THE ZAMBIA MULTIPLE CATEGORY TARGETING GRANT (MCTG)

DATA USE INSTRUCTIONS

OVERVIEW

This document provides information for using the Zambia MCTG data, a three-wave panel dataset that was created to analyse the impact of Zambia's MCTG cash transfer program. In addition to explaining the data structure, it provides brief information about the program and the evaluation.

This dataset is released by The Transfer Project, housed at the Carolina Population Center at the University of North Carolina – Chapel Hill. Additional information about the project not found here or without a direct link can be found on The Transfer Project's Website: https://transfer.cpc.unc.edu/.

The data package contains four longitudinal primary datasets (one community level, two household level, and one individual level). The surveys interviewed households, individuals, and community members at three time points in 2011, 2013 and 2014 (also referred to hereafter as baseline, midline and endline).

THE PROGRAMME

The Zambia's Multiple Category Targeting Grant (MCTG) was an unconditional social cash transfer run by the Ministry of Community Development, Women and Child Health (MCDMCH). The program targeted extremely vulnerable households taking care of orphans in two rural districts with some of the highest rates of food insecurity and poverty – Serenje and Luwingu. The overall goal of the MCTG was to reduce poverty and its intergenerational transfer; specific objectives relate to five primary areas: income, education, health, food security and livelihoods[1].

At the time of baseline data collection for this study in 2011, beneficiary households received a flat rate of 55 Zambia Kwacha (equivalent to roughly US$11) a month. At baseline, the transfer represented a 21 per cent increase to the household's baseline monthly expenditure. The benefit level had been set with the intent to cover one meal a day for everyone in the household for a month. To keep up with inflation, the transfer was increased to 60 and 70 ZMW by midline and endline respectively. Beneficiaries received the intended amount of funds through a local paypoint manager and according to schedule, regularly and on time; programme implementation largely functioned as expected.

THE IMPACT EVALUATION and THE SAMPLE

The impact evaluation was commissioned by the Government of Zambia and UNICEF as part of the Transfer Project. It was implemented by the American Institutes for Research (AIR) and designed as a longitudinal multisite cluster RCT with one baseline in 2011 and two follow-ups at 24- and 36-months (in 2013 and 2014 respectively). The ethical rationale for

[1] The six specific objectives of the programme were: 1) To supplement and not replace household income; 2) To increase the number of children enrolled in and attending primary school; 3) To reduce the rate of mortality and morbidity among children under five; 4) To reduce stunting and wasting among children under five; 4) To increase the number of households having a second meal per day: 5) To increase the number of households owning assets such as livestock.

the design was that the Ministry did not have sufficient resources or capacity to deliver the programme to all eligible households immediately, so communities who would entry the programme later on during the expansion phase could be used as control sites to measure impact.

In each of the two selected districts, 46 randomly selected communities (also referred to as clusters or CWACs – Community Welfare Assistance Committees) were randomly assigned to either treatment or delayed control status through public lottery. Within each of these 92 communities, roughly 33 households were randomly sampled for inclusion in the study, leading to a sample of over 3,000 households.

The full baseline sample contains 3,077 households and 15,370 individuals. Households in both arms were first interviewed – prior to learning whether they would be selected into the programme – at baseline in 2011 during the lean season (from September through February); the first transfer to MCTG beneficiaries was made short after baseline data collection. The sample was then interviewed again in a 24-month follow-up survey in 2013 (November-December), and last in 2014 (November-December), after the MCTG had been operating in study areas for three years. Table 1 shows the study sample by wave and treatment status.

Table 1: Household samples for the evaluation

	Baseline (2011)	Midline (2013)	Endline (2014)
Treatment	1,561	1,522	1,490
Control	1,516	1,489	1,481
TOTAL	3,077	3,011	2,971

The above table shows that overall 3,011 and 2,971 of the 3,077 households interviewed at baseline were tracked at midline and endline respectively. As a result, overall household sample attrition was low: 2% at midline and 3% at endline. There is no indication of differential attrition between treatment arms: official evaluation reports also indicate that successful randomization led to very similar control and treatment groups.

All evaluation reports (baseline, midline and endline) used for the study can be found here: https://transfer.cpc.unc.edu/countries-2/zambia/

Program Eligibility

The survey for the impact evaluation collected information for a sample of eligible households in treatment communities (*beneficiaries* − immediate entry in the programme) and eligible households in control communities (*would be beneficiaries* − delayed entry in the programme). Eligibility criteria to enter the MCTG required to meet any of the following conditions:

- A female-headed household keeping orphans;
- A household with a disabled member;
- An elderly-headed household (over 60 years old) keeping orphans;
- A special case: cases that do not qualify under any of the above listed categories but are considered critical. This could be for instance a couple of elderly people unable to look after themselves.

MAIN CHARACTERISTICS OF THE DATASETS

The study relied on three main type of instruments: 1) an extensive household survey (including a youth module); 2) a community survey; and 3) a health facility survey. All the instruments (baseline, midline and endline) used for the study can be found here: https://transfer.cpc.unc.edu/countries-2/zambia/

Data is released in four main longitudinal datasets:

- *2 household level* datasets
    - *Sections 15* (household expenditure) and *9 A-G* (agricultural production, livestock and animal production, related household expenses, land, and business modules)
    - all other household level data;
- *1 individual level* dataset;
- *1 community level* data including health facility data.

Each data is discussed in more details below.

Main household level dataset ['household_longitudinal']

The household dataset comprises the full list of (raw) variables included in the household survey and collected at the household level as well as a number of additional variables and/or aggregates computed by the evaluation team. These variables are reported at the end of the dataset. For greater ease of viewing, a variable:

ADDITIONAL 'VARIABLES----------'

was included to more clearly identify these sets of variables.

Among the identifiers, the variable **round** captures the survey wave, while the variable **qsn** is the unique household identifier within each wave. To uniquely identify each household over time, both round and qsn should be used.

Among other important variables provided by the evaluation team is:

- **treat**: a dummy variable capturing the treatment status (according to the community level randomization).
- **clid**: the cluster id (community or CWAC). There are 92 unique clusters in the data at baseline.
- **panel_overall_36**: a dummy variable that captures the survey status of the household across the three rounds (i.e. 1 if the household was surveyed, 0 otherwise). This household level dataset is square (wide), meaning that it contains 3,077 observations at each wave and can therefore be easily used for attrition analysis/computation.
- **ipw_24** and **ipw_36**: these are inverse probability weights computed at the household level based on the 24- and 36-month surveys respectively; computations were made by the evaluation team.

Among the additional variables provided by the evaluation team are also a set of household composition and demographic variables, already computed consumption aggregates (ZMW - 2011 units, i.e. the base year) and a set of baseline distances from the household to different services (school, health facility, food market).

Note: The household dataset also includes Sections 11 and 12; even though the questions in these modules may refer to specific individuals, data was collected at the household level and included in the household longitudinal data file.

Second household level dataset: sections 9 and 15 [HH_Sections9&15_longitudinal]

The data file "HH_Sections9&15_longitudinal" includes all household level data which relates to Sections 15 and 9A-G, namely 'Household Expenditure' and 'Agricultural/Livestock/Animal Production, related expenses, land and business modules'.

The identifiers are the same as for the 'household_longitudinal' data, namely *round* and *qsn*. The variable *treat* and *clid* are also included in this data file.

Similarly to the household longitudinal data file, this dataset is squared and each row in the data refers to a unique household at a specific wave. Indeed, to facilitate data management and use, some modules have been reshaped so that every row captures a household at a round (rather than an expenditure item, or livestock asset).

Note: Main household expenditure aggregates are included in the 'household_longitudinal' dataset although a number of additional variables computed by the evaluation team based on Sections 9A-G are provided at the end of this dataset.

Individual level dataset ['individual_longitudinal']

The individual dataset comprises the full list of (raw) variables included in the household survey at the individual level as well as a number of additional variables already computed by the evaluation team. Similar to the longitudinal household dataset, these variables are reported at the end of the dataset.

The unique individual identifier within each wave is **id**; id combines the unique household identifier **qsn** with **mid**, the unique individual identifier *within* the household. The variable 'mid' provides useful information. At baseline, and within each household, *mid* was assigned starting from 1 to n. At follow-ups, individuals keep their original baseline mid; however, individuals who joined the household after the baseline were assigned a new mid: starting with 201, 202, etc. for new household members joining at 24-month, or starting with 301, 302, etc for those who joined at 36-month.

To uniquely identify an individual within each wave, the data user should combine the following variables: **round** + **id** (equivalent to: round + qsn + mid). The individual level dataset includes the treatment dummy (*treat*) and the cluster IDs (*clid*).

The individual dataset includes all individual level data, as such it comprises also Section A1 and A2 on 'Household Composition Confirmation' and 'New Household Members Listing'; this basically means that it includes all household members who were ever surveyed as well as further information on those who attrited (see Section A1 for leave reasons, etc.).

Among the additional variables provided by the evaluation team are variables to easily identify new household members and individuals who are no longer household members (i.e. 'new_member' 'no_longer_member), gender and age. Beyond the raw gender (s1q5) and age (s1q3a s1q3b) variables, the evaluation data also provides the cleaned variables at the end of the dataset ('gender_r' and 'age_r').

Data from the Youth module (Section 16: OVC module for members age 15-19) contains sensitive information and is not released at this time.

Community level dataset ['community_longitudinal']

This longitudinal dataset includes not only all variables from the community survey but also those from the health facility questionnaire (collected only at baseline). For greater ease of viewing, a variable:

HEALTH 'FACILITY QUESTIONNAIRE (Baseline only)----------'

was included to more clearly identify these sets of variables.

The community identifier is **clid**, which should be combined with **round** to uniquely identify each community at each wave. The community level dataset includes the treatment dummy (*treat*).

This community level dataset is squared, meaning that it contains 92 observations (communities) at each wave.

Merging datasets and other useful STATA commands

To match the individual dataset with the household level dataset:

```
use individual_longitudinal.dta, clear
mmerge round qsn using household_longitudinal.dta
```

To match the household (or individual) dataset with the community dataset:

```
use household_longitudinal.dta, clear or use individual_longitudinal.dta,
clear
mmerge round clid using community_longitudinal.dta
```

To match the two household level datasets:

```
use household_longitudinal.dta, clear
mmerge round qsn using HH_Sections9&15_longitudinal
```

To reproduce the statistics in Table 1:

```
use household_longitudinal.dta, clear
bysort round: ta treat, mi
```

To quickly tabulate/check household level attrition:

```
use household_longitudinal.dta, clear
bysort round: ta panel_overall_36, mi
```

Other – Additional variables provided

Clean variables: Some variables are provided already cleaned by the evaluation team. These variables maintain the original questionnaire name/code but are recognizable as followed by '_r'. The label of these variables also indicates in squared brackets that the variables have been cleaned [cleaned].

Monetary values, when provided cleaned (i.e. '_r'): 1) Reported values from the baseline 2011 data (in kwacha - Kw) were rebased to ZMW (dividing by 1,000); 2) follow-up values were deflated to 2011 units using the all-Zambia consumer price index (CPI). In some cases, missing values were imputed and outliers replaced.

Important: For monetary values that are not provided clean, the data user should rebase the values at baseline (i.e. divide by 1000) and deflate follow-up values to 2011 units.

The formula for calculating the *Deflation Rate* is as follows: ((B - A)/B)*100

For midline values:

- "A" is the November 2011 CPI (116.9407) and "B" is the November 2013 CPI (133.82), so the deflation rate is computed as: ((133.82-116.9407)/133.82)*100=12.6134%

For endline values:

- "A" is the November 2011 CPI (116.9407) and "B" is the December 2014 CPI (145.74), so the deflation rate is computed as: ((145.74-116.9407)/145.74)*100=19.7607%

Hereafter the STATA code to rebase the kwacha Kw to ZMW, and deflate monetary values to 2011 units.

```
foreach x of varlist $money {
replace `x'=`x'/(1000) if round==1
replace `x'=`x'/(1+.126134) if round==2
replace `x'=`x'/(1+.197607) if round==3
}
```

De-identification and sensitive information

For security and privacy purposes, names, contact, GPS coordinates and any potentially identifying information of the individuals and households have been removed, and the names of any geographic units smaller than a district have been coded.

Caution: Some questions may not be available at each wave. Over time, some questions have been added, dropped and in a few cases slightly changed.

ANNEX - Table A1: Description of MCTG datasets

Dataset	Sections included	Additional variables	Identifiers
household_longitudinal	From household questionnaire: - Section 0 (cover page); - Section 6 (Inventory of household assets); - Section 7 (Household amenities and household conditions); - Section 8 (Access to facilities and programs); - Section 10 (Self-assessed poverty, food security and shocks to household welfare); - Section 10A (Credit); - Section 11 (Women's empowerment, savings and future expectations); - Section 11C (Women's health knowledge and affect); - Section 11D (Aspirations and expectations for children); - Section 11E (Social Support); - Section 12 (Deaths in the household); - Section 14B (Operational performance).	Treatment status; cluster id; panel variable; inverse probability weights; demographic variables; expenditure variables; distances from services.	'round' captures the survey wave. 'qsn' is the unique household identifier within each wave. 'round' + 'qsn' uniquely identify each household over time.
HH_Sections9&15_longitudinal	From household questionnaire: - Section 9A (Agricultural production); - Section 9B (Livestock and animal production); - Section 9C (Fish farming); - Section 9D (Household expenses for livestock and animal production); - Section 9F (Land module); - Section 9G (Business module); - Section 15 (Household expenditure).	Some agricultural production and business variables.	'round' captures the survey wave. 'qsn' is the unique household identifier within each wave. 'round' + 'qsn' uniquely identify each household over time.
individual_longitudinal	From household questionnaire: - Section A1 (Household composition confirmation); - Section A2 (New household members listing); - Section 1 (Household roster and OVC status); - Section 2 (Health for all persons); - Section 3 (Education - for all persons age 3 and above); - Section 4 (Main economic activity); - Section 4B (Economic activity - members aged 5-19 only); - Section 4C (Wage labour); - Section 5 (Income); - Section 9H (Family labour in family businesses and agriculture); - Section 13A (Child health and development); - Section 14 (Reproduction).	Treatment status; cluster id; clean gender and age variables; variables that capture whether the individual is a new household member or is no longer a household member.	'round' captures the survey wave. 'id' is the unique individual identifier within each wave ['id' combines the unique household identifier qsn with mid, the unique individual identifier within the household]. 'round' + ''id' uniquely identify each individual over time.
community_longitudinal	All community questionnaire Sections as well as Sections from the baseline Health Facility questionnaire.	Treatment status.	'round' captures the survey wave. 'clid' is the unique cluster (community) identifier within each wave. 'round' + 'clid' uniquely identify each community over time.