CFPS China Family Panel Studies

Supplement to the User's Manual for the 2012 CFPS Data

Qiong Wu, Lihong Dai, Yahong Cui, Wenjia Zhang,

Cong Zhang, Jia Chen, Hongwei Xu, Yu Xie

03/11/2015

Contents

1. Introduction 1

1.1 Adult Dataset 1

1.2 Child Dataset 2

1.3 Family Dataset 2

1.4 Family Roster Dataset 2

1.5 Cross-Wave ID Dataset 3

2. Changes in Data Structure from CFPS 2010 4

2.1 Combining Face-to-face and Telephone Interview Data 4

2.2 Combining Long and Short Forms of Questionnaires in Individual-level Data 4

3. Composite Variables 6

3.1 Family Income (in the Family Dataset) 6

3.2 Family Expenditure (in the Family Dataset) 6

3.3 Community Type (in the Family, Adult, and Child Datasets) 6

3.4 Personal Income (in the Adult and Child Datasets) 7

3.5 Educational Level (in the Adult and Child Datasets) 8

3.6 Employment Status (Adult Dataset) 8

3.7 Cognitive Function (in the Adult and Child Datasets) 8

4. Best Variables 10

5. Variable Coding 11

5.1 Occupational and Industry Codes 11

5.2 Address Codes 11

6. Weighting Variables 12

1. Introduction

The CFPS baseline survey in 2010 consists of five basic datasets: community, family roster, family, adults, and children. In 2012, the CFPS carried out the first full-scale follow up survey, which targeted all 14,960 families, and 57,155 members from the 2010 family roster, who were considered CFPS gene members. In addition, individuals newly born to the 57,155 gene members would also be considered the CFPS gene members, while individuals newly married to the existing CFPS gene members would be considered the CFPS core members. By design, gene members are followed regardless of their relationships with their original families, whereas core members are followed as long as they maintain their family relations with gene members. The 2012 questionnaires share similar structures to those in 2010, although some specific contents have been modified. For example, no community dataset is available in the 2012 wave since the community survey is scheduled to be carried out in every other wave. A cross-wave individual-level dataset file has been added, listing all the individuals who ever participated in the CFPS, as well as their vital records, survey participation status, respondent type, and family ID in each wave. Table 1 lists the sample size and number of variables for the 2012 datasets. A brief introduction to the structure of each dataset is given below.

Table 1 Sample Size and Number of Variables in Each of the Five CFPS 2012 Datasets[1]

Datasets	Sample Size	Number of Variables
Adult	35719	1740
Child	8621	811
Family Roster	55014	319
Family	13315	685
Cross-Wave ID	61423	18

1.1 Adult Dataset

The adult dataset includes adult respondents aged 16 and above, who were either repeat gene members from the 2010 baseline survey or new adult family members (gene or core members) added in 2012. These respondents were interviewed either face-to-face (IWmode=1) or by phone (IWmode=0), using long-form (longform=1) and/or short-form (shortform=1) questionnaires. Note that although most respondents received only one form of the questionnaire, it is possible for certain respondents to have both the long- and the short-form questionnaires (see Section 2.2 for more

[1] The statistics in this report are based on version 5.0 or 5.1 of CFPS2012 data.

details). The adult dataset includes 34,425 original gene members from the 2010 baseline survey, 23 new gene members in 2012, and 1,271 core members, who are immediate relatives of the gene members but are not gene members themselves.

1.2 Child Dataset

The child dataset includes children aged between 0 and 15 who were either repeat gene members from the 2010 baseline survey or new child family members (gene or core members) added in 2012. Data on children under 10 were collected through proxy reports by their adult guardians, whereas data on children aged 10 or above were collected through both self-reports and proxy reports by their guardians. Again, the interviews were completed through either face-to-face or phone interviews. The child dataset includes 7,254 original gene members from the 2010 baseline survey, 1,264 new gene members added in 2012, and 103 core members.

1.3 Family Dataset

The family dataset treats family as the unit of analysis. It covers the original families interviewed in 2010, as well as the new families that were split from original families since 2010 due to changes in marital status, children's economic independence and other reasons. It includes 12,625 original families and 690 newly split families.

1.4 Family Roster Dataset

The family roster dataset consists of individual level information on all the family members who were listed in the 2012 family rosters. The 2012 family roster dataset includes 55,014 observations from 13,453 households. It should be noted that these 55,014 observations do not represent 55,014 unique individuals. This is because the dataset lists the information of those who split from their original families twice (once as a member of the original family and once as a member of the new family). A dichotomous variable, "co_a12_p", is coded 1 if a "pid" (individual-level unique ID) is financially connected with a "fid12" (family ID in 2012), and 0 otherwise. By CFPS definition, an individual can belong to only one family at any given time point, so for any "pid," co_a12_p=1 for only one "fid12." Therefore, when the family roster dataset is restricted to individuals with co_a12_p=1, there would be no duplicate individual records. The rationale for retaining such duplicated individual records in the family roster data is that this better preserves the dynamic changes in family compositions between 2010 and 2012. Researchers may refer to the forthcoming document, entitled "Decomposing Family Roster Data and Constructing Family

Relations in the CFPS" (in Chinese) for details. The family roster dataset also recorded 639 deaths between 2010 and 2012. It includes 40,757 gene members (75.8%), 1,354 new gene members (2.5%), 2,713 new non-gene members (5.0%), 5,488 people who were not living in the household physically but were still financially connected to the family (10.2%), and 2,672 new gene members (5.0%) living in the family. By study design, a small number of gene members who were abroad, in prison, or for other reasons could not be followed in 2012 (n=136), are retained in the family roster for future tracking.

1.5 Cross-Wave ID Dataset

The cross-wave ID dataset includes all CFPS members (gene members, core members, and non-core member) in any wave, regardless of their interview status in any particular wave. Compared with the family roster dataset, the cross-wave ID dataset includes more individuals (e.g., those deceased cases and those who were not interviewed in a particular wave). The main purpose of the cross-wave ID dataset is to record the basic demographic information on all the CFPS individual participants and their interview status in every wave.

2. Changes in Data Structure from CFPS 2010

Two important changes were made in the field operation of the 2012 follow-up survey compared to the 2010 baseline, causing changes to the 2012 data structure.

2.1 Combining Face-to-face and Telephone Interview Data

In CFPS 2010, all the interviews were done in person. In CFPS 2012, telephone interviews were conducted whenever face-to-face interviews could not be completed. Telephone interviews account for about 0.89%, 2.8% and 0.44% of all the family, adult, and child interviews respectively.

The content of the telephone interview was quite similar to that of the face-to-face interview, except that certain infeasible questions (such as cognitive tests) were dropped in the former. In addition, certain complex questions were also simplified in the telephone interview to maintain a high response rate. In order to make the publicly released data more user-friendly, we have merged the telephone interview data and the face-to-face interview data and have added a dichotomous variable (IWmode) indicating whether the interview was conducted face-to-face (=1) or over the telephone (=0). It should be noted that although the face-to-face telephone interviews shared the same goal, the actual survey questions and hence variables were sometimes different. When combining the telephone and face-to-face interview data, we compared each question from both sources. We adopted the same variable name if the two sources employed exactly the same question and otherwise kept two different variables.

2.2 Combining Long and Short Forms of Questionnaires in Individual-level Data

In CFPS 2010, all the adult respondents were administered the same (long form) questionnaire. In CFPS 2012, a short-form adult questionnaire was added to supplement the long-form questionnaire so that when a gene or core member was away from home, some basic information about him/her would be collected through proxy-report by a local family member using the short-form questionnaire. Meanwhile, we would try to track the non-coresident member at his/her current residence, and if we succeeded, a self-report would be collected using the long-form questionnaire. Also, non-core family members were interviewed using the short-form questionnaire. In short, CFPS respondents might receive only the long-form

questionnaire (for most of the observations in the individual dataset) or both the short- and long-form questionnaires (for non-coresident family members successfully followed up in their current residence) or only the short-form questionnaire (for non-core members, whose data were not publicly released and non-coresident family members who could not be tracked). The proportions of those in which only either the long or short questionnaire were used in the adult dataset were 87.9% and 7.2% respectively, and the remaining 4.9% were interviewed using both questionnaires. The corresponding figures are 95.6%, 2.8% and 1.6% respectively for the child dataset.

Since the short-form questionnaires were mostly answered by proxy respondents, questions on such topics as subjective test and cognitive test were not included. Most of the questions in the short-form questionnaire were directly taken from the long-form questionnaire, so the same variable names were adopted. When we combined the two sources of data, we kept the self-reported data from the long-form questionnaire if both types of questionnaire were available for the same respondent. There were two conditions under which we retained the data from the short-form questionnaires when the long-form questionnaires were also available: (1) when a variable had a missing value in the long-form but not in the short-form questionnaire; and (2) when a variable was measured in different ways between the long- and short-form questionnaires (e.g., different question wording). In the latter case, we would retain both variables and assign a missing value to the one that was not directly collected from the targeted respondent.

3. Composite Variables

We have constructed a number of composite variables to facilitate data analysis.

3.1 Family Income (in the Family Dataset)

We have constructed multiple variables related to family income that suit different research needs, including total family income, family income per capita, and source-specific income (e.g., wage income, business income, transfer income, property income, etc.). These composite income variables are listed in Table 2. Because of several changes in the survey questions regarding income in CFPS 2012 as opposed to those in CFPS 2010, we have constructed another set of 2012 income variables that are comparable to those in 2010. We recommend using these comparable income variables in cross-wave comparative analysis.

3.2 Family Expenditure (in the Family Dataset)

Family expenditure variables include total family expenditure and specific types of expenditure. Total family expenditure consists of transfer, welfare, house mortgage and regular consumption of items such as food, clothes, etc. (see Table 2).

3.3 Community Type (in the Family, Adult, and Child Datasets)

In CFPS 2012, we provide three composite variables that capture the types of communities where respondents reside. The first variable refers to the Census Bureau's definition of whether a community was classified as urban or rural in 2012 ("urban12"). The other two variables are based on respondents' self-report and interviewer's observation: "urbancomm" is a dichotomous indicator of urban versus rural, and "typecomm" further distinguishes among four types of communities: city, town, village and suburb.

Table 2 Composite Variables in the Family Dataset

Variable Name	Variable Label
Family Income	
WAGE_1	Wage income
WAGE_2	Wage income (comparable with year 2010)
wage_1_adj	Wage income - Adjusted
wage_2_adj	Wage income - Adjusted (comparable with year 2010)

foperate_1	Business income
foperate_2	Business income (comparable with year 2010)
ftransfer_1	Transfer income
ftransfer_2	Transfer income (comparable with year 2010)
fproperty_1	Property income
fproperty_2	Property income (comparable with year 2010)
FELSE_1	Other income
FELSE_2	Other income (comparable with year 2010)
FINCOME1	2011-2012 net family income
FINCOME2	2011-2012 net family income (comparable with year 2010)
fincome1_adj	2011-2012 net family income - Adjusted
fincome2_adj	2011-2012 net family income - Adjusted (comparable with year 2010)
fincome1_per	2011-2012 net family income per capita
fincome2_per	2011-2012 net family income per capita (comparable with year 2010)
fincome1_per_adj	2011-2012 net family income per capita - Adjusted
fincome2_per_adj	2011-2012 net family income per capita - Adjusted (comparable with year 2010)
fincper_p	Percentile of family income per capita
fincperadj_p	Percentile of family income per capita - Adjusted
Family Expenditure	
PCE	Residents' consumption expenditure: Sum
FOOD	Expenditure on food: Adjusted
DRESS	Expenditure on clothing (yuan)
HOUSE	Expenditure on housing: Adjusted
DAILY	Expenditure on family equipment & daily necessities: Adjusted
MED	Medical and fitness expenditure
TRCO	Expenditure on communication and transportation: Adjusted
EEC	Expenditure on education and entertainment
OTHER	Other expenditure on consumption
EPTRAN	Transfer expenditure
EPWELF	Welfare expenditure-imputed
MORTAGE	Mortgage on housing: Estimate
EXPENSE	Total family expenditure
Other	
typecomm	Type of village/residential community (simplified)
urbancomm	Urban or rural (according to community type)
urban12	Urban area (Census Bureau's definition in year 2012)

3.4 Personal Income (in the Adult and Child Datasets)

We have constructed both unadjusted (“income”) and adjusted (“income_adj”) personal income variables. The main difference between the two is that we carried out imputations for persons whose wage income was 0 or missing for adjusted personal income.

3.5 Educational Level (in the Adult and Child Datasets)

Two composite variables on education are available: the highest level of education attained in 2012 (“edu2012”) and the highest level of school attended in 2012 (sch2012).

3.6 Employment Status (Adult Dataset)

Each adult’s employment status (“employ”) is classified as “employed,” “unemployed,” or “economically inactive” at the time of interview.

3.7 Cognitive Function (in the Adult and Child Datasets)

Cognitive function was measured through a word recall test and a number series test in 2012. The number series test involved an adaptive two-stage design, in which scores in Stage-I test items determine the difficulty level of Stage-II test items. We provide summary scores for both the word recall test and the number series test in the adult and child datasets. Due to its complex design, we have constructed two summary scores for the number series test: a Guttman score, based on original questions, and a W-score, based on the Rasch model of Item Response Theory. For more information about the number series test, see the CFPS technical report CFPS-31 (in Chinese).

Table 3 summarizes the aforementioned individual-level composite variables. 3.

Table 3 Composite Variables in the Adult and Child Datasets

Variable Name	Variable Label
Personal Income	
INCOME	Personal income
INCOME_adj	Personal income: Adjusted
Educational Level	
edu2012	Highest level of education attained in 2012
sch2012	Highest level of education attended in 2012
Cognitive Function	

NS_G	Number Series Test: Guttman scale (0-15)
NS_W	Number Series Test: W-score
NS_WSE	Number Series Test: Standard error of W-score
IWR1	Immediate Word Recall: 1st trial
IWR2	Immediate Word Recall: 2nd trial
IWR	Immediate Word Recall: Combined 1st and 2nd
DWR	Delayed Word Recall
Other	
Employ	Employment status

Note: Detailed information on constructing these composite variables will be available in corresponding technical reports.

4. Best Variables

Similar to 2010, we have also constructed a number of variables ending with "_best" in the publicly released 2012 data. The CFPS research team recommends these as the "best" among all other related variables when used in secondary data analysis. These best variables were created under the following circumstances: 1) when information on a variable came from multiple sources--for example, when information on gender and age was recorded in both the family roster questionnaire (proxy report) and the individual questionnaire (self-report) and might not be identical in the two sources; 2) when suspicious extreme values were identified during the data cleaning process. To deal with inconsistency across different data sources, we have adopted the basic principle that self-reports are preferred to proxy reports unless we have other reasons to believe that the former lack credibility. To deal with suspicious values for a given variable, we have borrowed information from other related variables. For instance, some asset values were initially deemed too high or too low, and we then looked at related variables such as types and total areas of real estate, in-house facilities, and family income and family expenditure, to assess the validity of the recorded values. Detailed information on property data cleaning can be found in the CFPS 2012 technical report "China Family Panel Studies: the Property Values in CFPS 2012 and CFPS 2010" (CFPS-29 in Chinese).

5. Variable Coding

5.1 Occupational and Industry Codes

CFPS 2012 collected respondents' detailed work information, including family agriculture work, agriculture work for other families, non-agriculture waged job, non-agriculture self-employment, private business, and unpaid help with family business. Most of the work-related variables contain responses recorded in verbal descriptions, which had to be recoded for data release. The detailed procedure on how the original verbal descriptions of responses were transcribed into standard and numeric occupational codes and industry codes can be found in the technical report CFPS-8 (in Chinese).

The design of the work module in CFPS 2012 is different from that in CFPS 2010. For the convenience of cross-year analysis, we have created another composite variable: Occupation of main job in 2012 ("job2012mn_occu"). Detailed information on constructing this variable can be found in the CFPS 2012 technical report "China Family Panel Studies: the Occupation of Current Main Job" (CFPS-30 in Chinese).

5.2 Address Codes

Similar to the 2010 baseline survey, address information in all the 2012 datasets consists of 3-level codes: province code, district/county code and village/community code. The province code follows the national standard, providing real province identifier; the district/county code and village/ community code are both recoded pseudo values so that the true county and community names are kept anonymous for confidentiality. For those who are interested in examining the possible contextual effect at the county level, a county-level dataset which includes a number of macro-variables (e.g., GDP, total population, employment rate, education levels etc.) is available upon further request and application. The detailed information can be found in the CFPS 2010 technical report "China Family Panel Studies: the Conversion of District/County Data" (CFPS-23 in Chinese).

6. Weighting Variables

CFPS 2012 provides both cross-sectional and panel weights at the family and individual levels. Table 4 presents the complete list of weighting variables for CFPS 2012.

Table 4 Weighting variables in CFPS 2012

Variable Name	Variable Label
Adult and Child datasets	
rswt_natcs12	Cross-sectional weight (individual level): total sample
rswt_rescs12	Cross-sectional weight (individual level): nationally representative subsample
rswt_natpn1012	Panel weight (individual level): total sample
rswt_respn1012	Panel weight (individual level): nationally representative subsample
Family Dataset	
fswt_natcs12	Cross-sectional weight (family level): total sample
fswt_rescs12	Cross-sectional weight (family level): nationally representative subsample
fswt_natpn1012	Panel weight (family level): total sample
fswt_respn1012	Panel weight (family level): nationally representative subsample

Total sample includes all families or individuals in the CFPS. The nationally representative subsample is a subsample from the total sample, where only parts of the five "large provinces" subsamples and the full 20 "small provinces" subsamples are included. For details about the sampling strategy of CFPS, users may refer to the User's Manual of CFPS 2010 (http://www.isss.edu.cn/cfps/EN/Documentation/js/236.html).

Cross-sectional weights are used to compute national estimates in 2012. Panel weights should be used when conducting longitudinal analysis; and they are available only for the families and individuals that participated in both waves.