Call for expression of interest of qualified individuals for cleaning and analyzing a medium sized data set in STATA

1. Background

1.1 Integrated Seed Sector Development Project

The Integrated Seed Sector Development (ISSD) Plus project is a four year project coordinated by Wageningen Centre for Development Innovation and funded by the Kingdom of the Netherlands, Kampala. In Uganda Wageningen UR Uganda (WUU) implements the project in collaboration with National Agricultural Research Organisation (NARO) for public varieties and food crops and the private sector for vegetable seed. The programme aims to strengthen the development of a vibrant, pluralistic and market-oriented seed sector that is able to address key challenges that hamper the seed sector development in Uganda. ISSD Plus project has four components: a) promotion of uptake of quality seed, b) enhancing the Quality Declared Seed system through supporting Local Seed Businesses, c) addressing bottlenecks in early generation seed (EGS) and creating an enabling environment for the seed sector, and d) promote the use of advanced vegetable varieties.

1.2 The quality seed uptake component of the project

Under component a) above, the ISSD Plus project aims that by the end of the project 300,000 smallholder farmers use/buy quality seed in the North, South West, West Nile, South Western Highlands, Western Highlands and East of the Country. Interventions under the component assumed that the causes of low quality seed are i) lack of awareness on availability of quality seed; ii) real or perceived lack of quality seed available at convenient locations; iii) lack of cash to purchase seed and; iv) lack of knowledge of economic benefits of investing in quality seed. The interventions to be implemented will therefor include; i) increase awareness on quality seed and its benefits; ii) increase access to quality seed in diversified locations; and iii) increase effective demand for seed by smallholder farmers.

1.3 Promoting drought tolerant maize varieties

In a bid to increase access to affordable quality seed of preferred improved crop varieties ISSD-Uganda is doing a research project to investigate the adoption for drought tolerant maize varieties looking into barriers for uptake of advanced seed technology. Specifically focussing on downside risk and willingness-to-pay for hybrid drought tolerant maize varieties. ISSD Uganda implemented an impact analysis using a randomised controlled trial to measure the effect of awareness campaigns and proximity marketing on farmers knowledge of quality seed and seed purchase behaviour. Information of 2000 respondents have been collected in three rounds, a baseline before the intervention, an end-line after the intervention and follow up session via telephone.

2. Scope of the assignment

2.1 Objective of the assignment

The goal of data cleaning is to clean individual data points and to make the dataset easily usable and understandable for the research team and external users. The second goal of the data cleaning is to code and document the dataset to make it as self-explanatory as possible. The overall objective of this assignment is to create a Master Do file using Stata 13 (or higher version saving the work under version 13).

The consultant will build on the work already done by the ISSD team, which includes recoding of all variables in the baseline data set and programming of tables.

2.2 Description of tasks

Task 1: Create a dictionary for all variables in the baseline data set, endline data set and follow-up data set (variable name, link to questionnaire number, including coding for baseline, endline, follow-up, type of data format, descriptions of numeric codes).

Task 2: Prepare data cleaning and re-coding for all three data sets in separate do files; a) baseline separate (split the do file into two) variable recoding from the statistical testing and other actions in the already prepared do file. Check all recoding and proceed with the data data cleaning, b) clean and recode end-line database, using the same variables as baseline where necessary for easy reference; c) clean and recode follow-up survey database. Data cleaning should include anything that biases a mean will bias a regression: outliers, missing values, typos, erroneous survey codes, illogical values, duplicates, etc. See appendix for more details on data cleaning requirements. See annex 1 for more details.

Task 3: merge the three data bases into one database using a separate do-file. Perform checks and balances to make sure the merging is done properly. Use assert etc to verify.

Task 4: Verify if sample is balanced on key variables and create all output tables that are described in annex 2. Do this in a separate do-file and create sub folder for the output tables and graphs and number them appropriately.

2.3 Method of work

The consultant will use his/her own STATA software and use the dropbox location so that the work and progress is accessible to the ISSD team and easy to review progress.

Organization of the folder: Do-files, datasets, graphs, logs, tables, pdfs, etc. should be saved in different sub-folders. Each table, graph etc should have a unique number that refers to the data set and do file in which it is created and the numbers should be written in the do file as well for easy reference.

Master file: The first thing to do is to write a do-file which is not really a do-file, but a master-file. This is a do-file that calls other do-files. The purpose of the master file is to organize the sequence in which your do-files are executed and it describes what each do file does and why.

Your do-file should be full of comments. The beginning of your do-file should be a description of what it does. Do not hesitate to be very descriptive. Include comments at each important step or operation that your do-file does. Explain why you are getting some variables from other data files, or why you created such a variable, etc. Adding comments can also helps to find errors in codes. Once you are done with writing your do-file, there could (should) be more comments than actual coding.

For every single do-file that you write, you should open a log file in order to keep track of everything you're doing. Ideally, your do-file should start by opening a new log-file, and end with closing this log file.

The consultant will create separate do files for the different tasks and payment will be based on completion and delivery of the tasks.

2.4 Deliverables

The consultant will deliver the following data sets and do-files and will be paid per completed task using the methodology described in section 2.3 and annex 1 and 2. It is anticipated that the assignment takes no more than 25 days and is divided over the deliverables as follows:

- Deliverable 1: detailed STATA do-file for the data cleaning and recoding and dictionary (in word) dictionary and data cleaning and re-coding as described in task 1 and 2. The deliverable is equivalent to a payment of 16 days
- Deliverable 2: STATA do-file with merged data sets and error and quality checks as described in task 3. The deliverable is equivalent to a payment of 2 days.

- Deliverable 3: Do file and exported tables as described in task 4. The deliverable is equivalent to a payment of 7 days.

2.5 Payment schedule and accountability

No advance payment will be provided and is based on a 'no cure, no pay' principle. Payment is based on submission of the deliverable(s) and an invoice. This consultancy is a fixed pay contract based on the number of days stipulated under deliverables and agreed fee.

The assignment has to be completed within 2 months from start date as stipulated in the contract.

3. Requirements: **Qualification of a consultant**

- Proven experience with STATA, including cleaning, checking, merging data sets, preparing do-files, exporting tables and figures (automatically)
- In possession of a laptop with STATA
- Access to dropbox and ability to work with Dropbox folders on laptop
- Access to internet to work online and synchronize the files in the dropbox folders continuously.

Contact details

The consultant will report directly to:

- Admin manger on contract and administrative matters
- Ms Astrid Mastenbroek, principal researcher, and Emma Letaa, Ag economist ISSD Uganda on progress of the assignment and quality checks

Annex 1 – Data cleaning

Source: https://dimewiki.worldbank.org/wiki/Data_Cleaning

ID Variables

Observations in the dataset should be uniquely and fully identifiable by a single ID variable. Often, raw primary data includes duplicate entries. Carefully document these cases. To ensure accuracy, only correct them after discussing with the Ag Economist and PI what caused them. ieduplicates, a command in ietoolkit identifies duplicated entries, while iecompdup helps to correct them. Once duplicates are corrected, the observations can be linked to the master dataset.

Outliers

While there are many rules of thumb for how to define an outlier, there is no silver bullet. Some consider an outlier to be any data point that is three standard deviations away from the mean of the same data point for all observations. This may be a starting point, but one needs to qualitatively consider if this is a correct approach. Approaches to outliers include, but are not limited to:

1. Replacing the outlier values with a missing value.
2. Winsorization, or replacing any values bigger than a certain percentile, often the 99th, with the value at that percentile. This prevents very large values from biasing the mean. It also maintains an equality of impact aspect. For example, if all project benefits go to a single observation in the treatment group, then the mean would still be high, but that is rarely a desired outcome in development. Winsorization thus penalizes inequitable distribution of the benefits of a project.

The consultant will identify outliers (being 3 standard deviations away from the mean) and document the outliers in the do-file.

Illogical Values

In theory, good questionnaire design should include logic checks that prevent illogical values. For example, if a respondent is male, then the questionnaire should not allow the respondent to answer that he is pregnant. However, no questionnaire ever can be pre-programmed to control for every such case. Check for illogical values and ddiscuss with the research team on the best approaches to illogical values found in the raw dataset.

Typos

If it is obvious beyond any doubt that the response is incorrect due to a simple typo, then correct the typo using the do-file. Make sure to document the change in a replicable way.

Survey Codes and Missing Values

Almost all data collection done through surveys of any sort allows respondents to answer something like "Do not know" or "Decline to answer" for individual questions. These answers are usually recorded using survey codes in the format -999, -88 or something similar. If left as such, these numbers will bias means and regressions. Accordingly, they must be replaced with missing values in Stata.

Stata has several missing values. The most well-known is the regular missing value represented by a single "." but Stata also offers extended missing values: ".a", ".b", ".c" etc. all the way to ".z". Stata handles these values the same as "." in commands that expect a numeric value. Conveniently, these extended missing values accept value labels that allow you to distinguish between, for example, "Do not know" and "Decline to answer." You might label ".d", for example, as "Decline to answer", and ".k" as "Do not know." Make sure to always assign value labels to extended missing values so that they can be precisely interpreted. Finally, make sure to consistently use the same letter ".a", ".b" etc. to represent only one response across your project. See Stata Manual Missing Values for more details on missing values.

Missing values can be used for much more than just survey codes. Any value that we remove because it is incorrect should be replaced with a missing value. In a master dataset, there should be no regular missing values. All missing values in a master dataset should contain an explanation of why there is no information for that value.

Use: . for missing answers due to skip logic

Use: .d for decline to answer

Use .k for don't know

Use .n for not applicable

Strings

All data should be stored in numeric format because (1) Stata stores numbers more efficiently than strings and (2) many Stata commands expect values to be stored numerically. During the data cleaning process, make sure to clean categorical string variables and convert them into numeric codes. Then assign value labels for clarity. The commands destring and/or encode may be useful during this process.

There are two exceptions in which string variables are acceptable:

1. If the number cannot be stored correctly numerically. This may occur in two scenarios:
    1. If the number is more than 15 digits long. For obvious reasons, an ID cannot be rounded and may remain a string. However, if a continuous variable has more than 15 digits, round it and convert it to a different scale. After all, a precision of 16 digits is not even possible in natural sciences.
    2. If the number begins with 0, as is sometimes the case for national IDs and telephone numbers. In this case, continue storing the number as a string, as Stata would remove any leading zeros when destringing.
2. Non-categorical text. It is acceptable to store text answers that cannot be converted into categories as strings. A few examples follow:
    1. Open-ended questions: open-ended questions should, in general, be avoided, but sometimes the questionnaire asks the respondent to answer a question in his or her own words.
    2. Other specifications: the respondent is asked to specify the answer after answering *other* in a multiple choice question.
    3. Proper names: names of people, etc. Note that not all proper names should be stored as string as some can be made into categories. For example, if you collect data on respondents and multiple respondents live in the same villages, then convert the variable with the village names into a categorical numeric variable and assign a value label.
3. There are several ways to add helpful descriptive text to a data set in Stata, but the two most common and important ways are variables labels and value labels.

4. Variable Labels

5. All variables in a clean data set should have variable labels that explain what the variable represents. In addition to a brief explanation of the variable and perhaps the question number from which it comes, you may also decide to include information such as the unit or currency used in the variable. The label can be up to 80 characters long.

6. Value Labels

Always store categorical variables numerically and use value labels to indicate what the numeric code represents. For example, yes and no questions should be stored as 0 and 1 with the value labels *No* for data cells with 0, and the label *Yes* for all data cells with 1. This same concept applies to multiple choice variables. There are tools in Stata that convert categorical string variables into categorical numeric variables and automatically apply the string as value labels.

7. Assert

Assert is Stata's most useful command for data checking. If the statement is true, Stata continues; if it is false, Stata issues an error message, and everything stops right there.

Annex 2 **Descriptive statistics**

Control variables

- Agro ecological zone
- Sex of respondent
- Age respondent
- Level of education (needs to be recoded from the poverty score card question)
- Marital status
- Relationship with head of household
- Type of household
- Number of household members
- Age of head of household (needs recoding from poverty score cards (if needed)
- Education level head of household
- Main occupation head of household
- Distance to nearest food market (in km; recoding miles)
- Distance to nearest agro-input shop/seed company (in km; recode miles)
- Membership of farmer group; main purpose of farmer group
- Membership of LBS
- Poverty score
- Means of transport
- Ability to read
- Land ownership
- Land size for cultivation
- Area cultivated in 2017 (2017 A +B)

Descriptive, Ttest for balance

This section provides background information on the demographic and socioeconomic characteristics of the sampled households. These include:

Crop production

This section covers information on the crops grown in the study are, before and after awareness activities interventions. This is followed by an in-depth analysis of bean and maize varieties grown, sources of seed and quantities of seed planted by source. Where possible, the data will be disaggregated by gender and zone, further specifications are provided for each line, if needed

- Crops grown in 2017A and 2017B
- Crops grown in 2018A and 2018B
- Crops grown in 2019A
- Maize and beans varieties grown in 2017, 2018 and 2019A, by variety type (local, OPV, hybrid, other-unclassified)
- Purpose for growing maize and bean variety, by seed source and seed type
- Reasons for growing particular variety
- Sources of seed of maize and beans in 2017, 2018 and 2019A, by variety type
- Quantity of seed of maize and bean in 2017, 2018, 2019A by variety type (local, OPV, hybrid, other)
- Area planted with maize and beans
- Yield of maize and beans, by seed type and seed source, and kg seed planted
- Perception about maize and bean yield by seed source
- Input use per input (pesticides, herbicides, labor, fertilizer) in maize and beans

Farmer knowledge on quality seed and its benefits

This section covers the baseline on farmers' knowledge on quality seed and its benefits, resulting from the seed awareness and access interventions by sex and zone, using baseline information.

- Farmers understanding of quality seed
- Attributes of quality seed for farmers when buying
- Attributes of home saved seed for farmers planting
- Knowledge on improved varieties
- Understanding on improved varieties
- Knowledge on blue and green labels

Access to extension, using baseline

- Access to information on crop production (by zone and sex)
- Access to seed related information (by zone and sex)
- Key message related to seed received by farmers (by zone and sex)
- Sources of information related to seed (by zone)
- Practices applied when using quality seed (by zone)

Farming experiences in the last 12 months

- Challenges faced with seed planted (by zone and gender)
- Crop production losses experienced (by zone and gender)
- Frequency of crops most affected (by zone and gender)
- Post-harvest losses experienced (by zone)

Access to credit in the past 12 months using baseline information

- Accessibility to credit (by zone and sex)
- Sources of credit (by zone)
- Reasons for borrowing (by zone)

Farmer knowledge on quality seed and its benefits

This section covers changes in farmers' knowledge on quality seed and its benefits, resulting from the seed awareness and access interventions by sex and zone using baseline and end line data by control and treatment, sex and zone.

- Farmers understanding of quality seed
- Attributes of quality seed for farmers when buying
- Attributes of home saved seed for farmers planting
- Knowledge on improved varieties
- Understanding on improved varieties
- Knowledge on blue and green labels

Knowledge about quality seed campaign using end line, by control & treatment, district and sex

- Access to information of quality seed (Golden Harvest campaign)
- Sources of information on quality seed
- Message received on quality seed by information source
- Attended

Knowledge about seed access points using baseline and end line, by control and treatment, district and sex

- Knowledge on seed sources for high yielding maize (OPV and hybrid varieties)
- Knowledge on seed sources for high yielding bean varieties

How to apply

Interested firms or individuals should submit their Proposals to hr@issduganda.org not later than 9th April, 2020.