UWEZO DATA CLEANING PROTOCOL

18th April 2013

This note briefly explains the protocol used to clean all the Uwezo datasets (rounds 1-3). It also responds to some common questions about how the data can be used.

Objectives

The primary objectives of the cleaning process are to:

- remove information from the data that could lead to individuals or specific schools/villages being located by end-users;
- ensure a minimum degree of consistency between datasets across countries and over time;
- removing observations that are redundant or manifestly problematic;
- address missing test score data in a simple and transparent manner; and
- facilitate analysis of the datasets (individually or in conjunction) by end-users.

Following the above, the main actions undertaken in the cleaning process involve:

- establishing a "clean" set of geographical identification variables for each country;
- ensuring that household, school and village datasets can be correctly and simply "joined";
- renaming and labeling core variables that are common to each dataset;
- creating consistent categorical / numeric codes for common variables where possible;
- dropping observations that do not have a minimum amount of usable information (e.g., no age);
- ensuring consistency across school-level info (e.g., someone enrolled in school cannot simultaneously have dropped out);
- calculating sample weights; and
- imputing missing test scores via a multiple regression procedure.

Note that the intention of the cleaning process is minimal as opposed to comprehensive. Therefore, no claims are made that the data is free of errors. Moreover, the focus of attention in the cleaning process is

on the household/child level data, not the school or village level information. Aside from ensuring correct geographical identification and unique identifiers for each enumeration area, this data is left untouched. As a consequence these datasets are somewhat less easy to use.

Cleaning steps

The cleaning process was developed/run in Stata and applied consistently (in exactly the same fashion) to all datasets. The steps in the process are as follows:

1. Drop redundant or non-anonymous variables
2. Clean geographic identifiers and unique EA codes;
3. Drop redundant or non-anonymous variables;
4. Rename (and labeling) variables to a consistent codebook;
5. Drop observations without sufficient information;
6. Clean school-level information to ensure consistency across responses;
7. Calculate sample weights;
8. Report missing observations and "balance" by gender/location;
9. Impute missing test score values;
10. Creation of a control file including dataset checksums and a master list of EAs.

Points 7 and 9 merit further comment; they are discussed in the next sections.

Sample weights

Overview of sampling method

The sampling method employed in each of the Uwezo surveys has followed a three stage process, entailing:

a. Selection of districts (strata) by random sample (with each district given an equal probability of selection);

b. Selection of enumeration areas (EAs) by probability proportional to size (PPS); and

c. Selection of households – by random systematic sampling

Calculation of design weights

The first stage in the calculation of design weights reflects differences in the probability of inclusion of each household in the sample. Based on the above three stage process, the inclusion probability for household *i* located in EA *j* within strata *k* and country *l*, is the multiple of three probabilities – the probability of selecting that household within the enumeration area in which it is located, the probability of selecting that enumeration area given the stratum, and the probability of choosing that stratum. Algebraically this can be stated as follows:

$$P_{ijkl} = \frac{\alpha_j}{E_j} \times \frac{\beta_k E_j}{S_k} \times \frac{\gamma_l}{N_l} = \frac{\alpha_j \times \beta_k \times \gamma_l}{S_k \times N_l}$$

where α_j gives the number of households sampled in EA *j*; β_k gives the number of EAs sampled from strata *k*; γ is the number of strata in sampled; E_j is the total number of households in EA *j*, S_k is the total number of households in strata *k*; and *N* is the total number of strata (districts) in the country. The numbers in the denominator all derive from the sample frame.

Note that if all districts are included in the sample frame (i.e., with probability equal to one), then the last term in the middle part of equation (1) falls out (i.e., is set to one). In turn, this means that the inclusion probability becomes:

$$P_{ijkl} = \frac{\alpha_j}{E_j} \times \frac{\beta_k E_j}{S_k} \times 1 = \frac{\alpha_j \times \beta_k}{S_k}$$

In other words, the probability of inclusion is defined as the number of households sampled in a given stratum divided by the total number of households in that stratum. The design weight is then just the inverse of the inclusion probability:

$$w_{ijkl} = 1/P_{ijkl}$$

Note that in the design of the survey it may be the case that the design coefficients for α_j and β_k are held constant across EAs and strata. However, during implementation of the survey it is likely that these values vary somewhat. Discrepancies can also arise once a cleaning procedure has been applied, which may

remove certain observations (and are assumed to be removed at random). Thus, in calculating final design weights, we apply values for the coefficients (α_j and β_k) based on the actual number of households remaining in the sample after some observations are removed.

Imputation of missing test scores

An objective of the Uwezo surveys is to monitor levels of actual skills acquired, both for children enrolled in school and those not enrolled (but of school age). For a variety of reasons, including variation in the quality of enumerators, some children included in the survey either were not administered the test scores or they did not answer certain tests. In statistical analysis, the existence of missing observations would not be material if it occurred purely at random, in the sense that the existence of a missing observation was unrelated to other characteristics such as the child's age or location. However, analysis of the data suggests that observations are not 'missing at random' but rather are concentrated in particular locations or subpopulations. For instance, children never enrolled in school have a much higher probability of not answering the tests.

In order to ensure that test score results are representative of the entire school age population, it is therefore necessary to impute test scores where they are missing. Given the selective pattern of non-response to the tests, if we would treat non-response on test results simply as missing values then test scores likely would be overestimated. A simple solution would be to assume the lowest score in cases of non-response; however, we feel that a better estimate of true test scores is obtained using imputation. The latter is done based on an ordinary least squares regression of the observed test scores on a range of "core" characteristics of the child. These include the child's age (included as fixed effects), her gender, the grade in which she is enrolled (included as fixed effects), whether she is currently in preschool, whether she has dropped out, the interaction of age and dropping out, whether she is in secondary school, her household size and her geographic location (included as district fixed effects). To impute test scores we take the fitted value from this model for children where the test score is missing and add a random error term, the location of which is held fixed for the same child across different tests.

Note that the original test scores (which include the missing values) are retained in the published datasets. Imputed scores are identified with a suffix to the variable name ("_imputed"). Users should be aware that imputed values are estimates which are subject to error; thus, they should be used with due caution.

Outputs

For each dataset (i.e., a Uwezo survey in a given country and year [denoted below as XXYY]), there are five main outputs of interest. They are as follows:

1. Child/household level data file: observations are reported at the child level, although some variables are constant for multiple children in the same household. [File is located in "Data/XXYY_hhld.dta"]
2. School-level data file: corresponds to a single government primary school in the EA. Not all children interviewed will attend this school. [File is located in "Data/XXYY_school.dta"]
3. Village-level data file: summary data about the EA. [File is located in "Data/XXYY_village.dta"]
4. Control file: this is an excel workbook containing two sheets. The first sheet reports the file names of the datasets, their time of creation and their checksums. The second sheet is a master list of EAs and includes dummy variables indicating whether a given EA is contained in any of the underlying datasets. [File is located in "Codebook/XXYY_EaCodeList.xls"]
5. A code book: this summarises the main child/household variables and reports their "original" names as per the raw data, their labels and data type. This can be used for reference purposes. [File is located in "Codebook/XXYY_Codebook.xls"]

Other documentation, where available, including sample questionnaires is provided with the public datasets.

Dataset use FAQs

1. **How can I merge the school-household-village datasets?** The unique identifier on which to join the datasets is the variable "id_village", which represents the relevant EA code. Be aware that this will merge school-level information to all children in the EA, regardless of whether they attend the school itself.

2. **How can I match children to schools?** Only children that (a) attend a primary school; and (b) attend the same primary school that was surveyed (as per the school file) can be matched. In most datasets (but not all) a dummy variable is included "schoolmatch" that takes the value of one if conditions (a) and (b) hold.

3. **How do I use sample weights?** Sample weights are included as the variable "weight". These refer to all children in the sample NOT the sub-sample of children attending school. The latter would need to be calculated separately, adjusting for the proportion of non-enrolled children in the district. No sample weights are given in the village or school-level datasets as it would be less meaningful to do so.

4. **How do I tell what form/grade/class a child attends?** If a child attends either primary or secondary school (not pre-school) the variable "enr_ans" will take the value of one. The numeric class is given by the variable "grade". The user should interpret this as equal the completed years of schooling minus one. Therefore a child in the first year of school (e.g., Standard one) takes a value of one. The numeric value applies similarly to secondary school. Thus, Kenyan pupils in the first year of secondary school take a value of 9 (as primary school is eight years); Ugandan and Tanzanian children in the first year of secondary school take a value of 8 (as primary school is seven years).

5. **How do I tell what a variable refers to?** In principle the variable name and label should be informative (as per the codebook). If not, then the user must resort to the questionnaires and figure it out. The latter exercise is requisite when using the school and village data, as no cleaning of variable names has been attempted. This is not to make life difficult for users, but simply reflects resource limitations on the part of Uwezo. Some datasets include a codebook for the school/village data and this is included (without warranty) where available in the public release bundle.

6. **How should I cite the data?** Use of the data in reports or publications should include acknowledgement of the specific source/site where the data was obtained. We also recommend citation of one/more of the Uwezo publications (found at: www.uwezo.net) to point the reader to relevant background documentation.

7. **What restrictions are there on using the data?** The data remains the property of Uwezo and is provided for non-profit, research purposes only. Commercial or for-profit uses are strictly prohibited without prior consent. We advise all potential and effective users of the data to read the data use agreement found with the data.

8. **Who is responsible for analysis undertaken with the data?** All data is provided on an "as is" basis. This means Uwezo provides absolutely no warranty as to the underlying data quality or its fitness

for purpose. Users of the data must take full and independent responsibility for the analysis they undertake and conclusions they reach.