THE ZAMBIA CHILD GRANT PROGRAMME (CGP)

DATA USE INSTRUCTIONS

OVERVIEW

This document provides information for using the Zambia CGP data, a five-wave panel dataset that was created to analyse the impact of Zambia's CGP 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 niner longitudinal primary datasets (five community level, two household level, one individual level and one facility level). The surveys interviewed households, individuals, and community members at five time points in 2010 (baseline), 2012, 2013 (June and October) and 2014.

THE PROGRAMME

The Zambia Child Grant Programme (CGP) was an unconditional social cash transfer run by the Ministry of Community Development, Women and Child Health (MCDMCH). The program was a demonstration that was implemented I the three districts of Kalabo, Shang'ombo and Kaputa between 2010-2014, targeting rural households with a child under age five years. The overall goal of the CGP was to reduce poverty and its intergenerational transfer.

At the time of baseline data collection for this study in 2010, beneficiary households received a flat rate of 55 Zambia Kwacha (equivalent to roughly US$11) a month. At baseline, the transfer represented a 28 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 during the life of the study. Beneficiaries received the intended amount of funds bimonthly (six times a year) through a local paypoint manager and according to schedule, regularly and on time; programme implementation largely functioned as expected.

The baseline report, which provides details of the sampling and targeting and coverage of the programme, is available here:

https://transfer.cpc.unc.edu/wp-content/uploads/2021/04/Zambia-CGP-Baseline.pdf

The associated report for each of the study waves, which describes attrition and programme impacts, are available here:

https://transfer.cpc.unc.edu/countries/zambia/#reports

Please review these reports carefully prior to using the data.

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 2010 and four follow-ups as described above. The ethical rationale for the design was that the Ministry did not have sufficient resources or capacity to deliver the programme to all eligible households within a district (the intervention was a demonstration). It was decided to choose the clusters (also referred to as CWACs – Community Welfare Assistance Committees) that would receive the program randomly, with other CWACs then serving as controls. At the end of the

demonstration phase in 2024, eligible households in control CWACs were provided a lump-sum payment for their participation in the study.

In each of the three selected districts, 30 randomly selected CWACs were randomly assigned to either treatment or delayed control status through public lottery. Within each of these 90 CWACs, roughly 25 households were randomly sampled for inclusion in the study, leading to a sample of 2,519 households.

The full baseline sample contains 2,519 households and 14, 345 individuals and by 2014 there were 17,051 individuals in the study sample. Households in both arms were first interviewed – prior to learning whether they would be selected into the programme – at baseline in 2010 from October to early December; the first transfer to CGP beneficiaries was made short after baseline data collection. The sample was then interviewed again in the four subsequent waves. Table 1 shows the study sample by wave and treatment status.

Table 1: Household samples for the evaluation

	Treatment	Control	Total
2010	1260	1259	2519
2012	1153	1146	2299
2013-June	1182	1185	2367
2013-October	1221	1239	2460
2014	1202	1227	2429

MAIN CHARACTERISTICS OF THE DATASETS

The study relied on three main type of instruments: 1) an extensive household survey; 2) a community survey; and 3) a health facility survey conducted at baseline only. 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 nine 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;
- *5 community level* data, one for each survey wave.

Each data is discussed in more details below.

Main household level dataset ['household_longitudinal_allwaves']

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 _time captures the survey wave, while the variable **qsn** is the unique household identifier within each wave. To uniquely identify each household over time, both _time 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 unique cluster id (or CWAC). There are 90 unique clusters in the data at baseline.
- **panel_overall_48**: a dummy variable that captures the survey status of the household in ech of the five rounds (i.e. 1 if the household was surveyed, 0 otherwise). This household level dataset is square (wide), meaning that it contains 2,519 observations at each wave and can therefore be used for attrition analysis/computation.
- **ipw_24, ipw_30, ipw_36 and ipw_48**: these are inverse probability weights computed at the household level based on the associated survey wave; 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 - 2010 units, i.e. the base year) and a set of baseline distances from the household to different services (school, health facility, food market).

Any variable with the suffix _b is a baseline measure.

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 *time* and *qsn*.

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_allwaves’ dataset as well as a number of additional variables computed by the evaluation team based on Sections 9A-G.

Individual level dataset [‘individual_panel’]

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. 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.

The unique individual identifier within each wave is id which 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 and so on. Tabulating *mid* quickly shows how many individuals joined at each wave.

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’).

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.

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.

Community level dataset [‘Community_0m, Community_24m, Community_30m, Community_36m, Community_48ml’]

These datasets include all variables from each of the five community survey.

The community identifier is **clid.** This community level dataset is squared, meaning that it contains 90 observations (clusters) at each wave.

Health facility survey [Health Facility_merged]

The health facility survey was only conducted once, at baseline. Not each cluster had a health facility, there were 41 facilities within the catchment area of the 90 clusters, of which three facilities are from the same cluster. The study team took the average of the three facilities when merging facility information to the household. The facility data set can be linked to the household data set using *district* and *cwac*.

Merging datasets and other useful STATA commands

To match the individual dataset with the household level dataset:

```
use individual_panel.dta, clear
mmerge _time qsn using household_longitudinal_allwaves.dta
```

To reproduce the statistics in Table 1:

```
use household_longitudinal_allwaves.dta, clear
bysort _time: tab treat, mi
```

To quickly tabulate/check household level attrition:

```
use household_longitudinal_allwaves.dta, clear
bysort _time: ta panel_overall_48, 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 brackets that the variables have been cleaned (cleaned).

Monetary values, when provided cleaned (i.e. '_r'): 1) Reported values from the baseline 2010 data (in kwacha - Kw) were rebased to ZMW (dividing by 1,000); 2) follow-up values were deflated to 2010 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 2010 units.

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

For 2012 values:

- "A" is the October 2010 CPI (109.44) and "B" is the October 2012 CPI (124.8), so the deflation rate is computed as: ((124.8-109.44)/124.8)*100=12.308%

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. Please check the questionnaires on the Transfer Project website.