth encouraged farmers to order lime immediately "Order immediately", and the seventh indicated that acidity was a problem in the farmer's area "Your cell is acidic + yield impact". All messages were either framed positively (gain) or negatively (loss). The number of repetitions (from 1 to 4) Finally, message receiving farmers in half of the treated groups received an additional message encouraging them to spread the information to others in their group, especially those without phones (Social nudge message).

- [T1-G: General promotion (gain)] Many fields in Rwanda have acidic soil and need TRAVERTINE to increase yields. Order from TUBURA now.
- [T1-L: General promotion (loss)] Many fields in Rwanda have acidic soil and need TRAVERTINE to avoid a yield loss. Order from TUBURA now.
- [T2-G: Specific+ yield impact (gain)] Many fields in Rwanda have acidic soil. Applying 25 kg/are of TRAVERTINE will increase yields by 20%.Order from TUBURA now.
- [T2-L: Specific+ yield impact (gain)] Many fields in Rwanda have acidic soil. Applying 25 kg/are of TRAVERTINE will prevent a yield loss of 20%. Order from TUBURA now.
- [T3-G: Self-diagnosis (gain)] Do you have fields with poor harvests even when you use fertilizer? You probably have acidity and need TRAVERTINE to increase yields. Order from TUBURA now.
- [T3-L: Self-diagnosis (loss)] Do you have fields with poor harvests even when you use fertilizer? You probably have acidity and need TRAVERTINE to avoid a yield loss. Order from TUBURA now.
- [T4-G: Soil test (gain)] Ask your Field Officer for a free soil test to learn if your fields are acidic and you need to order TRAVERTINE to increase yields.
- [T4-L: Soil test (loss)] Ask your Field Officer for a free soil test to learn if your fields are acidic and you need to order TRAVERTINE to avoid a yield loss.
- [T5-G: How travertine works (gain)] Many fields in Rwanda have acidity, which blocks fertilizer uptake. Applying TRAVERTINE solves the problem, increasing crop yields. Order from TUBURA now.
- [T5-L: How travertine works (loss)] Many fields in Rwanda have acidity, which blocks fertilizer uptake. Applying TRAVERTINE solves the problem, preventing a yield loss. Order from TUBURA now.
- [T6-G: Order immediately (gain)] Many fields in Rwanda have acidic soil and need TRAVERTINE to increase yields. Order it immediately, when signing your TUBURA order form.
- [T6-L: Order immediately (loss)] Many fields in Rwanda have acidic soil and need TRAVERTINE to avoid a yield loss. Order it immediately, when signing your TUBURA order form.
- [T7-G: Your cell is acidic + yield impact (gain)] In your cell the soil is acidic. If you apply 25 kg/are of TRAVERTINE you can boost yields by 20%. Order from TUBURA now.
- [T7-L: Your cell is acidic + yield impact (loss)] In your cell the soil is acidic. If you apply 25 kg/are of TRAVERTINE you can avoid a yield loss of 20]%. Order from TUBURA now. Social nudge message:
- [SN] Please share this information about TRAVERTINE with your group members and neighbors, especially those who don’t have phones!

2.12 References

Aker, J. C. (2017). Using digital technology for public service provision in developing countries. Digital Revolutions in Public Finance, pages 201–24.

Aker, J. C., Collier, P., and Vicente, P. C. (2017). Is information power? using mobile phones and free newspapers during an election in mozambique. Review of Economics and Statistics, 99(2):185–200.

Aker, J. C., Ghosh, I., and Burrell, J. (2016). The promise (and pitfalls) of ict for agriculture initiatives. Agricultural Economics, 47(S1):35–48.

Aker, J. C., Ksoll, C., and Lybbert, T. J. (2012). Can mobile phones improve learning? evidence from a field experiment in niger. American Economic Journal: Applied Economics, 4(4):94–120.

Anderson, J. and Feder, G. (2007). Agricultural extension. Handbook of agricultural economics.

Anderson, J. R. and Feder, G. (2004). Agricultural extension: Good intentions and hard realities. The World Bank Research Observer, 19(1):41–60.

Brady, N. and Weil, R. (2004). Soil acidity, alkalinity and salinity. Elements of the Nature and Properties of Soils, pages 266–315.

Camacho, A. and Conover, E. (2010). The impact of receiving price and climate information in the agricultural sector.

Casaburi, L., Kremer, M., Mullainathan, S., and Ramrattan, R. (2014). Harnessing ICT to increase agricultural production: Evidence from Kenya. Unpublished Working Paper.

Cole, S. A. and Fernando, A. N. (2016). 'mobile'izing agricultural advice: Technology adoption, diffusion and sustainability.

Cunha, N., Lichand, G., Madeira, R., and Bettinger, E. (2017). What is it about communicating with parents. Technical report, October 2017 Mimeo.

Duflo, E., Kremer, M., and Robinson, J. (2008). How high are rates of return to fertilizer? evidence from field experiments in Kenya. The American economic review, 98(2):482–488.

Duflo, E., Kremer, M., Robinson, J., and Schilbach, F. (2018). Learning about appropriate technology use: Evidence from fertilizer use in Western Kenya. Working Paper.

Evenson, R. E. and Gollin, D. (2003). Assessing the impact of the green revolution, 1960 to 2000. Science, 300(5620):758–762.

Fabregas, R., Kremer, M., Robinson, J., and Schilbach, F. (2017a). The effectiveness of public agricultural extension: Evidence from two approaches in Kenya. Unpublished Working Paper.

Fabregas, R., Kremer, M., Robinson, J., and Schilbach, F. (2017b). Netflix for agriculture. Unpublished Working Paper.

Fafchamps, M. and Minten, B. (2012). Impact of sms-based agricultural information on Indian farmers. The World Bank Economic Review, 26(3):383–414.

FURP (1995). New fertilizer recommendations for small-scale farmers in kenya. KARI Information Bulletin - No. 12.

Gudu, S., Okalebo, J., Othieno, C., Obura, P., Ligeyo, D., Shulze, D., and Johston, C. (2005). Response of five maize genotypes to nitrogen, phosphorus and lime on acid soils of Western Kenya. In African Crop Science Conference Proceedings, volume 7, pages 1109–1115.

Hall, A. K., Cole-Lewis, H., and Bernhardt, J. M. (2015). Mobile text messaging for health: a systematic review of reviews. Annual review of public health, 36:393–415.

Head, K. J., Noar, S. M., Iannarino, N. T., and Harrington, N. G. (2013). Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Social science & medicine, 97:41–48.

Kiplagat, J. K., Okalebo, J. R., Serrem, C. K., Mbakaya, D. S., and Jama, B. (2014). Determination of Appropriate Rate and Mode of Lime Application on Acid Soils of Western Kenya: Targeting Small Scale Farmers, pages 177–186. Springer International Publishing, Cham.

Kisinyo, P., Opala, P., Palapala, V., Gudu, S., Othieno, C., and Ouma, E. (2015). Micro-dosing of lime, phosphorus and nitrogen fertilizers effect on maize performance on an acid soil in Kenya. Sustainable Agriculture Research, 4(2):21.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. Econometrica, 75(1):83–119.

Ksoll, C., Aker, J., Miller, D., Perez-Mendoza, K., and Smalley, S. (2014). Learning without teachers? A randomized experiment of a mobile phone-based adult education program in Los Angeles.

Larochelle, C., Alwang, J., Travis, E., et al. (2016). Did you really get the message? using text reminders to stimulate adoption of agricultural technologies.

NAAIAP (2014). Soil suitability evaluation for maize production in Kenya. Manual.

Nakasone, E., Torero, M., and Minten, B. (2014). The power of information: The ICT revolution in agricultural development. Annu. Rev. Resour. Econ., 6(1):533–550.

1AF (2014). Managing soil acidity with lime (2014 phase 2 trial report). Technical report, One Acre Fund.

1AF (2015). Managing soil acidity with lime (2015 trial report). Technical report, One Acre Fund.

1AF (2017). Annual report.

Pauw, E. F. D. (1994). The management of acid soils in Africa. Outlook on agriculture, 23(1):11–16.

The, C., Calba, H., Zonkeng, C.and Ngonkeu, E. L. M., and Adetimirin, V. O. (2006). Response of maize grain yield to changes in acid soil characteristics after soil amendment. Plant Soil.

Tisdale, S. L., Nelson, W. L., and Beaton, J. D. (1990). Soil fertility and fertilizers. New York: Macmillan.

Wasonga, C. J., Sigunga, D. O., and Musundu, A. O. (2008). Phosphorus requirements by maize varieties in different soil types of Western Kenya. Africa Crop Science Journal, vol. 16:161–173.

3 Behavioral Spillovers via SMS: Experimental evidence from Rwanda

with Tomoko Harigaya, Michael Kremer, Matthew Lowes, and Giulia Zane[76]

3.1 Abstract

We examine the effects of an SMS campaign in Rwanda which promoted the adoption of lime, a relatively new agricultural input, among farmers in a group-based credit and extension program. The campaign experimentally varied message diversity and intensity within farmer groups as well as message content, framing,and repetition. On average, SMS treatments increased the likelihood of ordering lime through the program by 20% over the adoption rate of 5.2% in the controlgroup. Sending diverse messages to farmers within a group increased lime adoption by as much as sending the most effective, identical message to all farmers. Furthermore, diverse messages generated spillover effects, increasing lime adoption among those who did not receive SMS messages by 9-14%. These effects are equally large across farmer groups that received varying intensities of treatment. An additional SMS encouraging farmers to share information had no spillover effect. Our findings suggest that when relative effects of different messages are unknown, diversifying messages in a group setting could effectively facilitate behavioral spillovers.

3.2 Introduction

ICT interventions offer promising potential to improving the delivery of agricultural extension in rural areas. Digital technology allows delivery of customized information at low cost and at scale, overcoming many barriers present in a traditional form of in-person extension. In fact, recent studies demonstrate that mobile phone-based agricultural extension could increase agricultural knowledge (Cole and Sharma, 2018) adoption of recommended practices (Fabregas et al., 2018), and yields (Casaburi et al., 2014; Cole and Fernando, 2016). However, while evidence supports the potential of digital agricultural extension, little is known about how and when these interventions yield better results.

This study evaluates the effects of an SMS campaign to promote the adoption of an agricultural input among over 216,475 smallholder farmers in Rwanda. One Acre Fund (1AF), a non-profit offering a group-based agricultural credit and extension program to smallholder farmers in East Africa, implemented a field experiment, encouraging its member farmers to apply agricultural lime, a soil additive that neutralizes acidity in soil and improves its capacity to absorb nutrients. Five campaign features were randomized. First, 1AF varied message diversity and intensity at the farmer-group level: farmer groups were assigned to one of the following experimental arms: 1) identical messages for all phone owners (G1), 2) diverse messages for all phone owners (G2), 3) diverse messages for 50% of phone owners (G3), or 4) no message (G0). Second, farmers groups in G1 and individual farmers in G2 and G3 were randomized to receive one of seven messages, which encouraged farmers to order lime from 1AF because many soils in Rwanda are acidic (basic message), provide specific information on dosage and impact, help

[76] Harigaya: Precision Agriculture for Development (tharigaya@precisionag.org); Kremer: Harvard University (mkremer@fas.harvard.edu); Lowes: One Acre Fund (mlowes@oneacrefund.org); On: University of California, Berkeley (roberton@gmail.com); Zane: Inter-American Development Bank (giuliaz@iadb.org)

farmers diagnose whether their field needs lime, encourage farmers to request a free soil test to diagnose the need for lime, explain how lime could improve fertilizer uptake and yields, promote urgency, or increase relevance by emphasizing the local soil acidity. Third, farmer groups in G1 and individual farmers in G2 and G3 were randomly assigned to receive either a gain-framed ("to increase yields") or loss-framed ("to avoid yield loss") message. Fourth, farmers were randomized to receive a varying number of message repetition between one and four. Finally, randomly selected farmers in G1-G3 received an additional message encouraging them to share the information in SMS with other farmers in the group.

Four key insights emerge from the analysis of 1AF's administrative data on input orders. First, SMS messages are effective in facilitating behavior change: the SMS-based promotion increased the likelihood of ordering lime from 1AF by 13.6% among existing 1AF members with a mobile phone. Second, message framing matters: the messages that encouraged lime adoption "to increase yields" resulted in a 21% increase in lime adoption, whereas the messages that encouraged adoption "to avoid yield loss" had no effect. The difference between the two messages is statistically significant at 10%. Third, farmers in the same group share information sent via SMS: sending diverse messages to farmers within a group was as effective as sending the best message to all farmers in a group. Fourth, diverse messages are effective in generating spillover effects within groups: non-phone owners and those assigned not to receive messages had a significantly higher lime adoption rate than their counterparts in the control group. Furthermore, this effect was equally large and robust across groups with varying shares of farmers receiving SMS messages. Overall, these findings suggest that small tweaks in message and dissemination design influence individual's behaviors, highlighting the importance of testing.

3.3 Background

3.3.1 Soil acidity and agricultural lime

Soil acidity is a widespread problem for agricultural production in East Africa, and Rwanda is no exception. Acidic soil limits the capacity of plants to absorb nutrients, reducing the potential benefits of fertilizers and thereby land productivity. Many agronomic experts recommend the use of agricultural lime, also called "travertine" in Rwanda, to neutralize acidity in soil. Agronomic trials in this region (Kisinyo, 2015) and 1AF's own trial plots (Owino, 2016) have demonstrated that the application of lime could significantly increase maize yields (see Fabregas et al., 2018 for a detailed discussion on the issues around soil acidity in East Africa). However, lime is a relatively new technology in East Africa and the adoption rate in Rwanda has remained low. For example, fewer than 5% of 1AF members ordered lime in 2016.

3.3.2 Tubura

One Acre Fund (1AF), locally known in Rwanda as Tubura, is a non-profit organization providing agricultural credit and training services to smallholder farmers in East Africa. Under the core program, members form groups of 5-17 and undertake a number of activities together. Group members place orders for agricultural inputs and other household products offered by 1AF several months before the start of the agricultural season. Once the agricultural inputs are delivered, members prepare their fields and plant together. Members then meet regularly and discuss agricultural practices and loan payments with a Field Officer (FO) from 1AF throughout the season. Importantly, farmers are liable for other group member's loan payment and will be disqualified from participating in the program in the subsequent season if one or more farmers in the group default.

In Rwanda, over 200,000 farmers enroll in the 1AF program every season, majority of whom grow maize as a primary crop during the main agricultural season between March and May (called "Long Rains (LR)"). The basic maize package included seeds, planting fertilizers, and topdressing fertilizers. In addition, 1AF offers other agricultural and non-agricultural products, of which most popular are agroforestry trees and solar lamps.

In 2016, the Rwandan government started promoting lime adoption through district-level subsidy schemes. Even though 1AF had been offering lime for many years at that point, take-up was low at below 3%. In parallel to the government's initiative, 1AF increased its lime marketing effort, providing training and marketing materials and a financial incentive for lime sales to its field officers.

1AF sold lime by acre: a farmer specifies a plot size for which she or he is purchasing lime, and 1AF delivers the recommended quantity for the ordered plot size. During this experiment, 1AF sold 25 kg of lime per acre at 2,500 RWF (≈$2.83), or 100 RWF per kg.

3.4 Experimental design

During the enrollment period for the LR season in 2018, 1AF discontinued the lime incentive scheme for field officers, and instead implemented an SMS campaign to encourage farmers to order lime. 1AF designed a large-scale experiment to test the overall effectiveness of the SMS campaign and the relative effects of several message design features. The experiment took place in all districts where 1AF had existing members in 2017, involving over 200,000 farmers.

3.4.1 Experimental treatments

The experiment varied five features of the SMS campaign. Some of the features were randomized at the group-level, while others at the individual farmer-level. We describe each of the five experimental features below.

Group message type: First, 1AF randomly assigned farmer groups to one of four experimental arms: (1) "same message" arm where all farmers with a registered mobile phone ("phone owners") in a group receive an identical message, (2) "diverse messages" arm where phone owners within a group receive different messages, (3) "diverse messages (low-intensity)" arm where phone owners within a group receive different messages at a 0.5 probability and no message otherwise[77] and (4) control arm where none of the farmers receive a message. Each farmer group had a 0.5 probability of being assigned to the diverse messages (low-intensity) arm and a one-sixth probability of being assigned to one of the remaining experimental arms. Throughout the rest of the paper, we refer to all farmers that were assigned to receive any SMS messages in G1, G2, and G3 as "treated farmers".

Message variation: Second, 1AF created an exogenous variation in message content across farmer groups in the "same message" arm and among individual treated farmers in the two

[77] Appendix Figure A1 shows the distribution of the share of farmers in a group that were assigned to receive an SMS for the sample of farmer groups in this treatment arm.

"different messages" arms. The message variation included seven different ways of encouraging farmers to adopt lime. A basic message (T1) read: "Many fields in Rwanda have acidic soil and need TRAVERTINE to increase yields. Order from TUBURA now." In comparison, other messages were designed to provide specific information on dosage and impact (T2), help farmers diagnose whether their field needs lime (T3), encourage farmers to request a free soil test to diagnose the need for lime (T4), explain how lime could improve fertilizer uptake and yields (T5), promote urgency (T6), and increase relevance by emphasizing the local soil acidity (T7). (See Appendix Table A1 for the exact texts of SMS messages.)

Message framing: Third, in addition to varying message content, 1AF randomized message framing: a "gain" framing encouraged lime adoption "to increase yields" while a "loss" framing encouraged lime adoption "to avoid yield loss". Again, this randomization was carried out at the farmer group level in the "same message" arm so that all farmers in the same 1AF farmer group received an exactly identical message while treated farmers in "different message" arms were randomized at the individual level.

Message repetition: Fourth, 1AF randomly varied the number of messages sent to the treated farmers. Across all farmer groups, each treated farmer had an equal probability of receiving one, two, three, or four messages. Repeated messages were delivered 3-4 days apart.[78]

Social nudge: Finally, half of all treated farmer groups were randomly assigned to receive an additional message which encouraged information sharing among group members. This message was specifically designed to encourage treated farmers to share the message with those who do not have a registered phone: "Please share this information about TRAVERTINE with your group members and neighbors, especially those who don't have phones."

3.4.2 Sample Frame

The full sample of this study consists of 216,475 farmers who were enrolled in the 1AF program in 2017. Table 2 provides the summary of baseline characteristics of our sample. At baseline, 114,582 farmers (52.9%) had registered a phone number with 1AF. With the average group size of 10, this means that an average farmer group had 5-6 members that had a registered mobile phone. Even though 1AF has been operating since 2009, active members in 2017 were relatively new, with the average of 1.6 years of enrollment. The lime adoption rate at baseline is low at 3.9%. The average quantity of lime order was 1.49 kg for the full sample in the control arm, implying that an average adopter purchased 38 kg of lime at 100 RWF/kg of lime, or 380 RWF in credit, which represents a small proportion (1.9%) of the average credit volume of 19,990 RWF (≈$22). Finally, note that 16% of the farmer groups had at least one member who purchased lime from 1AF in 2017.

3.4.3 Mobile phones registered by multiple farmers

One complication arose in the analysis of the data from this experiment because an insignificant number of farmers had registered a mobile phone number used by other 1AF members. Out of 114,582 farmers who had a registered phone number, only 85,160 farmers (74%) had unique phone numbers. This means 29,422 (26%) were sharing it with other members of the program; among those, 11,427 farmers (10%) had a phone number shared with 1AF members from other farmer groups. This resulted in 6.3% of farmers in treated farmer groups (or

[78] Note that the SMS delivery rate through 1AF's SMS platform is roughly 60%.

11.8% of treated farmers) receiving multiple different messages, and 7.9% of farmers in the control arm receiving at least one message. Furthermore, 18.8% of farmer groups in the same message arm received more than one type of message.

This problem may affect both the accuracy and precision of our impact estimates. But, because the marginal effects of an additional SMS and increasing message diversity may not be constant, the direction of the overall bias is ambiguous. For example, the effect of the SMS treatment may be attenuated under this contamination (or "mixed treatment") problem given that nearly 20% of farmer groups in the control arm received at least one message. However, if SMS messages had an increasing marginal effect over the total number of SMS messages, this problem would have a larger effect on lime adoption in SMS treatment arms (G1 - G3) than in the control arm, resulting in an overestimation of the true effect. Similarly, the comparison between the identical and diverse messages is imperfect as 18.8% of farmer groups in the "same message" arm received multiple types of messages, seemingly reducing the difference between the two treatment arms. On the other hand, the proportion of farmers that received unassigned messages is actually higher in the "diverse messages" arms (32.7-34.7% in G1 and G2 as opposed to 18.8% in G1), potentially increasing message diversity by a greater extent. In this way, we cannot pin down in which direction the mixed treatment problem affects our point estimates. We argue, however, that the intended variations across different treatments are still retained: SMS treatment arms received more SMS messages than the control arm; the diverse-messages arms received more diverse messages than the same-message arm; and over 81.7% of treated farmers received the assigned number of messages. Therefore, the Intent-to-Treat (ITT) effects we estimate provide qualitatively valuable insights into the effects of message features.

3.4.4 Randomization balance

Randomization was carried out independently for each of the five message features. Table 2 presents the differences in baseline characteristics across experimental groups, estimated in an OLS model with standard errors clustered at the farmer-group level. For each randomization, we report at the bottom of the panel the p-value from the joint significance test across all experimental groups. Out of 40 F-tests, 3 are significant at the 5% level.

We further investigate the baseline balance in the following exercise. First, we simulate randomizations and obtain the distribution of the number of pairwise tests across all experimental groups for baseline outcomes that fail at a given significance level. Second, we compare this distribution to the binomial distribution: the exercise with 25,000 simulations reveals that our outcomes are highly interdependent (See Figure A2), and therefore using binomial or order statistics to assess baseline balance would be inappropriate. Given these results, we use an approach similar to the permutation test and compare the number of pair-wise balance tests that fail at 1% and 5% significance levels in the observed experimental data to the distribution of the equivalent numbers in the 25,000 iterations of simulated randomizations.[79]

Out of 103 tests, our observed randomization yields 1 significant result at the 1% level and 17 at the 5% level: the likelihoods of observing these numbers of significant tests in the

[79] In this approach, we re-randomize the observed data to generate simulation data: this is in practice slightly different from the standard permutation test. We cannot permute some of the randomizations which were carried out at the group-level for some and at the individual-level for others.

simulated data are 28.5% and 8.5%, respectively. While these results suggest that we were somewhat unlucky in our random draws, these likelihoods are not alarmingly high. In our analysis, we control for the presence of any lime adopter, group size, and total credit volume in the previous season.

3.4.5 Data and outcome measures

We use the administrative data from 1AF to examine the effects of SMS messages on lime adoption. The database of all active 1AF members in 2017 contains information on location, farmer group, detailed credit portfolio for the two seasons in 2017, and the tenure at 1AF. We use the equivalent database for the 2018 season as our outcome data: 65% of members in the baseline data appear in the 2018 data, which we consider as a retention rate between the two seasons.[80] We assume that all farmers that do not merge with the 2018 enrollment data received no credit or inputs from 1AF.

The analysis uses two measures of lime adoption as primary outcomes of interest: whether a member ordered any lime from 1AF and how much lime she ordered during the 2018 enrollment period. While a small proportion of enrolled farmers drop out between enrollment and start of the agricultural season, the attrition rate is generally small. %at around XX.

3.5 Empirical strategy

We assess the impact of the SMS messages on lime adoption using the following OLS model:

$$Y_{ig} = \alpha + \sum_{j=1}^{n} \beta_j T_{jg} + \delta_i X_i + \gamma_g Z_g + \varepsilon_i$$

where Y_{ig} is the outcome measure of lime adoption for an individual i in group g and T_g the vector of indicators of the group-level assignments for message features where n indicates the number of experimental groups. For example, when testing the effects of message diversity and intensity treatments, $n = 3$; when comparing the relative effects of seven individual message types in G1 to the two diverse messages arms, $n = 9$. Finally, X_i and Z_g are the vectors of individual and group characteristics, respectively. Since random assignment ensures that the error term ε_i is orthogonal to T_{ij}, β_j measures the unbiased intent-to-treat (ITT) effect of the SMS treatment j.

3.6 Results

3.6.1 Treatment effects among phone owners

3.6.1.1 Effects of message diversity, content and framing

We first estimate the ITT effect of receiving any SMS messages by comparing lime adoption among all treated farmers across three group-level treatment arms (i.e., all phone owners in G1 and G2 and 50% of phone owners in G3 that were assigned to receive SMS

[80] Note, however, that this figure is likely an underestimate of the actual retention as an insignificant number of members receive a new client ID every season, and we are not able to link them between the seasons.

messages) to that of mobile phone owners in the control arm. On average, the SMS treatment increased the likelihood of ordering lime by 0.96 percentage point, or 18% over the control mean of 5.2% adoption rate. We observe a similar effect on the amount of lime ordered: an increase by 0.42 kg, or a 14.3% increase over the control mean of 2.86 kg. We show in Columns (3) to (5) that lime campaign did not substantially affect the overall enrollment, credit portfolio or other input orders likely because lime is a very small proportion of credit portfolio.

In Panel B, we provide suggestive evidence that message diversity among farmer groups facilitated lime adoption more effectively than identical messages. The point estimates in Columns (1) and (2) indicate that farmer groups that received diverse messages saw a somewhat larger increase in lime adoption and that the effect of sending diverse messages to 50% of phone owners is comparable to that of sending identical messages. However, the F-tests for the equality of coefficients on G1 and G2 are insignificant, and therefore the difference in the average effects of identical and diverse messages is only suggestive.

We next test whether any one message type or framing in the same-message arm is as effective as sending diverse messages. We do this by comparing the effects of different message types and framing in the same-message arm to the average effects of sending diverse messages. Table 4 present the results. Overall, diverse messages is no more effective than sending the best message to all farmers in the same group. The coefficients suggest that the messages that conveyed the benefits of lime and used gain framing were effective, increasing lime adoption by 14 and 11 percentage points, respectively, in comparison to the control arm that received no messages. These coefficients are comparable to the point estimates for the diverse-message arms, providing no evidence that message diversity affects the intensity or the pattern of information flow within a group.

To assess the relative effectiveness of different message types, we conduct F-tests for the equality of coefficients. The test fails to reject the null hypothesis of the equality of coefficients for the message that emphasized the benefits of lime (T5) and the basic message (T1). On the other hand, the equality test of coefficients for the gain and loss framing is rejected at the 10% significance level. In fact, none of the message types in loss framing increased the likelihood of ordering lime in the same-message arm (results not shown). These results together suggest that the content of SMS matters: in this particular setting, farmers responded to the message that encouraged adoption "to increase yields", but not to the message that appealed to loss aversion - that lime could help "avoid yield loss".

3.6.1.2 Effects of message repetition

Turning to the variation in the number of messages, Table 5 shows a large marginal effect of sending a second message but no additional effect beyond the second. The point estimates suggest that the effect of the second message, an increase by 0.7 percentage points in the likelihood of ordering lime, was nearly 170% of the first message. Message repetition may help reduce the SMS non-delivery rate.[81] However, the relative magnitudes of the coefficients on the first and second messages imply that the effect of the second message is driven at least partially by message repetition and not simply due to reducing non delivery. The results are consistent for the quantity of lime ordered.

[81] The platform 1AF uses to communicate with its farmers in Rwanda has an average non-delivery rate of 60%.

3.6.2 Spillover effects

The earlier results - that message diversity within a group could achieve an increase in lime adoption as large as the identical, best message - suggest that farmers share information delivered via SMS. Our experimental design allows us to formally test the spillover effects of the SMS campaign in two ways. First, we compare the effects of group-level treatments on lime orders among farmers who had no registered phone, and therefore did not receive SMS messages; second, taking advantage of the fact that 50% of phone owners in G3 were assigned not to receive SMS treatments, we compare lime orders between these particular phone owners in G3 and phone owners in the control arm.

3.6.2.1 Effects of message diversity

Table 6 show that SMS messages influenced behavior of non-SMS receivers and that diverse messages, on average, were somewhat more effective in generating spillover effects than the identical messages. The point estimates reported in Column (1) show that the identical messages had no economically meaningful effect, while diverse messages increased the likelihood of ordering lime among non-phone owners by roughly 0.3 percentage points, a 13.6% increase over the control mean of 2.2%. The effects on the quantity ordered are even larger: the point estimates on the diverse messages groups suggest 22.5-27.9% increases over the control mean of 0.93 kg. We report the p-values from F-tests comparing the effect of identical messages and the average effect of diverse messages (G1 vs. G2 or G3) in Table 6. While not consistently robust, p-values of 0.124 for the likelihood of ordering lime and 0.048 for the quantity ordered provide a suggestive indication that diverse messages, compared to identical messages, generated larger spillover effects.

Interestingly, we find no effect of social nudge - an additional SMS message that encouraged farmers to share the information on lime with other farmers in the group. In Columns (3) and (4), we show that non-phone owners in the groups assigned to receive the nudge message were no more likely to order lime than those in other groups even though the message specifically encouraged the sharing of information with non-phone owners. In a group setting where members are already sharing information with others, an extra nudge via SMS may have little effect on the behavior of non-SMS receivers. This is also corroborated by the observed relative effects of the two diverse messages arms - the difference in the intensity of the SMS treatment did not affect the magnitude of the spillover effect.

In Columns (4)-(8) we provide consistent results on spillovers in a different sample frame - among phone owners. Diverse messages increased lime adoption among phone owners who didn't receive messages by 4.9 percentage points, or nearly 20% of the control mean (Column 4). These estimates are less precise likely because of the smaller sample size. In Columns (4)-(5), we again find no evidence for the effect of social nudge messages.

3.6.2.2 Effects of message content and framing

Finally, we analyze the spillover effects by message content and framing. In Table 4, we showed that the gain framing was particularly effective in the same message arm and that diverse messages within a group increased lime adoption by a similar magnitude as the best, identical message treatment. We use the same specification to gain insights on whether we observe a similar pattern in the spillover effects. Table 7 presents the results. Unfortunately, observed point estimates in this sample are substantially smaller than those for the treated farmers reported in Table 3, limiting our ability to detect small effects of individual message type. However, in Columns (3) and (4), we provide suggestive evidence that gain framing is more effective than loss framing in generating spillover effects in the same message arm - the difference between the coefficients for the two types of messages is marginally significant at 10%. Furthermore, the point estimates for gain framing in G1 are similar to those for G2 and G3. Even though these results are not statistically robust, they provide a consistent indication that sending diverse messages to farmers in the same group was as effective as sending the best identical message to all farmers in the group in increasing the likelihood of ordering lime both for treated farmers as well as their peers.

3.7 Conclusion

This study examined the SMS trial on lime promotion conducted by 1AF in Rwanda. Our analysis provides evidence that SMS communication could improve the impact of agricultural extension, increasing the adoption of recommended inputs. More importantly, we show that small differences in message and dissemination designs generate substantially varying effects. In our context, messages that encouraged lime adoption to increase yields were significantly more effective than the messages that encouraged adoption to avoid yield loss. Our findings also provide preliminary evidence that diversifying messages may be an effective strategy for a mobile phone-based promotion where individuals are interacting frequently and undertaking relevant decision-making in a group setting. One productive avenue for future research may be to explore how message content, diversity, and intensity affect the way information flows within different social structures and networks. Insights on these questions would not only contribute to the growing literature on social learning in agriculture, but also inform the design of digital interventions that are rapidly changing the landscape of agricultural extension around the world.

3.8 Tables

	Control	Message diversity and intensity		
		Same message	Diverse messages	Diverse messages (low intensity)
	(G0)	(G1)	(G2)	(G3)
Num. farmers	36,873	35,781	35,921	107,885
Num. phone owners	19,743	18,988	18,821	57,017
Num. groups	3,805	3,678	3,657	11,141
Panel A. SMS treatment				
Assigned to SMS	None	100%	100%	50%
Received SMS	7.89%	100%	100%	58.9%
Mixed treatment (Individual)	7.89%	7.05%	23.1%	17.7%
Mixed treatment (Group)	19.3%	18.8%	34.7%	32.7%
Panel B. Message content				
T1: General promotion		5,005	5,105	15,374
T2: Specific dosage and impact		5,000	5,107	15,340
T3: Self-diagnosis		5,379	5,121	15,297
T4: Soil test		5,199	5,125	15,733
T5: How travertine works		5,103	5,104	15,318
T6: Order immediately		4,827	5,246	15,425
T7: Your cell is acidic		5,268	5,113	15,398
Unit of randomization		Group	Individual	Individual
Panel C. Message framing				
Gain framed		17,939	17,874	27,123
Loss framed		17,842	18,047	26,847
Unit of randomization		Group	Individual	Individual
Panel D. Number of messages				
One message		8,843	8,962	13,478
Two messages		8,858	9,095	13,443
Three messages		9,129	8,982	13,416
Four messages		8,951	8,882	13,632
Unit of randomization		Individual	Individual	Individual
Panel E. Social nudge				
Social nudge		18,309	18,302	53,342
No social nudge		17,472	17,619	54,543
Unit of randomization		Group	Group	Group

Table 1: Summary of experimental treatments[82]

[82] This table presents summary statistics on the experimental assignments. Panel A reports the share of phone owners that were assigned to treatments and that of those that actually received treatments. "Mixed treatment" identifies individuals and groups that received any message to which they were not assigned. Panels B - E summarize treatment assignment across different randomizations.

	Ordered lime (1)	Lime quantity (kg) (2)	Total credit (RWF) (3)	Total enrolled seasons (4)	Group size (5)	# phones per group (6)	Any adopter in group (7)
Control mean	0.039	1.494	19,993.9	1.691	10.944	5.746	0.161
Panel A. Group-level assignment							
I. Message diversity							
G1: Identical	0.002	-0.252	-85.1	-0.006	0.038	-0.026	-0.005
G2: Diverse	-0.007	-0.627	-390.7	-0.013	0.132	-0.042	-0.016**
G3: Diverse (low intensity)	-0.005	-0.428	-226.5	-0.037	-0.006	-0.070	-0.017***
P-val joint F-test	0.184	0.339	0.741	0.406	0.205	0.783	0.032
II. Message content (G1 only)							
T1: General	0.016	-0.104	493.4	0.018	-0.010	0.092	0.029
T2: Specific dosage	-0.009	-0.097	-160.4	0.023	0.018	0.013	-0.015
T3: Self-diagnosis	-0.002	0.080	-228.4	0.023	0.143	-0.088	-0.010
T4: Soil test	0.001	-0.632	-374.9	0.002	-0.027	0.124	0.005
T5: How lime works	-0.002	-0.839*	-952.1	-0.034	0.067	-0.112	-0.017
T6: Order immediately	0.006	-0.251	1,101.6	-0.020	-0.056	-0.051	-0.007
T7: Your cell is acidic	0.001	0.068	-396.1	-0.051	0.119	-0.157	-0.022
P-val joint F-test	0.754	0.467	0.671	0.980	0.974	0.921	0.372
III. Message framing (G1 only)							
Gain framing	0.006	-0.088	-104.3	-0.001	0.069	-0.018	0.003
Loss framing	-0.002	-0.416	-65.8	-0.011	0.006	-0.034	-0.013
P-val jointly 0	0.500	0.507	0.974	0.962	0.768	0.950	0.317
IV. Social nudge							
Social nudge	0.003	-0.305	103.7	0.007	0.090	0.032	-0.004
No social nudge	0.001	-0.197	-280.7	-0.019	-0.016	-0.087	-0.006
P-val joint F-test	0.907	0.763	0.764	0.836	0.585	0.604	0.809
Panel B. Individual-level assignment							
V. Message content							
T1: General	-0.005	-0.417	182.2	-0.018			
T2: Specific dose	-0.008**	-0.551*	75.7	-0.014			
T3: Self diagnosis	-0.002	-0.317	75.9	0.017			
T4: Soil test	-0.007*	-0.443	68.5	0.004			
T5: How lime works	-0.008*	-0.696**	103.3	0.003			
T6: Order immediately	-0.008*	-0.661**	-50.3	-0.013			
T7: Your cell is acidic	-0.006	-0.275	458.1	0.021			
P-val joint F-test	0.240	0.136	0.675	0.478			
VI. Message framing							
Gain framing	-0.006	-0.382	5.1	0.008			
Loss framing	-0.005	-0.539**	171.0	-0.012			
P-val jointly 0	0.344	0.119	0.590	0.379			
VII. Number of messages							
One message	-0.006	-0.350	-163.3	-0.004			
Two messages	-0.001	-0.037	264.2	0.006			
Three messages	0.001	-0.486	-217.6	-0.005			
Four messages	-0.006	-0.527	-68.5	0.012			
P-val joint F-test	0.021	0.024	0.139	0.850			

Table 2: Randomization balance[83]

[83] This table reports the differences in baseline characteristics across experimental groups, estimated in an OLS model. Standard errors are clustered at the farmer group level. Panel A reports the results for message features randomized at the group level. The sample for (II) message content and (III) message framing is limited to G1 and the control group since they were randomized at the individual-level in G2 and G3. Panel B reports the results for the features randomized at the individual level.. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	Ordered lime (1)	Lime quantity (kg) (2)	Enrolled (3)	Total credit (RWF) (4)	Credit for other inputs (RWF) (5)
Panel A. Average effects of SMS messaging					
Get SMS	0.00959***	0.416**	0.00890	408.7	365.1
	(0.00251)	(0.210)	(0.00679)	(288.3)	(284.5)
Observations	86,049	86,048	86,049	86,049	86,048
R-squared	0.028	0.022	0.002	0.133	0.131
Has phone	Yes	Yes	Yes	Yes	Yes
Control mean	0.0523	2.864	0.647	16578	16292
Panel B. Effects of within-group variation in messages					
G1: Identical messages	0.00714**	0.239	0.0110	419.0	388.3
	(0.00320)	(0.247)	(0.00854)	(360.1)	(355.7)
G2: Diverse messages	0.0113***	0.582**	0.0111	478.8	420.6
	(0.00324)	(0.264)	(0.00858)	(360.1)	(354.6)
G3: Diverse (low intensity)	0.00756***	0.296	0.00469	271.2	241.6
	(0.00257)	(0.217)	(0.00702)	(303.2)	(299.4)
Observations	114,569	114,568	114,569	114,569	114,568
R-squared	0.027	0.023	0.002	0.136	0.134
Has phone	Yes	Yes	Yes	Yes	Yes
Control mean	0.0523	2.864	0.647	16578	16292
P-val G1=G2	0.215	0.156	0.988	0.867	0.927
P-val G1=G3	0.873	0.760	0.361	0.621	0.620
P-val G2=G3	0.175	0.170	0.355	0.487	0.543

Table 3: Effects of SMS and message diversity[84]

[84] This table reports estimates of the main effects of the program on lime adoption. All models are estimated using OLS. Standard errors are clustered at the farmer group level. Regressions control for baseline credit amount, group size, and whether or not the group included any farmers who had purchased lime the previous year. The sample frame consists only of farmers who registered a phone number with Tubura in the previous season. The sample for Panel A excludes farmers in G3 that were assigned not to receive SMS messages. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	Ordered lime (1)	Lime quantity (kg) (2)	Ordered lime (3)	Lime quantity (kg) (4)
G1: General promotion (T1)	0.00807	0.271		
	(0.00679)	(0.469)		
G1: Specific dosage and impact (T2)	0.00503	-0.192		
	(0.00590)	(0.361)		
G1: Self-diagnosis (T3)	0.00216	0.0595		
	(0.00589)	(0.407)		
G1: Soil test (T4)	0.00576	0.109		
	(0.00623)	(0.434)		
G1: How travertine works (T5)	0.0148**	0.956*		
	(0.00719)	(0.535)		
G1: Order immediately (T6)	0.00708	0.605		
	(0.00701)	(0.571)		
G1: Your cell is acidic (T7)	0.00738	-0.0986		
	(0.00663)	(0.425)		
G1: Gain-framed			0.0112***	0.441
			(0.00401)	(0.299)
G1: Loss-framed			0.00302	0.0360
			(0.00389)	(0.288)
G2: Diverse messages	0.0113***	0.582**	0.0113***	0.582**
	(0.00324)	(0.264)	(0.00324)	(0.264)
G3: Diverse messages (low intensity)	0.00756***	0.296	0.00756***	0.296
	(0.00257)	(0.217)	(0.00257)	(0.217)
Observations	114,569	114,568	114,569	114,568
R-squared	0.027	0.023	0.027	0.023
Has phone	Yes	Yes	Yes	Yes
Control mean	0.0523	2.864	0.0523	2.864

Table 4: Effects of message content and framing[85]

[85] This table presents estimates of the treatment effects of different message contents and framings on farmers enrolled in Tubura who own a phone. All models are estimated using OLS. Standard errors are clustered at the farmer group level. Regressions control for 2017 credit, group size, and whether or not the group included any farmers who had purchased lime the previous year. *** p < 0.01, ** p < 0.05, * p < 0.1.

	Ordered lime (1)	Lime quantity (kg) (2)
First message	0.00470	0.166
	(0.00290)	(0.249)
Second message	0.00701***	0.307
	(0.00254)	(0.204)
Third message	-0.00201	-0.0550
	(0.00257)	(0.204)
Fourth message	0.00202	0.165
	(0.00262)	(0.218)
Observations	86,061	86,060
R-squared	0.028	0.022
Has phone	Yes	Yes
Control mean	0.0523	2.864

Table 5: Effects of message repetition[86]

[86] All models are estimated using OLS. Standard errors are clustered at the farmer group level. Regressions control for 2017 credit, group size, and whether or not the group included any farmers who had purchased lime the previous year. The sample includes all farmers enrolled in Tubura in 2018 who have a phone and have been assigned either to receive an SMS message or to the control group. Farmers in G3 who did not receive SMS messages are removed from this analysis. Estimates reported in this table are the incremental effect of receiving each additional message. *** p < 0.01, ** p < 0.05, * p < 0.1.

Table 6: Spillover Effects

	No Phone				Has Phone (no message)			
	Ordered lime (1)	Lime quantity (kg) (2)	Ordered lime (3)	Lime quantity (kg) (4)	Ordered lime (5)	Lime quantity (kg) (6)	Ordered lime (7)	Lime quantity (kg) (8)
G1: Identical messages	0.000699	0.0725	0.00127	0.0973				
	(0.00195)	(0.103)	(0.00205)	(0.112)				
G2: Diverse messages	0.00332	0.190	0.00389*	0.214*				
	(0.00208)	(0.117)	(0.00220)	(0.128)				
G3: Diverse (low intensity)	0.00293*	0.260***	0.00349**	0.284***	0.00487*	0.157	0.00587*	0.0930
	(0.00165)	(0.0936)	(0.00177)	(0.0986)	(0.00271)	(0.228)	(0.00312)	(0.250)
Social nudge			-0.00113	-0.0488			-0.00202	0.130
			(0.00127)	(0.0817)			(0.00318)	(0.245)
Observations	101,891	101,891	101,891	101,891	48,263	48,263	48,263	48,263
R-squared	0.015	0.008	0.015	0.008	0.025	0.029	0.025	0.029
Control mean	0.0221	0.931	0.0221	0.931	0.0523	2.864	0.0523	2.864
P-val joint F-test (G1-G3)	0.175	0.0267	0.135	0.0192				
P-val G1 = G2/G3	0.124	0.0478	0.126	0.0486				

Note: All models are estimated using OLS. Standard errors are clustered at the farmer group level. Regressions control for 2017 credit, group size, and whether or not the group included any farmers who had purchased lime the previous year. Columns (1)-(4) report the spillover effects among those who had not registered a phone across all experimental groups; Columns (5)-(8) report the spillover effects among phone owners (G0 vs. phone owners assigned not to receive SMS in G3). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	Ordered lime (1)	Lime quantity (kg) (2)	Ordered lime (3)	Lime quantity (kg) (4)
G1: General promotion (T1)	0.00456	0.348		
	(0.00453)	(0.257)		
G1: Specific dosage and impact (T2)	-0.000702	-0.0417		
	(0.00379)	(0.189)		
G1: Self-diagnosis (T3)	0.00150	0.181		
	(0.00356)	(0.216)		
G1: Soil test (T4)	-0.00354	-0.0944		
	(0.00342)	(0.196)		
G1: How travertine works (T5)	-0.000702	-0.0153		
	(0.00336)	(0.166)		
G1: Order immediately (T6)	0.00111	-0.0604		
	(0.00400)	(0.180)		
G1: Your cell is acidic (T7)	0.00257	0.174		
	(0.00394)	(0.204)		
G1: Gain-framed			0.00282	0.195
			(0.00247)	(0.135)
G1: Loss-framed			-0.00144	-0.0513
			(0.00225)	(0.115)
G2: Diverse messages	0.00332	0.190	0.00332	0.190
	(0.00208)	(0.117)	(0.00208)	(0.117)
G3: Diverse messages (low intensity)	0.00293*	0.260***	0.00293*	0.260***
	(0.00165)	(0.0936)	(0.00165)	(0.0936)
Observations	101,891	101,891	101,891	101,891
R-squared	0.015	0.008	0.015	0.008
Has phone	No	No	No	No
Control mean	0.0221	0.931	0.0221	0.931
P-val jointly equal (G1)	0.774	0.698	0.109	0.0833

Table 7: Spillover effects by message content[87]

[87] This table presents estimates of spillover effects on farmers without phones by message content and framing. All models are estimated using OLS. Standard errors are clustered at the farmer group level. Regressions control for 2017 credit, group size, and whether or not the group included any farmers who had purchased lime the previous year. This table presents estimates of the treatment effects of different message contents and framings on farmers enrolled in Tubura who registered a phone with Tubura. P-values for joint significance within G1 are presented comparing T1 - T7 and gain vs. loss. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

3.9 Appendix

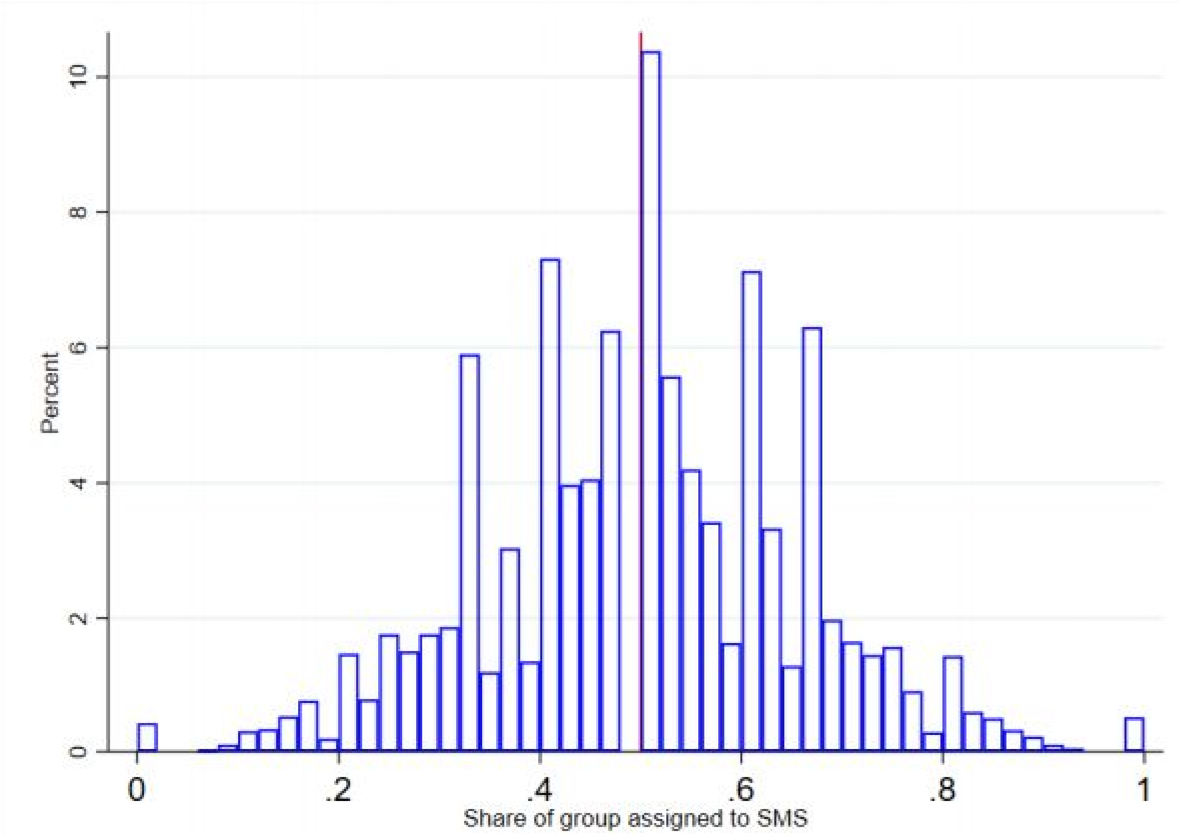
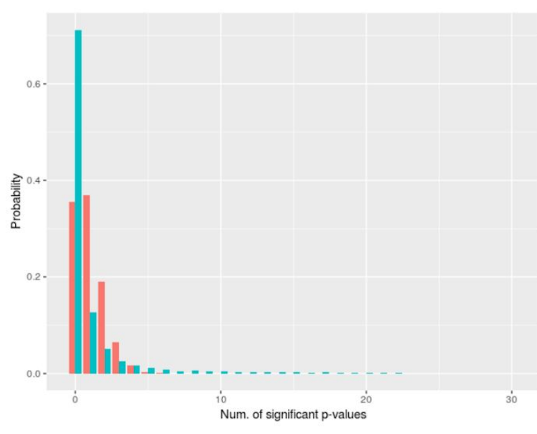
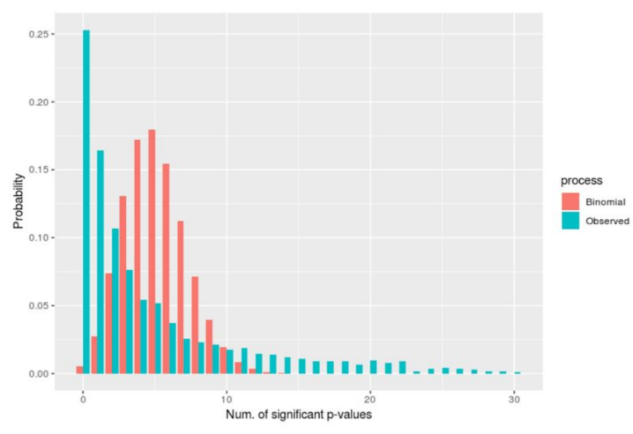


Figure A1: Share of farmers in G3 assigned to receive a message



(a) 1% significance level



(b) 5% significance level

Figure A2: Randomization balance: simulated vs. binomial distribution of significant p-values (out of 103)

Table A1: SMS Content

Code	Description	Gain-framed	Loss-framed
T1	General Promotion	Many fields in Rwanda have acidic soil and need TRAVERTINE to increase yields. Order from TUBURA now.	Many fields in Rwanda have acidic soil and need TRAVERTINE to avoid a yield loss. Order from TUBURA now
T2	Specific dosage and impact	Many fields in Rwanda have acidic soil. Applying 25 kg/are of TRAVERTINE will increase yields by 20%. Order from TUBURA now.	Many fields in Rwanda have acidic soil. Applying 25 kg/are of TRAVERTINE will prevent a yield loss of 20%. Order from TUBURA now
T3	Self-diagnosis	Do you have fields with poor harvests even when you use fertilizer? You probably have acidity and need TRAVERTINE to increase yields. Order from TUBURA now.	Do you have fields with poor harvests even when you use fertilizer? You probably have acidity and need TRAVERTINE to avoid a yield loss. Order from TUBURA now
T4	Soil test	Ask your Field Officer for a free soil test to learn if your fields are acidic and you need to order TRAVERTINE to increase yields.	Ask your Field Officer for a free soil test to learn if your fields are acidic and you need to order TRAVERTINE to avoid a yield loss
T5	How travertine works	Many fields in Rwanda have acidity, which blocks fertilizer uptake. Applying TRAVERTINE solves the problem, increasing crop yields. Order from TUBURA now.	Many fields in Rwanda have acidity, which blocks fertilizer uptake. Applying TRAVERTINE solves the problem, preventing a yield loss. Order from TUBURA now
T6	Order urgency and details	Many fields in Rwanda have acidic soil and need TRAVERTINE to increase yields. Order it immediately, when signing your TUBURA order form.	Many fields in Rwanda have acidic soil and need TRAVERTINE to avoid a yield loss. Order it immediately, when signing your TUBURA order form
T7	Your specific site	In your site the soil is acidic. If you apply 25 kg/are of TRAVERTINE you can boost yields by 20%. Order from TUBURA now.	In your cell the soil is acidic. If you apply 25 kg/are of TRAVERTINE you can avoid a yield loss of 20%. Order from TUBURA now.
SN1	Social Nudge	Please share this information about TRAVERTINE with your group members and neighbors, especially those who don't have phones!	

Note: This table presents the message content for the experiment. Social nudge messages were sent to assigned farmers in a second SMS immediately following the main message.

3.10 References

Casaburi, L., Kremer, M., Mullainathan, S., and Ramrattan, R. (2014). Harnessing ICT to Increase Agricultural Production: Evidence From Kenya. Harvard Business School Working Paper.

Cole, S. and Sharma, G. (2018). The Promise and Challenges in Implementing ICT for Agriculture Introduction and Motivation.

Cole, S. A. and Fernando, A. N. (2016). Mobile'izing Agricultural Advice: Technology Adoption, Diffusion and Sustainability.

Fabregas, R., Kremer, M., and Zane, G. (2018). Leveraging ICT-Extension to Increase Farmer Experimentation: Evidence from Kenya. Unpublished working paper.

Kisinyo, P. O., Opala, P. A., Palapala, V., Gudu, S. O., Othieno, C. O., and Ouma, E. (2015). MicroDosing of Lime, Phosphorus and Nitrogen Fertilizers Effect on Maize Performance on an Acid Soil in Kenya. Sustainable Agriculture Research, 4(2):21.

Owino, K. (2016). Managing Soil Acidity with Lime (2015) Farmers First. Technical report.