Evaluation of the Millennium Challenge Corporation's Electricity-Transmission and Distribution Line-Extension Activity in Tanzania: Baseline Data User's Manual

February 10, 2014

Kathy Buek
Xiaofan Sun
Duncan Chaplin
Arif Mamun
John Schurrer

MATHEMATICA
Policy Research

Contract Number:
MCC-10-0114-CON-20-TO05

Mathematica Reference Number:
06919.325

Submitted to:
Millennium Challenge Corporation
875 15th St., NW
Washington, DC 20005
Project Officer:
Shreena Patel

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Arif Mamun

Evaluation of the Millennium Challenge Corporation's Electricity-Transmission and Distribution Line-Extension Activity in Tanzania: Baseline Data User's Manual

February 10, 2014

Kathy Buek
Xiaofan Sun
Duncan Chaplin
Arif Mamun
John Schurrer

MATHEMATICA
Policy Research

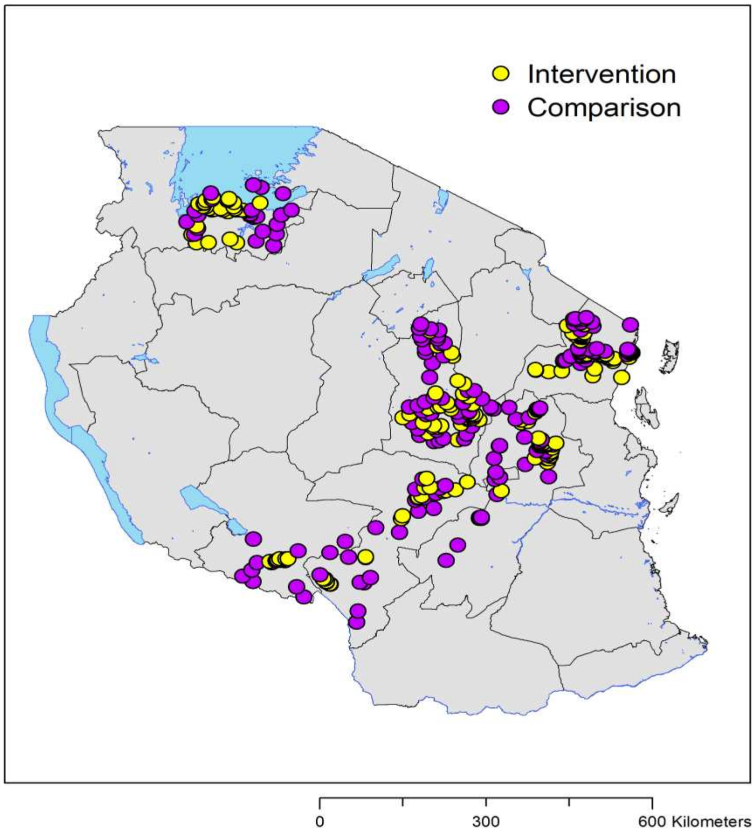
CONTENTS

TABLES

ACRONYMS

FS	Customer-connection financing scheme
GPS	Global positioning system
IGA	Income-generating activity
MCA-T	Millennium Challenge Account—Tanzania
MCC	Millennium Challenge Corporation
NBS	National Bureau of Statistics, Tanzania
NRECA	National Rural Electric Cooperative Association International
PSU	Primary sampling unit
TANESCO	Tanzania Electric Supply Company
T&D	Transmission and distribution systems rehabilitation and extension

TANZANIA MAP: AREAS OF BASELINE DATA COLLECTION



Sources: Tanzania Energy Sector Baseline Household Survey and Global Administrative Areas Database.

I. INTRODUCTION

In an effort to promote economic growth and reduce poverty in Tanzania, the Millennium Challenge Corporation (MCC) is funding an energy sector project that is being implemented by the Millennium Challenge Account–Tanzania (MCA-T). The project has a number of key components, including rehabilitation and extension of the transmission and distribution (T&D) network, a customer-connection financing scheme initiative to facilitate lower-cost electricity connections in selected areas (referred to more succinctly as the financing scheme [FS] initiative), installation of a new submarine cable connecting Zanzibar's Unguja Island to the mainland, and promotion of solar power systems in the Kigoma region of mainland Tanzania. Together, these activities are intended to increase the availability of reliable and high quality electricity to people in mainland Tanzania and Zanzibar.

MCC has contracted with Mathematica Policy Research to carry out rigorous evaluations of the T&D activity and FS initiative.[1] These evaluations are designed to enable MCC to understand more fully how the T&D activity and FS initiative affect the well-being of the target populations. MCA-T contracted with NRECA International (NRECA) to carry out the instrument design, data collection, and data entry for the T&D activity impact evaluation. Mathematica worked closely with NRECA and oversaw data collection from communities, households, and businesses in six regions of mainland Tanzania.

This manual provides information about the sample design, questionnaire design, data collection, data entry and cleaning, and response rates for the surveys, as well as a description of the content and format of the internal use data file that Mathematica submitted to MCC.

II. SAMPLE DESIGN

To provide data for the T&D evaluation, three baseline surveys were implemented: a community survey, a household survey, and an enterprise survey. In this section, we describe the sampling strategies we applied for each of these surveys.

A. Sampling for the Baseline Community Survey

The community survey was conducted in 182 intervention communities and 546 potential comparison communities in six regions. The primary sampling unit (PSU) for the community survey was a village (*kijiji*) in rural areas and a *mtaa* in urban areas.[2] These are the smallest administrative units for which it was possible to develop a sampling frame. The rural and urban communities covered by the community survey were selected in three steps. First, the evaluation team worked with MCA-T and Tanzania Electric Supply Company (TANESCO) to finalize a list of communities (villages or *mitaa*) that are likely to receive new lines; we found a total of 337 communities that are receiving the new lines (Table II.1). Second, we randomly selected 182 of the 337 villages and *mitaa*

[1] Mathematica is also conducting an evaluation of the Zanzibar cable activity, as discussed by Chaplin et al. (2011).

[2] The Swahili word *kijiji* (plural *vijiji*) means village and refers to a rural administrative unit; *mtaa* (plural *mitaa*) translates to "street" and refers to the smallest urban administrative unit. Villages can be further subdivided into subvillages (*vitongoji*, singular *kitongoji*), which is the smallest rural administrative unit. Because the English word "street" could be confusing when referring to a geographic area, throughout this report, we use the Swahili words *mtaa* or *mitaa* to refer to the urban communities in the evaluation. For the rural communities, we use villages and subvillages to refer to *vijiji* and *vitongoji*, respectively.

to represent the intervention communities in the evaluation. This number was chosen to achieve the desired level of precision, as explained in the evaluation design report (Chaplin et al. 2011). Third, we identified 546 potential comparison villages using propensity score matching based on existing data, including 2002 census data global positioning system (GPS) data from the National Bureau of Statistics (NBS) as well as data from TANESCO. Table II.1 shows the number of intervention and comparison communities sampled for the community survey.

Table II.1. Community Survey Sample Size by Intervention Status

	Number of Communities (Villages/*Mitaa*)
Intervention group	182
Comparison group	546
Total	728

B. Sampling for the Baseline Household Survey

Using the data collected in the community survey, we performed a second round of propensity score matching to identify a single comparison community match for each of the 182 intervention communities (for more details on matching, see Chaplin et al. 2012). Then the baseline household survey was conducted in 182 intervention communities and 182 matched comparison communities.

The PSU for the household survey was villages, subvillages, and *mitaa*. In urban areas, we continued to use a *mtaa* as the PSU. In rural areas, when a village had multiple subvillages, we used a subvillage (*kitongoji*) as the PSU; when a village did not have subvillages, we used the village as the PSU. In the community survey, we collected information on the number of subvillages within each village, the number of households in each subvillage, and (in the intervention communities) the estimated proportion of households in each subvillage that would likely be eligible to connect to the planned T&D lines, as reported by the community leaders. We then ranked subvillages according to the number of households in each. In the intervention group, for a village with multiple subvillages, we selected the subvillage with the largest percentage of households eligible to connect to the new T&D lines.[3] In each comparison village with multiple subvillages, we selected a subvillage that was matched to the household population rank of the corresponding intervention subvillage. We did not need to identify a smaller PSU in urban areas because we expected that almost all households will have access in urban areas receiving new lines.

For the baseline household survey, after identifying the communities as the PSU, we sampled households from these communities. For each intervention and comparison community (village, subvillage, and *mtaa*) selected for the baseline household survey, a census of all households residing in the community (also referred to as a household listing) was created. This list included the name and gender of the household head and also identified whether a household was already connected to the grid or near an existing line. Households that were already connected or within range to connect

[3] Here, access to the electricity lines implies that the household is within 30 meters from the new low-voltage lines. Households or businesses within this distance are eligible for connection at a basic rate. Entities farther away must pay for additional poles.

to an existing line were excluded from the household survey sampling frame.[4] The household listing forms are shown in Appendix A.

The remaining households on the list constituted the sampling frame for each community. We separated these households into three strata: (1) comparison group households, (2) intervention group households residing in small houses (estimated to contain no more than two rooms), and (3) intervention group households residing in larger homes.[5] Within each of these strata, we sampled the same fraction of households from each PSU, which meant that we interviewed more households in the larger communities. Table II.2 shows the number of intervention and comparison households sampled (more details on sample size are provided in Section IV, below).

Table II.2. Household Survey Sample Size by Intervention Status

	Number of Households
Intervention Group	4,767
Comparison Group	5,531
Total	10,298

There was a potentially important difference in how the intervention and comparison group household surveys were conducted. The data collection team prepared lists of all households in the sampled intervention and comparison communities; these lists were used to produce the sampling frame for the household survey. In the intervention communities, the list of households was prepared when the community survey was being fielded, which was as long as several months earlier than the household survey; for the comparison communities, it was prepared the day before the household survey was administered.[6] The lag time between the household listing and the survey in the intervention group meant that during the interim some households may have moved out of the community. This would imply that the intervention could potentially have fewer relatively mobile households than the comparison group, and consequently this could affect the equivalence of the two groups at baseline; however, our analysis of the data suggests that differential migration did not lead to differences between the samples in the intervention and the comparison group.

[4] During the household survey, we had to replace seven comparison communities because all households in those communities were within 30 meters of existing lines or were already connected and thus were not eligible for the survey.

[5] The households with smaller houses were being considered for a targeted subsidy pilot activity that was not implemented. We oversampled those households, so 40 percent of the sample selected for the survey was in these smaller homes, compared to 25 percent in the sampling frame.

[6] The difference in the timing of the household listing in the intervention and comparison communities occurred for a number of reasons. The community and household surveys were conducted in all intervention group communities. Consequently, for the intervention group, NRECA was able to carry out the household listing and the community survey at the same time. Moreover, we needed to identify households that resided in small (no more than two rooms) versus large houses for a planned subsidy pilot activity in the intervention communities, so that we could oversample subsidy-eligible households. As a result, the listing of households in the intervention communities had to be carried out long before the fielding of the household survey. In contrast, for the comparison group, the community survey was conducted in three times as many communities as the household survey (546 versus 182), and data from the community survey were used to select the 182 communities where the household survey was administered. Thus, comparison communities where the household survey was carried out were not identified at the time of the community survey. Consequently, we decided not to do the household listing at the same time as the community survey; this required fewer resources than would have been needed to list households in all 546 comparison communities.

C. Sampling for the Baseline Enterprise Survey

The baseline enterprise survey was conducted only in the Tanga region.[7] The target sample size for the survey was 32 enterprises in seven intervention communities and another 32 enterprises in seven comparison communities. The communities where the enterprise survey was administered were selected randomly from all intervention and comparison communities in the Tanga region. We listed all stand-alone businesses (businesses not occupying the same premises as a household) in each community, regardless of whether they were already connected to or within connecting range of an existing power line. The listing form used to create the sampling frame of enterprises eligible for the survey is included in Appendix B.

There were two types of enterprises included in the sampling frame for the enterprise survey: those that are not connected to the grid and those that are already connected. In the intervention group, enterprises in the first set (not connected to the grid) were sampled with certainty because of the small number of them in the sampling frame; the data collection team randomly selected from the remaining enterprises in the sampling frame (that were not connected to the grid) to meet the target sample size. In the comparison group, sampling was driven to match the number of enterprises in each set (not connected and already connected). Considering the relatively small sample size of the enterprise survey, the enterprise study is being considered as a case study only. Table II.3 shows the number of intervention and comparison enterprises sampled in Tanga.

Table II.3. Enterprise Survey Sample Size by Intervention Status

	Number of Enterprises
Intervention Group	32
Comparison Group	32
Total	64

III. QUESTIONNAIRE DEVELOPMENT AND DATA COLLECTION

Three questionnaires were developed for the baseline data collection: a community questionnaire, a household questionnaire, and an enterprise questionnaire. The community questionnaire captured information about community composition, access to water and electricity, access to civil services and transportation, and business activity and markets. The household questionnaire contained questions about household composition, health and education outcomes, energy and electricity use, household income and expenditures, time use, and assets. The enterprise survey collected data about business characteristics, capital investments, labor inputs, revenues and expenditures, and energy use (electric and non-electric). The final questionnaires are included in Appendix C.

The community, household, and enterprise questionnaires were developed by NRECA based on guidance provided by Mathematica, MCC, and MCA-T regarding the outcomes and indicators of interest. Mathematica also provided NRECA with examples of existing survey items from energy surveys that had been carried out in Tanzania and other developing countries by the World Bank,

[7] The enterprise survey was implemented in the Tanga region because the region was expected to have larger businesses. However, the communities in the Tanga region where the enterprise survey was administered were selected based on the household survey sampling, and there were no large businesses in those communities.

MCC, and others, as well as Tanzanian income and enterprise surveys conducted by the NBS, including the National Panel Survey, Demographic and Health Survey, and the Household Budget Survey. Items and formats from these questionnaires were adapted and tailored to the needs of the T&D evaluation, and additional items were created to ensure that all research questions were adequately and accurately addressed in the surveys.

A. Community Questionnaire

The purpose of the community questionnaire was to capture key background information about the community and also collect information about other community-level characteristics that might mediate, or confound, the effects of the interventions observed at the household and enterprise levels. Additionally, this information was used for the purposes of matching intervention and comparison communities. The community questionnaire was administered to community leaders—the village/*mtaa* chairperson, executive officer, or council member—in a 30-minute in person interview that was usually held in the village/*mtaa* government office. The respondent was encouraged to use government records or statistics, when available, in answering survey questions. The community survey consists of the following modules:

- Subvillage information
- Background characteristics
- Transportation, communication, and water supply
- Access to electricity
- Civil services
- Development projects
- Health services
- Business activity
- Energy/fuel prices

B. Household Questionnaire

The purpose of the household questionnaire was to capture information about baseline characteristics of the household as well as the household's socioeconomic status and energy consumption—the main outcomes of interest for the T&D evaluation. The questionnaire was designed to capture information primarily at the household level, with "household" defined as a group of people (related or not) living, eating, and sharing expenses together. The questionnaire was designed to be administered to the female spouse of the male head of household or the female head of household if there was no male present.[8] However, the pre-test revealed that female respondents often required the assistance of their male counterparts to answer some of the questions in the questionnaire. Thus, the questionnaire was divided into sections that were to be completed by either the female respondent alone, the female respondent with male assistance, or the male respondent alone. This strategy ensured that both female and male respondents were able to respond for

[8] The male head of household was defined, according to tradition in Tanzania, as the male with primary decision-making authority in the household.

themselves on questions relating to work and income (which are unreliable when reported by a proxy), and that both had a measure of privacy when reporting on these items. This enabled us to capture a more accurate picture of gender-specific income contributions to the household, and provided important information for the evaluation to be able to assess differences in outcomes by gender, as gender is an area of strategic priority for MCC. In households where no female head was present, the entire questionnaire was administered to the male head of household. The household survey consists of the following modules:

- Characteristics of household members (female respondent)
- Health (female respondent)
- Household electrical and non-electrical energy devices and appliances (female respondent)
- Household-owned businesses/income-generating activities (female respondent)
- Household consumption and expenditure (female respondent)
- Use of telephones (female respondent)
- Time use of household members (female respondent)
- Household assets and non-wage income (female and male respondents)
- Wage income (female and male respondents)
- Household energy use (female and male respondents)
- Use of electricity (female and male respondents)
- Wage income—male/spouse (male respondent)
- Time use—male/spouse (male respondent)
- Businesses/income-generating activities—male/spouse (male respondent)

C. Enterprise Questionnaire

The purpose of the enterprise questionnaire was to collect information on baseline characteristics of businesses located in the intervention and comparison communities, with regard to revenues, expenditures, and energy use. This information will be used in assessing the impacts of the program activities on business income and growth. The enterprise questionnaire was based largely on the household questionnaire and was modified and updated to address the specific characteristics (such as sources of finance, inputs, and business revenue) of the businesses that were located in our sampling areas (the same areas where the household survey was conducted). The target respondent for the enterprise survey was the owner of the business, if present, or the day-to-day manager/operator on site if the owner was not present. The enterprise questionnaire consists of the following modules:

- Basic characteristics
- Operation, net asset, and capital investment
- Sources of finance
- Non-energy inputs and enterprise revenue

- Use of hired labor
- Energy use (non-electricity)
- Use of electricity
- Electrical and non-electrical energy devices and appliances
- Mobile phones for enterprise purposes

All three surveys were drafted in English and then translated into Kiswahili. Once the questionnaires were translated, they were tested through a pilot data collection effort carried out by NRECA. The pre-test exercise for the community survey was conducted in February and March 2011 in one rural and one urban community (located in Morogoro Region and Dwani Region, respectively) that were not part of the study sample. The protocol for household listing was pilot tested in these same communities to ascertain its feasibility and the amount of time it would take. The pre-testing of the household and enterprise surveys was carried out in June 2011 in a rural community of Morogoro Region with 30 households and four enterprises and in an urban community of Dodoma Region with another 30 households. Based on the pre-tests, several changes were made to the instruments including clarification of instructions, corrections to translations, re-ordering of modules to improve the flow of the interview, and removing some items to reduce the length of the interview.

IV. DATA COLLECTION AND DATA ENTRY

NRECA International was responsible for hiring and training data collectors and supervisors; conducting household and enterprise listing; data collection for the community, household, and enterprise surveys; and entering and cleaning all data. The respective trainings, led by NRECA staff in Kiswahili, introduced data collectors to the purpose and design of the T&D evaluation, provided in-depth explanation of the questionnaire items and interviewing techniques, explained important aspects of research ethics, and allowed trainees the opportunity to practice administering the instruments in mock interviews and group practice sessions.

Using the sampling strategy described in Section II, three baseline surveys were conducted at the community, household, and enterprise levels to support the evaluation of the T&D activity. The community survey was conducted first, over a seven-week period in April and May 2011. Data collection for the household and enterprise surveys started in August 2011. The enterprise survey, a much smaller data collection effort, was completed within three weeks, in September 2011. The household survey required a total of 14 weeks of field work and was completed in November 2011. Table IV.1 below summarizes the purpose, respondents, target sample size, and timing of the three surveys.

Table IV.1. Purpose, Respondents, Target Sample Size, and Timing of Baseline Surveys for the Tanzania Energy Sector Evaluation

Survey	Purpose	Regions	Target Sample Size	Respondent	Start and End Date
Baseline Community Survey	Collect community-level data at baseline; also used to identify matched comparison communities for the T&D evaluation	Dodoma, Iringa, Morogoro, Mbeya, Mwanza, Tanga	182 intervention, 546 comparison communities	Community leaders	April 18– May 28, 2011
Baseline Household Survey [a]	Collect baseline data on households for the T&D and subsidy pilot evaluations	Dodoma, Iringa, Mbeya, Morogoro, Mwanza, Tanga	11,648 households in 182 intervention and 182 comparison communities	Key female and male members of household	Aug 15– Nov 20, 2011
Baseline Enterprise Survey	Collect baseline data on enterprises for the T&D evaluation	Tanga	32 intervention and 32 comparison enterprises	Owner/ operator of the business	Aug 15– Sep 3, 2011

[a] All households in the sampled intervention communities were listed, and information on eligibility for a planned subsidy pilot activity was completed, when the baseline community survey was administered in April–May 2011. This list was used to produce the household survey sampling frame for the intervention group.

The baseline community survey was administered as planned and data were collected from all but three comparison communities where seasonal rains rendered villages inaccessible to data collection teams (that is, from 725 communities). For the baseline household and enterprise surveys, interviews were completed with a total of 59 businesses and 10,298 households. Table IV.2 shows the response rates for all three surveys.

Table IV.2. Baseline Survey Response Rates by Treatment Group

	Target Sample Size		Completed Interviews		Response Rate	
	Intervention	Comparison	Intervention	Comparison	Intervention	Comparison
Community Survey	182	546	182	543	100%	99.45%
Household Survey	5,824	5,824	4,767	5,531	81.85%	94.97%
Enterprise Survey	32	32	32	27	100%	84.38%

V. DATA CLEANING

The raw data files provided by NRECA contained data for all completed interviews from the community, household, and enterprise surveys. For the household survey, raw data files were exported from CSPro in several segments. These segments were reformatted and combined by Mathematica to produce a single case for each household. In addition, the survey variables for household and enterprise surveys were cleaned to recode numeric missing code and logical skips to lettered indicators. The community survey variables were not recoded for missing or skips.

For data from the community and household surveys, some observations were excluded from the final constructed files for a number of reasons (Chaplin et al. 2012). We excluded 361 potential comparison communities from the analysis, as these communities were not selected as matched comparison communities for the evaluation. Also, four intervention communities were excluded from the baseline analysis because they were no longer receiving new lines. For data from the household survey, we dropped 88 intervention households with completed surveys from the baseline analysis for a number of reasons (Table V.1). Of these 88 households, 38 were dropped because they were in the four intervention communities that were not receiving new lines. Another 41 intervention households were dropped because they could not be matched to the household listing.[9] Six other households were dropped because they were duplicates. The sample size before household-level matching was 10,213 households from 178 intervention and 182 comparison communities. We then dropped three more intervention group households after conducting propensity score matching at the household level because we could not find suitable matches for them in the comparison group. Thus, sample size for the baseline analysis of household survey data is 10,210 households.

Additional adjustments and assumptions were made during analysis of the household survey data to account for inconsistencies in the data and/or errors in translation that were discovered after the completion of the field effort. A discussion of these adjustments is included in Appendix C.

Table V.1. Baseline Household Survey: Matched Intervention and Comparison Sample Versus Data from NRECA

	Intervention Group		Comparison Group	
Region	Total Number of Villages/*Mitaa*	Number of Households Interviewed	Total Number of Villages/*Mitaa*	Number of Households Interviewed
Data from NRECA	182	4,767	182	5,531
Not receiving new lines under the T&D activity	4	38	0	0
Could not be merged to household listing	0	41	0	0
Duplicate records	0	6	0	0
Not matched in propensity score analysis	0	3	0	0
Matched Sample for T&D Evaluation	**178**	**4,679**	**182**	**5,531**

Sources: NRECA (2012) and Tanzania Energy Sector Baseline Household Survey

[9] We needed to merge the household survey data with the household listing in order to calculate the households' selection probabilities, which depended on eligibility for the subsidy pilot intervention as estimated during the household listing.

VI. FILE CONTENT AND SPECIFICATIONS

There are six data files containing data collected during the baseline surveys and the variables constructed for the baseline analyses:

1. Household survey file
2. Household constructs file
3. Enterprise survey file
4. Enterprise constructs file
5. Community survey file
6. Community constructs file

The "survey" files contains survey items in the questionnaires, while the "constructs" files include constructed variables used in the analysis only, plus the identifying variables (MPRVID and FORM_NO) and the weight variables, if applicable. The survey files and constructs files can be merged using the MPRVID and FORM_NO, which together perform as the unique identifiers for each observation in the files.

The survey files have been processed and cleaned based on raw survey data files provided by NRECA. Processing and cleaning has been done in the following ways:

- Cases identified as survey incompletes in the raw data files from NRECA have been dropped from the survey files. Cases were identified as incompletes when all three survey results variables were not equal to 1 (RESULT_1ST not equal to 1 and RESULT_2ND not equal to 1 and RESULT_3RD not equal to 1). The numbers of observations of complete cases in the survey files are 725 for the community survey, 10,298 for the household survey, and 59 for the enterprise survey.
- The community and household survey files each include an "InAnalysis" flag to indicate observations that are excluded from the analysis for the T&D baseline report (Chaplin et al. 2012). In the community file, InAnalysis = 1 if the community was included in the baseline analysis; InAnalysis = 2 if the community was excluded from the baseline analysis because it was not receiving new lines as of August 1, 2012; and InAnalysis = 3 if the community was excluded from the baseline analysis because it was not covered by the household survey. In the household file, InAnalysis = 1 if the household was included in the baseline analysis; InAnalysis = 2 if the household was excluded from the baseline analysis because it was in a community not receiving new lines as of August 1, 2012; and InAnalysis = 3 if the household was excluded because it was a duplicate record, the household could not be identified in the listing file (possibly because of new households moving into the community after the sample listing was created), or the household did not match with a comparison household.
- Certain variables in the survey files have been excluded. Some have been dropped because they contain identifying information, like person and village names. Others have been excluded because they are survey administration variables, like data entry dates and interviewer ID.
- Survey variable names correspond to the sections and the numbering of the questions in the survey instruments.

- The survey variables for household and enterprise surveys have been cleaned to recode numeric missing code and logical skips to lettered indicators (., .M, .N for "missing data"; .D for "don't know," and .S for "logical skips"). The community survey variables were not recoded for missing or skips, as these types of data cleaning were done at the stage of creating the community constructs for analysis. In the community survey variables, any numeric value that contains only 9s or 98s may indicate a missing value, especially if it is the maximum value.
- "TZ_HH_Base_v1a.sas," "clean_enterprise.sas" and "06919_Baseline_Community_ Measures_Prep.do" document cleaning for the household survey file, the enterprise survey file, and the community survey file, respectively.

The "constructs" files includes constructs created based on the survey items.

- Names of the construct variables have distinctive prefixes for each file: "cbc_" for community constructs, "cbh_" for household constructs, and "cbe_" for enterprise constructs.
- In the household construct file (N = 10,210 for InAnalysis = 1), "matchwt" is the final weight variable used in the baseline analysis and "fwt" is the weight to adjust for nonresponse and sampling. Neither the community construct file (N = 360 for InAnalysis = 1) nor the enterprise construct file (N = 59) has weight variables. For more information on the weights, see Chaplin et al. (2012).
- Creation of the community constructs are documented in the program file "06919_Baseline_Community_Measures_Prep.do." "TZ_HH_Base_v1b.sas," "TZ_HH_Base_v1c.sas," and "TZ_HH_Base_v1d.sas" document how the household constructs were created and "Constructs_enterprise.sas" documents the enterprise constructs.

All survey and construct variables can be found in the accompanying codebooks. Entries for each variable include the variable name, question text, universe, variable type, mean, minimum and maximum value for numeric variable, frequency of binary and categorical variables, and number of nonmissing responses.

VII. DATA ANONYMIZATION

We reviewed the risks to the rights and privacy of individual respondents to the baseline surveys and determined the appropriate data anonymization and data access strategies that balance the need to minimize such risks with MCC's need to be transparent and provide adequate access to the data for policy research and the public good. The draft guidelines for data documentation and anonymization from MCC, as well as additional consultation with MCC, guided our review and decisions about potential risks and the level of anonymization for each survey. In the process, we took into account issues such as the likelihood of compromise of the data and the value and potential uses of the data if compromised. We discuss these risks and risk mitigation strategies for each survey in this section.

A. Baseline Community Survey

The baseline community survey was administered to community leaders (for example, village or *mtaa* council chairs and members). The survey data included geographic identifiers such as region, district, ward, village/*mtaa*, subvillage, and GPS coordinates, as well as individual identifiers such as name and position of the community leaders who responded to the survey. We expect almost no privacy risks for individuals given that the survey questions are about the community and not about individuals. Nevertheless, to make it more difficult to identify a community, and thereby, minimize the risk to privacy of individuals in these communities, we removed all geographic identifiers (except region) and all respondent identifiers (name and position) from the data file submitted to MCC. We assigned an identification number for each community (MPRVID) to facilitate merging these data with the non-anonymized household data for users who may gain access to those data under license from MCC. It is unlikely that someone could glean any individual-level information from the remaining anonymized data available in the community survey and construct files; therefore, the risk to the rights and privacy of individuals is very low. Consequently, no additional data reduction or perturbation strategies were applied to the community survey data files.

B. Baseline Household Survey

The baseline community survey was administered to female and male heads of the household. The survey data included geographic identifiers of the household such as region, district, ward, village/*mtaa*, subvillage, and GPS coordinates as well as individual identifiers such as head of household name, address, telephone number, name of other household members, and contact information for the reference person for contacting the respondent in the future for a follow-up survey. As part of the standard procedure to limit risks to privacy of the household survey respondents, we removed all direct geographic identifiers (except region) and individual identifiers from the data files submitted to MCC.[10] However, given the relatively large amount of information collected on a comprehensive set of household characteristics and outcomes, it might still have been possible to identify individual households and their businesses with the remaining information in the data set. In particular, it might be possible to identify individual households using outlier values of continuous variables, especially those related to financial well-being. Consequently, for potentially sensitive continuous variables in the survey and constructs files, we top/bottom coded the top/bottom 100 cases with the mean response of those cases and dropped such variables if less than 100 households had nonmissing values.[11] A list of the top-/bottom-coded variables is provided in Appendix D, Tables D.1 and D.2, and a list of the variables that were dropped from the survey file is shown in Appendix D, Table D.3.

While we do not include any direct geographic identifiers in the household survey, we do include a randomly generated community identifier. This will enable users to calculate variation within and between communities, and to control for clustering by community. In the non-

[10] This includes the variables identifying which communities were covered by the financing scheme initiative and which will be receiving new lines as part of the T&D activity. Once the follow-up data are available we will explore options for producing anonymized data that includes these identifiers so that users can estimate impacts of these activities. This is less important for the baseline data.

[11] We use 100 cases because this is approximately 1 percent of our sample. MCC guidelines mention the use of percentage rules for top and bottom coding. We use a rule based on a number rather than a percentage because some variables in the raw data are missing for most households but might still be used for identifying individual households—for example the amount of debt.

anonymized version of the data, we include a crosswalk from the randomly generated community ID to the MPRVID variable so that users who spent time processing the anonymized version of the household survey data can easily link that to the community survey data if/when they get access to the non-anonymized version of the household data.

The steps implemented and described above to anonymize the survey and constructs data files for the household survey reduce the risks to the rights and privacy of individuals but do not eliminate them. It may still be possible for a persistent intruder to use the remaining information in the data files to individually identify a respondent. In particular, in our effort to anonymize the data sets, we did not thoroughly investigate whether unique and rare responses to some categorical items would permit identification of individual households or household members. As a result, there may exist some situations where individual households could be identified using a combination of responses to several items. However, the chance of this happening is low for a number of reasons. First, the potential benefits of doing such identification are limited because we have top/bottom coded the sensitive financial data an intruder might want in order to benefit financially from such information. Second, while our household sample size is large, it covers a small fraction of households in the communities covered by our household survey and a very small fraction of all households in the regions covered by our survey. Hence, it would only be potentially valuable for intruders if they happened to know that the household they wanted information on was covered by this survey. Third, while some households might stand out for a given set of characteristics within a given community, it is far less likely that an intruder could rule out the possibility that other households with the same characteristics might have been covered by our survey in other communities. Fourth, even under ideal conditions, deductively identifying individual households would require considerable effort on the part of an intruder. Therefore, while the risk of deductive disclosure of identity of individual subjects is not completely removed, it is reasonably low.

Considering the low risk of violation of rights and privacy of individual respondents, and the need to invest substantial amount of resources for identifying and anonymizing unique and rare responses to categorical items, we believe the data reduction strategies implemented for the household survey data files are adequate. Nevertheless, the files may contain indirect identifiers. Consequently, MCC may want to consider giving only licensed-use access to some of these data, especially the survey household data in the "survey" file. After discussing with MCC's Disclosure Review Board, it was decided that MCC will make the anonymized version of the household survey and construct data files available to the public.

C. Baseline Enterprise Survey

The baseline enterprise survey was administered to the business' owners, managers, or operators. The enterprise survey data included geographic identifiers such as region, district, ward, village/*mtaa*, subvillage, and GPS coordinates, as well as individual identifiers such as business name, business address, business telephone number, owner's/manager's name, owner's/manager's phone number, and name and contact information for the reference person for contacting the respondent in the future for a follow-up survey. To limit risks to privacy of the enterprises included in the survey, we removed the GPS coordinates and individual identifiers from the data files submitted to MCC. However, given the relatively small sample size of 59 enterprises who responded to the survey, it might still be possible to identify individual respondents and their businesses with the remaining information in the survey and construct data files. Considering the sensitive nature of some of the information on the businesses (such as sources of finance, cost of inputs, and business revenue), and the possibility of unassisted identification of individual respondents and their businesses, Mathematica recommended and MCC agreed not to release the enterprise survey data as public use data. Consequently, no additional data reduction or perturbation strategies were applied to the enterprise data files.

REFERENCES

Chaplin, Duncan, Arif Mamun, and John Schurrer. "Evaluation of the Millennium Challenge Corporation's Electricity-Transmission and Distribution Line-Extension Activity in Tanzania: Baseline Report." Washington, DC: Mathematica Policy Research, November 20, 2012.

Chaplin, Duncan, Arif Mamun, Thomas Fraker, Kathy Buek, Minki Chatterji, and Denzel Hankinson. "Evaluation of Tanzania Energy Sector Project: Updated Design Report." Washington, DC: Mathematica Policy Research, March 16, 2011.

NRECA International. Household and Enterprise Surveys Data Collection Completion Report. Report submitted to the Millennium Challenge Account–Tanzania. Arlington, VA: NRECA International, Ltd., April 9, 2012.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC