

Guide to Questionnaires and Supporting Materials
Ghana Cocoa Farmers Survey 2006

Centre for the Study of African Economies
and
Ghana Cocoa Board

Contents

1. Sampling and survey methods	1
2. Supporting materials	3
3. Overview of GCFS 2006 structure	3
4. Data format	5

The Ghana Cocoa Farmers Survey 2006 was jointly conducted by a team comprising members of the Centre for the Study of African Economies (CSAE), the Ghana Cocoa Board (COCOBOD), and the Ghana Statistical Service (GSS). Thanks are also due to Harunah Maamah (ECAM Consultancy, Ltd) for his help in making logistical arrangements. The survey was conducted in September of 2006.

This document serves two functions. First, it provides information on the nature of the sample and questionnaires. Second, it provides a guide to the supporting materials and data itself, which can be used by those wishing to access the data.

1. Sampling and survey methods

Sampling frame

The GCFS 2006 represents an extension of a panel begun in 2002 by Marcella Vigneri, then at CSAE. The original sample was drawn as a random selection of cocoa-farming households in the 1998/1999 Ghana Living Standards Survey. In that year 25 villages from the GLSS sample were selected with probability proportional to the size of the cocoa-farming population in each village. Not all GLSS households could be traced; remaining households were drawn from a listing of cocoa-farming households in each village. The sampling unit within villages is the farmer, not the household, so it is possible (though quite rare) for multiple individuals to be observed within the same household. A total of 492 completed household interviews were conducted.

In 2004, the second round of the GCFS was conducted with the intention of revisiting all households from the previous round. A total of 49 households (from the original Households reported to have left the village were replaced as follows:

- In cases where the current owner or cultivator of the land previously cultivated by the household could be identified, the primary owner or cultivator of the land was used as a replacement.
- In addition, a number of additional households were sampled. However, requirements of time and departures from protocol mean that these additional households should be considered a non-random sample (e.g., dependent on turnout, etc).

The procedure by which each farmer was added to the sample is reported in the data (Section 0).

Similarly, the GCFS 2006 attempted to resurvey all households interviewed in 2004. Re-sampling procedures were improved somewhat relative to the previous round. Farmers could be added to the sample through one of two routes:

- First, as in 2004, farmers cultivating land previously owned or operated by a now-absent survey respondent were added to the sample in cases where their identity could be confirmed.
- Second, a complete list of cocoa farmers in the village was obtained from the registers of all purchasing clerks in the village. Farmers were selected at random from these registers, with those who had made sales but came from other villages or who had migrated (potentially seasonally) for

In addition, one new village was added in the Western region. Farmers in this village were selected by a listing exercise based on the registers of purchasing clerks. The procedure for which each new farmer was added to the sample is reported in the data (Section 0).

Survey methods

Data from the 2004 and 2006 rounds were collected on handheld computers using a questionnaire designed in PocketSurvey. (The raw file in PocketSurvey's .Q format is available for specific purposes on request.) Documentation of the questionnaire is described in Section 2.

Both the presence of CSAE researchers Andrew Zeitlin and Daniel Clarke in the field for data collection in 2004 and 2006 and, crucially, the inclusion of staff from Cocobod Research Department and CRIG staff in data collection allowed for close interaction between researchers and the population of farmers studied. Moreover, the electronic form of data collection allowed data to be analyzed on a regular basis while in the field, allowing quick correction of errors. Errors might be of two sorts: mis-entry of data or problems in the design/flow of a complex electronic questionnaire. Enumerators were equipped with paper copies of questionnaires and prompt sheets to address this problem.

One implication of the close interaction between research design and data collection is that some modifications were made to the questionnaires during the process of data collection. In most instances amendments to the questionnaire were minor and of a technical/programming nature. In a few cases more substantial additions were made to allow investigation of particular hypotheses. These modifications have in some instances

been fruitful (e.g., the addition of questions on reasons for and destinations of departed household members, and Cocobod's interest in investigating labor supply by 'young adult' farmers in 2006). They do however have the implication that certain variables may not be available for all villages in the sample. Such changes should be immediately identifiable by reference to date of survey or village codes, and will apply only to questions added since the 2004 questionnaire.

The data collection and processing have incorporated some consistency checks imposed on the data. This was aided by the use of physical (pen-and-paper) prompt sheets for each interview, upon which interviewers recorded key variables by hand, and which provided interviewers with data from the preceding round that could be used to match plots and household members across rounds of the survey.

2. Supporting materials

The supporting materials are provided alongside this document to facilitate access to the GCFS:

- (i) 2006 Questionnaire overview (excel format);
- (ii) Full questionnaires (html format)
- (iii) 2002 Questionnaire (pdf format)

The excel copy of the questionnaire was prepared largely by Daniel Clarke, to whom I am very grateful for his hard work on this (among other aspects of the survey design and oversight). This is intended and should be seen as indicative only, to provide a more traditional overview of the questionnaire and the way in which individual modules fit together.

Full questions and available answer codes as used in the field are available in the html-format questionnaire files. These files are broken down by module as indicated in the following Section. Users may find it easiest to navigate through the html files either by searching for particular questions – using variable names as provided in the excel-format questionnaire or keywords from questions themselves – or by following links to navigate through the questionnaire in sequence. Note that the html files are not in general organized in chronological order of the questionnaire as read to a respondent.

3. Overview of GCFS 2006 structure

The questionnaire and associated data consists of two types of files. The 'parent' questionnaire contains one observation per respondent. The html version of this questionnaire, supplied as an attachment to this document, is titled [gcfs06sept20.html](#). The corresponding data are in the file titled [gcfs06.dta](#).

Respondents are identified in all questionnaires by a unique identifying code, denoted `s0far`. There are several modules that allow many-to-one responses. Each observation in these files contains both the value of `s0far` to which it corresponds as well as a sub-questionnaire identifier. The combination of these two uniquely identifies each sub-

observation. Briefly, each of the many-to-one data files is described below, along with the file in which questions can be found, and the unique sub-questionnaire identifier.

- roster.dta: This file contains information on each current resident of the household. Household members are uniquely identified by variable `rosterid`. For the 2006 data, the values for `rosterid` correspond to the values of `rosterid` in the previous wave, so that individuals can be matched across rounds. The full set of questions and codes for this file is available in the questionnaire [roster_sept20.html](#).
- missing.dta: This file contains information on household members present in 2004 but not present in 2006. Household members are uniquely identified by the variable `rosterid`, as in the file [roster.dta](#); and as in [roster.dta](#), these ID numbers correspond to the 2004 roster IDs of household members – though in these case the individuals no longer belong to the household.
- plots.dta: This file contains information on a plot-by-plot basis, including tenancy status, physical characteristics, etc. Plots are uniquely identified by the variable `plotid`. In 2006, enumerators were provided with a vector of characteristics of each plot from the 2004 data in order to attempt to match plots over time. Thus values of `plotid` also correspond to their 2004 values. New or unmatched plots should take values of `plotid` not appearing in the 2004 data. This can be confirmed by an additional question explicitly asking enumerators to confirm that plots were matched to the previous round. Variable `matched` takes values of 1 for plots that were successfully matched to a plot in the previous data. Further, in a subset of villages the questions `s3_sizematch` and `s3_plotmatchwhynot` were asked. These explore, respectively, the whether and – if not – the why not aspects of plot sizes being the same across years.

Note that some information on inputs at plot level is actually collected in the main file (`gcfs06`). Doing so allows PocketSurvey to reference previously answered questions across plots in order to impose (or at least facilitate) consistency in answers regarding inputs applied to each plot and in aggregate. Where this was done, the plot IDs corresponding to each question suffix (`_1`, `_2`, ... , `_5`) were collected in variables `idplota`, `idplotb`, ... `idplote` (note these questions were prefixed `s4_plot*` in some of the initial villages).

- output.dta: This file contains information on all crops produced for sale by the farmer. Crops are uniquely identified by variable `cropid`. Please note that the cleaned *cocoa* output data is given in Section 2 of the questionnaire (in the file [gcfs06](#)). Where problems were encountered, information from both the general-purpose output section of the questionnaire and the physical prompt sheets (which asked enumerators to record total cocoa output for the 2005/06 season) were cross-referenced to provide the final value given in the variable `kg_cocoa` of the file [gcfs06](#).
- marketing.dta: Contains information on the farmer's relationship with each of the marketing outlets (Licensed Buying Companies) to whom he or she sold cocoa in the 2005/06 season. Each farmer-LBC relationship is uniquely identified by the

farmer ID, `s0far`, and the variable `lbcid` in this file. LBCs themselves are identified in variable `s8bq3`, but the names of particular LBCs have been encoded to ensure anonymity of respondents. Codes for LBC names are consistent across waves of the survey.

4. Data format

All data are provided ‘as is’ in Stata format. Not all variables have been encoded. A full set of codes (i.e., one which includes potential responses that were not in any instances supplied by respondents) can be ascertained from the .html version questionnaires.

For accuracy, some numeric variables were entered as “pre-formatted” strings in PocketSurvey. What this means is that enumerators asked for, say, the price they were paid for cocoa, would have been prompted with a value of “999,999,999” and would have entered the correct price by replacing these data with, e.g., “000,562,500”. This was deemed the best way of ensuring that orders of magnitude were preserved. The teams adopted the convention of leaving such strings as “999,999,999” in instances where the respondent did not know an answer. More generally, a series of 9s was used to denote “don’t know” for numeric questions.

Appendix 1: Structure of GCFS 2006 files

Primary questionnaire	Questionnaire file	Data file	Linking questions (sub-identifier)
Survey information	gcfs06sept20.html	gcfs06	
Household Roster	roster_sept20.html	roster	rosterid
	missing_sept20.html	missing	rosterid
Intro: Getting to know the Farmer	gcfs06sept20.html	gcfs06	
Plot details	gcfs06sept20.html#plots	plots	plotid
Labour inputs	gcfs06sept20.html	gcfs06	
Non-labour inputs	gcfs06sept20.html	gcfs06	
Sale of output	output_sept20.html	output	cropid
Access to Credit and Collateral	gcfs06sept20.html	gcfs06	
Transfers and other sources of income	gcfs06sept20.html	gcfs06	
Other assets	gcfs06sept20.html	gcfs06	
Responses to unexpected events	events_sept20.html	events	eventid
Investment choices	gcfs06sept20.html	gcfs06	
LBCs	marketing_sept20.html	marketing	lbcid