CFPS China Family Panel Studies

China Family Panel Studies

User's Manual

(3rd edition)

Yu Xie, Xiaobo Zhang, Ping Tu, Qiang Ren,

Yan Sun, Ping Lv, Hua Ding, Jingwei Hu, Qiong Wu

7/30/2017

CONTENTS

Foreword

After several years of preparation and two pilot studies in 2008 and 2009, China Family Panel Studies (CFPS) implemented its baseline survey in 2010 and three waves of full sample follow-up surveys in 2012, 2014, and 2016. In addition, a small-scale sample maintenance survey was conducted in 2011. Large-scale surveys are known to be complicated and involve multiple details in the initial conceptual design, survey technologies, interviewing processes, quality control, and data processing. Each individual aspect influences the academic value of the data. The CFPS baseline sample covers 25 provinces/municipalities/autonomous regions, representing 95% of the Chinese population. The 2010 baseline survey interviewed a total of 14,960 households and 42,590 individuals, and it is China's first large-scale academically-oriented longitudinal survey project. It aims to become the most authoritative survey project on Chinese family and society. While we are delighted to have collected several waves of comprehensive, high-quality, and valuable social science data, this accomplisment has come with a cost: due to the complexity in the design and implementation of the survey, and the data structure, users may experience difficulty in utilizing the data. Therefore, we have published this User's Manual to provide detailed information for the usage of the data in a systematic and user-friendly way. The first and second editions focused on the baseline survey and its database. This edition adds an introduction about the design, implementation and dataset of the follow-up survey. Specifically, we cover the following topics:

1. Conceptual design and methods, including those pertaining to sampling, weighting, survey instrument, questionnaire design, and follow-up strategies, etc.
2. Details of the actual implementation, such as map drawing, residence and household screening, standardization of interviewing procedures, data quality control, sample maintenance, etc.
3. Data management and datasets, including the structure and content of the data sets, data cleaning, construction of composite measures, occupational codes, etc.
4. Technical report index. Regarding certain substantive topics and professional fields, we provide a series of separate technical reports for users to better understand our project and our data. This manual refers to and cites some parts of the technical reports, but does not elaborate on the content of the technical reports. Instead, an index of the reports is provided toward the end of this manual.
5. Data quality assessment. Based on comparisons with Chinese censuses and other data, we provide a brief evaluation on the quality of the CFPS surveys.

This manual draws extensively on the conference minutes, official documents, manuals, and technical reports provided by the staff and researchers from the Institute of Social Science Survey (ISSS) at Peking University. The tables and graphs in the

preliminary findings and evaluations were prepared by Chunni Zhang, Qi Xu, Xiang Zhou, Hongwei Xu, and Guoying Huang. The first and second edition of the manual was organized and edited by Jingwei Hu and proofread by Chunni Zhang. The third edition was organized and edited by Qiong Wu and proofread by Jingwei Hu. The English version was proofread by Cindy Glovinsky. In addition, Xin Zhang, Wangyang Li, Wenshan Yu and Yongai Jin assisted with the writing of this manual.

We sincerely hope that data users find this manual helpful. We will continue to publish updated versions of this manual when changes are needed due to data or documentation updates. Please let us know if you identify any problems with this manual. Suggestions for improvement are also welcome.

Acknowledgements

A large number of people have made great and selfless contributions to the CFPS survey. Our accomplishments are the results of a collective effort. Here we would like to acknowledge and express our appreciation for the entire CFPS team.

Led by Peking University, many researchers and scholars from home and abroad joined the design of the CFPS baseline survey questionnaires. The main contributors are Jianjun Bai, Yude Chen, Yuyu Chen, Xiaohao Ding, Jiafeng Gu, Zhigang Guo, Guitian Huang, Guoping Li, Jianxin Li, Qiang Li, Shiding Liu, Yunfeng Lu, Xiaochun Qiao, Zeqi Qiu, Mingming Shen, Yan Shen, Yan Sun, Ping Tu, Qiong Wu, Qun Xiao, Yu Xie, Xianglin Xu, Jie Yan, Boxu Yang, Yang Yao, Ruijun Yuan, Changjun Yue, Chunni Zhang, Qianfan Zhang, Tuohong Zhang, Xiaobo Zhang, Yaohui Zhao, Xiaolin Zhou, Yanhui Zou, He Cai (Sun Yat-Sen University), Youde Guo (Fudan University), Hong Lei (Huazhong University of Science and Technology), Lulu Li (Renmin University of China), Peilin Li (Chinese Social Science Academy), Shi Li (Beijing Normal University), Youmei Li (Shanghai University), Jingming Liu (Tsinghua University), Yuzhao Liu (Shanghai University), Liping Qiu (Shanghai University), Jingxian Ren (Tsinghua University), Song Zhe (Tsinghua University), Guangzhou Wang (Chinese Social Science Academy), Zhengwei Wang (Tsinghua University) , Wenfang Tang (The University of Iowa), Dingjun Weng (Shanghai University), Xiaogang Wu (Hong Kong University of Science and Technology), Weiqiang Zhang (Tsinghua University), Congyi Zhou (Tsinghua University), Jianhua Zhu (City University of Hong Kong), Colette Browning (Monash University), Michael Carter (University of Wisconsin-Madison), Robert Hauser (National Research Council and the University of Wisconsin-Madison), David Lam (University of Michigan), James Lepkowski (University of Michigan), Arland Thornton (University of Michigan), Donald Treiman (University of California-Los Angeles), Nora Schaeffer (University of Wisconsin-Madison), Robert Willis (University of Michigan), and Jean Yeung (National University of Singapore). The CFPS questionnaires cover almost all fields of social science research. We would like to thank the abovementioned researchers and scholars for their constructive opinions and suggestions, which were invaluable to enriching, improving, and optimizing the contents of the questionnaires.

At the same time, we also thank all those who had contributed to the implementation of the survey. During the busy survey seasons, they have often worked overtime, even on holidays, to effectively deal with all kinds of problems that have emerged from the field. We owe the successful implementation to the hard work by He Cai, Yulong Cao, Minyan Chen, Wei Cong, Jiapo Chen, Lijuan Ci, Hua Ding, Xinxing Ge, Bin Ge, Chunjie Gu, Jiafeng Gu, Zhenwei Guo, Junli Han, Changqun Huang, Yang Hong, Danli Jia, Xiaojing Jia, Guohua Li, Ran Li, Li Li, Shengwen Li, Youmei Li, Yucheng Liang, Yue Liu, Ping Lv, Jie Lv, Yun Ma, Tengyu Ma, Wenting Ma, Chao Ma, Xia Meng, Dejin Peng, Ping Qian, Zeqi Qiu, Xinchen Qiu, Liyin Ren, Shibin Song, Wei Si, Yufang Shen, Na Shao, Ting Sun, Shuai Sun, Yan Sun, Yi Sun,

Yuhuan Sun, Caiqin Sun, Xueliang Teng, Xu Yang, Rong Shen, Ting Wan, Tao Wang, Yanmei Wang, Qiyao Wang, Jing Wang, Kun Wang, Xueyin Wang, Xiaowen Wei, Qi Xu, Jie Yan, Qian Yang, Sijia Yang, Jiahui Yao, Xue Ye, Jing Yi, Wenmao Yin, Shuang Yu, Xuejun Yu, Haobin Zang, Haidong Zhang, Lanxin Zhang, Man Zhang, Yaxin Zhang, Yongjian Zhang, Yun Zhou, Hongping Zhou, Yingying Zhou, Tingwei Zhu, Chenling Zhu, and Yanhui Zou.

Due to the large sample size and various types of data, the management and cleaning of CFPS data have involved a large workforce. The CFPS data team has worked extremely hard to provide reliable and user-friendly data, for which we would like to thank Ling Bai, Yahong Cui, Jia Chen, Lihong Dai, Jingwei Hu, Guoying Huang, Yongai Jin, Li Li, Wangyang Li, Weixiang Luo, Ping Lv, Chao Ma, Sha Ni, Liyin Ren, Qianping Ren, Qiang Ren, Yuhuan Sun, Zhibo Tan, Jia Wang, Longyu Wang, Xueyin Wang, Xiao Wang, Yulei Wang, Lingwei Wu, Qiong Wu, Jun Xiang, Yu Xie, Hongwei Xu, Qi Xu, Jie Yan, Xu Yan, Jiahui Yao, Jia Yu , Chunni Zhang, Jingshen Zhang, Cong Zhang, Wenjia Zhang, Xin Zhang, Duan Zhao, and Fangyuan Zhao.

The implementation of CFPS has been assisted by the former National Population and Family Planning Commission, State Bureau of Statistics, Ministry of Civil Affairs, Shanghai University, and Sun Yat-Sen University. The Institute for Social Research at the University of Michigan, an important collaborative unit, has provided us with much guidance and advice on survey design and technological support. Our survey is also supported by the National Natural Science Foundation of China and Peking University, to whom we are deeply grateful.

Finally and most importantly, we would like to thank the interviewers, numbering over 2000, who conducted interviews in various study sites across the country throughout these years. Fieldwork is the most important but laborious part of survey research. Our interviewers have overcome numerous transportation difficulties and achieved excellence in their work. We are also indebted to our respondents, whose understanding and cooperation made our study possible. Without our respondents, it would not have been possible to acquire such precious empirical material that truly reflects China's social conditions.

1. Introduction

1.1 Background

China Family Panel Studies (CFPS) is a national longitudinal general social survey project. By collecting data at three levels (i.e., individual, family, community), the project aims to document changes in Chinese society, economy, population, education, and health, so as to provide data for academic research and public policy analysis (Xie, Hu & Zhang, 2014; Xie & Hu, 2014).

CFPS focuses on both the economic and non-economic well-being of the Chinese people, covering substantive areas such as economic activities, educational attainment, family relationships and dynamics, population migration, and physical and mental health. The target sample of CFPS consists of 16,000 households in 25 provinces/municipalities/autonomous regions in China (excluding Hong Kong, Macao, Taiwan, Xinjiang, Tibet, Qinghai, Inner Mongolia, Ningxia and Hainan) (Xie & Lu, 2015). All eligible households and household members are subjects of the survey. An eligible household refers to an independent economic unit that lives in a residential community with one or more family members of Chinese nationality.[1] Family members are defined as financially dependent immediate relatives,[2] or non-immediate blood/marital/adoptive relatives who have lived with the household for more than three consecutive months and are financially related to the sampled household.

The preparatory work of CFPS started in 2007. Pilot studies were conducted with 2,400 households in Beijing, Shanghai and Guangdong, including a baseline survey in 2008 and a follow-up survey in 2009. In 2010, the baseline national survey was officially launched in 25 provinces/municipalities/autonomous regions. We visited 19,986 households and successfully interviewed members of 14,960 households. Within these households, a total of 33,600 adults and 8,990 youths were interviewed. At the household level, the response rate, cooperation rate, contact rate, refusal rate are 81.25%, 96.58%, 84.13%, 2.67%, respectively. At the individual level, the rates are 84.14%, 87.01%, 96.7%, and 8.47%, respectively. [3]

For the CFPS 2010 baseline survey, face-to-face interviews were conducted with the sampled households' family members living in the sampled communities. Family members who were elsewhere in the same county were also interviewed. For those who were not present at home at the time of interview, basic information was collected from the present family members. All family members who were identified at baseline to have blood/marital/adoptive ties with the household were identified as CFPS gene members. In the follow-up surveys, newly born or adopted children of

[1] Initially at least one family member had to live in the sampled community for six months consecutively to qualify for our survey. We later dropped this requirement during survey implementation because most of our sampled households fulfilled it.
[2] See Sun et al. (2011) for definition of immediate family members
[3] Based on AAPOR estimation. See Technical Report: CFPS-5.

gene members were also considered CFPS gene members. All gene members would be tracked in the follow-up survey. In the follow-up survey, gene members' non-gene immediate relatives (i.e., parents, spouses, children) in the same household were defined as the core members of the household in the survey year. Family members who were neither gene members nor core members were defined as non-core members. In CFPS, only gene members are to be tracked permanently; core members are interviewed when they maintain their ties with gene members. Non-core members are not interviewed, but their basic information is collected in the family roster questionnaire.

The CFPS 2010 baseline survey used face-to-face interviews aided by Computer Assisted Personal Interviewing (CAPI) technology. Subsequent surveys starting from CFPS 2012 used a mixed mode, in which CAPI was supplemented by Computer Assisted Telephone Interviewing (CATI) technology. CFPS 2010 used five questionnaires targeted at the community, family members, households, adults, and youths. The CATI questionnaire for CFPS 2012 and CFPS 2014 was a simplified version of the corresponding CAPI questionnaire. CFPS 2016 began using the same questionnaire for both CAPI and CATI, with the exception that the cognitive module was available only in the CAPI mode. Also, starting with the CFPS 2012 follow-up survey, proxy questionnaires were used to collect information about family members who were not physically present in the household at the time of the interview.[4]

In addition to the full-sample follow-up every two years, CFPS conducted a small-scale sample maintenance survey in 2011. This report focused on the baseline survey and the full-sample follow-up surveys. The maintenance survey in 2011 is only briefly discussed.

CFPS was designed by the Peking University research team and supported by Peking University and the National Natural Science Foundation of China. The Institute of Social Science Survey (ISSS) at Peking University is responsible for its implementation, and has received great support from former Chinese National Population and Family Planning Commission and Ministry of Civil Affairs.

1.2 Design

1.2.1 Social Changes in China[5]

China has been undergoing a social transformation in which scope, rapidity, and influences have been unprecedented in human history. China's ongoing social transformation since the late twentieth century is no less consequential in the course of world history than events considered historical watersheds, such as the Italian

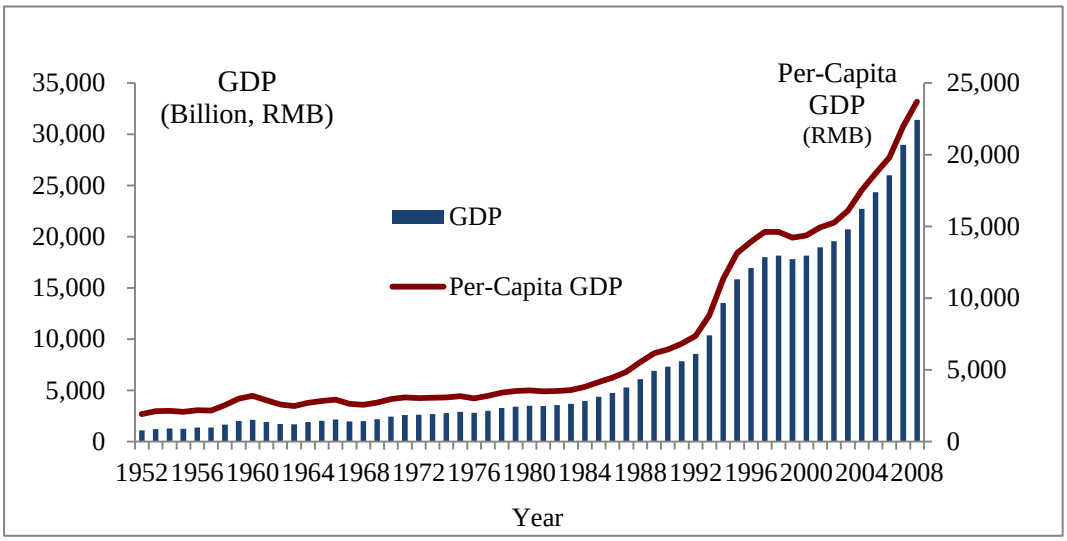
[4] Not physically present in the household means the family member is financially connected to the family (defined in CFPS as a family member) but does not live at the same address.
[5] Sections 1.2.1、1.2.2、and 1.2.3 draw partially on Xie (2011).

Renaissance, the Protestant Reformation, or the British Industrial Revolution. The rapid, large-scale, and irreversible social changes that have been occurring in China in recent decades are multifaceted. We may gain a better appreciation of these social changes by looking at three highly important aspects:economic growth, education expansion, and demographic transition.

Chinese economic output has grown tremendously and sustainably since the 1980s. Figure 1 shows a sharp increase in total gross domestic product (GDP) and GDP per-capita after the initiation of economic reforms in 1978. The inflation-adjusted GDP per-capita increased at an annual growth rate of 6.7% between 1978 and 2008. Compared to the rapid, yet sustained, economic development in China, the 1.5% annual growth rate in inflation-adjusted GDP per-capita during the Gilded Age in the United States was significantly smaller.[6]

In addition, educational attainment in China has significantly improved in recent years, most dramatically at the postsecondary level. Figure 2 shows trends in the numbers of enrolled and newly admitted college students. While the figure shows a gradual increase in college enrollments over time—except for a downturn associated with the Cultural Revolution (1966–1976)—the enrollment rate started to increase drastically in the late 1990s. The rapid increase in the number of young Chinese receiving higher education is both a cause and a consequence of China's tremendous economic growth in recent decades.

[6] Calculation based on data from Measuring Worth (2011) http://www.measuringworth.com/.



Note: Adjustment has been made for the data of 2005–2008, on the basis of the 2d Economic Census.

Sources: State Bureau of Statistics. 2010. *China Statistical Abstracts 1949-2008*. China Statistics Press. State Bureau of Statistics. 2010. *China Statistical Abstracts 1949-2008*. China Statistics Press.

Figure 1. Trends in GDP and Per-Capita GDP, 1952–2008 (in 2008 RMB).

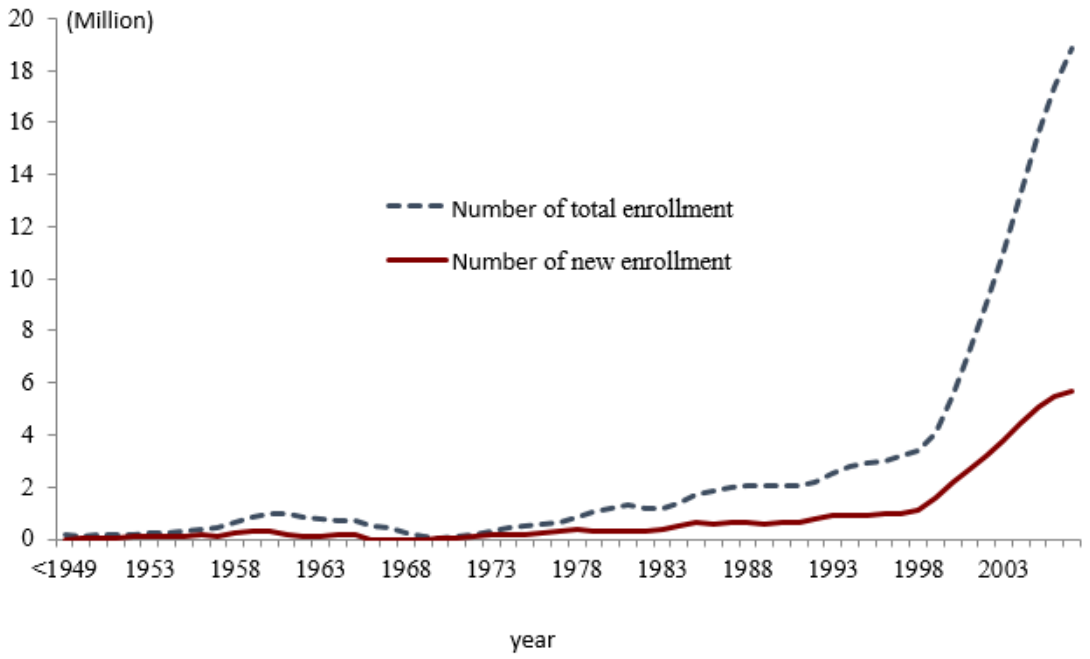
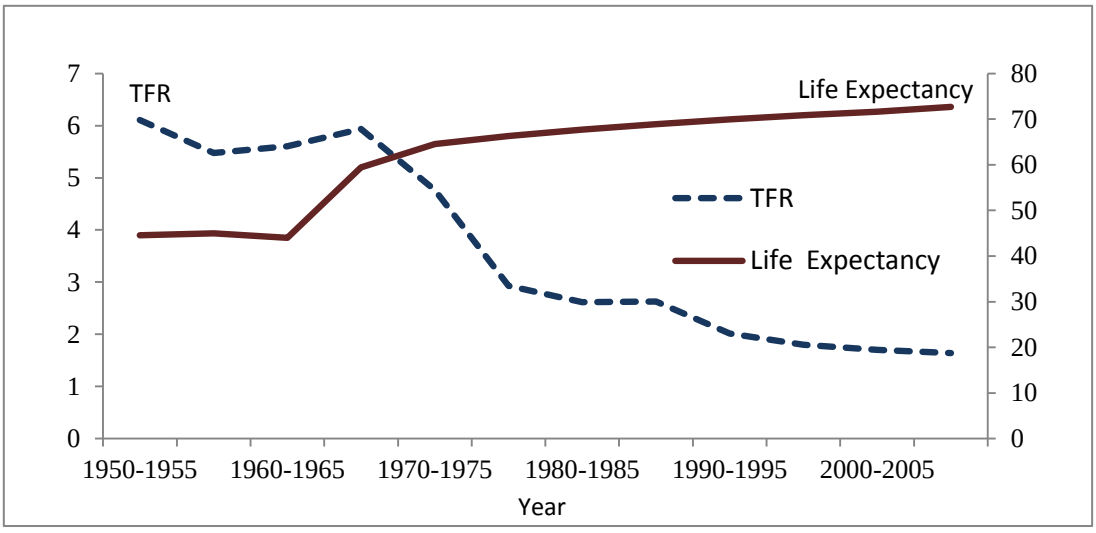


Figure 2. Higher Education Expansion in China, 1949–2007. Source: China Education Yearbook Editorial Board (1984, 1986–1988, 1989–2008).

Furthermore, China has completed its demographic transition from a high-fertility, high-mortality pattern to a low-fertility, low-mortality pattern over the past few

decades. Clearly in Figure 3, there has been a sharp decline in total fertility rate (TFR) since the late 1970s, as well as a steady improvement in life expectancy since the 1950s. By 1990–1995, China had already completed its demographic transition, with TFR having fallen to replacement level at 2.0, and a life expectancy of 70 by around 1970, which is comparable to demographic trends in developed countries and way ahead of those in other underdeveloped countries.



Sources: United Nations, Department of Economic and Social Affairs, Population Division (2011). World Population Prospects: The 2010 Revision, CD-ROM Edition.

Figure 3. Total Fertility Rate and Life Expectancy, 1950-2010.

The aforementioned social changes are probably the most significant ones in China over the past thirty to forty years. Of course, many other important social changes have taken place that merit social science research. Examples include increasing social inequality, rising divorce and premarital cohabitation, and massive labor migration. It is CFPS's goal to provide social scientists with high-quality, comprehensive, and longitudinal data to conduct in-depth research on such social phenomena and their changes.

1.2.2 The Importance of Empirical Research in China

Research on the unique social changes occurring in China not only provides valuable findings, but also presents the potential to change the course of world development. For the last three centuries or so, the West has been leading the world up an ascendant path, which has commonly been referred to as "modernization," "development," or "progress." The premises behind the development of Western countries are that (1) democracy is the only legitimate system and (2) the free market is the only viable economic system. For the first time in three centuries, China is now

presenting a serious challenge to the West as this modernization theory fails to explain the rapid development of China. China neither is a democratic country nor operates under a proper free market economic system. However, its economy has been developing rapidly and steadily for more than 30 years. In contrast, Western economies have not seen steady growth, with severe recessions taking place, such as the Great Recession in the US in 2008. Is it possible that China's model may also be a feasible, perhaps even superior, path to development?

China's developmental model is an interesting topic that drew our attention immediately. The study of such social science issues requires objective judgment without prejudice, which is often grounded in projections of experiences from other countries or from theoretical speculations. We have to understand China's situation from its history, culture, politics, and economy, and to analyze it based on empirical evidence rather than purely on speculation.

Social theory is best when based on a social context, and social research is best when conducted within a specific social context. Thus, the large-scale rapid social changes in contemporary China, which have unique characteristics compared to existing and previous social reforms, are especially worth studying. China's current development may be considered a social phenomenon shaped by its current political, economic, cultural, and social environments. Therefore, we can expect that innovative frameworks, both theoretical and methodological, will be designed specifically for studying the social phenomena in contemporary China.

Furthermore, social science in China is experiencing rapid changes along with Chinese society. Empirically-based research is developing and gaining importance among social scientists over traditional opinion-based, ideological and speculative discussions. This trend is inevitable. The public, government, and academy all welcome higher quality empirical studies, as these studies serve their interests. What is going on in China exceeds the public's understanding; therefore, many people are interested in learning about the changes that impact their daily lives. Also, Chinese policy makers benefit from having more accurate information and evidence to make more rational policy decisions. Finally, China is one of the world's most powerful countries contributing to the arts, sports, finance, the natural sciences, technological advances, and world peace. We sincerely hope that our empirical social studies in China will not only make significant contributions to social science, but also receive world-wide recognition and appreciation.

1.2.3 Design Goals of CFPS

The massive social transformations in China have generated both challenges and opportunities for today's social scientists. To completely understand Chinese society, we must first understand these changes. History calls for empirical social science research, and empirical research relies on high quality survey data. From a long-term perspective, it is impossible for contemporary social scientists to fully understand the

ongoing social changes in China. This is exactly why we are making every effort to collect empirical materials: to help future social scientists better understand China today. Thus, for both current empirical research in the social sciences and future research into contemporary matters, we should value this unique opportunity to document the historical changes in China. Such is the principal mission of the CFPS. With this in mind, we designed the CFPS based on the following three social characteristics:

First, social phenomena have multiple domains, but many of these different domains are closely related. Many aspects of social life, such as family background, social network, housing, education, occupation, income, and health, are interdependent.

Second, social phenomena have multiple levels. At the macroscopic level, there are state policies, cultural traditions, and historical events; at the intermediate level, there are regional economies, urban facilities, and community environments; at the microscopic level, there are family structures, intergenerational relations, individual educations and careers, etc. All these factors at different levels contribute to shaping the course of an individual's life.

Finally, social phenomena are continuous over time. Present behaviors depend on past events, and future decisions depend on today's experiences. In fact, seemingly simple social matters, such as population migration and family expenditures, are consequences of complex causal connections and dynamic processes.

Based on the multi-dimensional and multi-level nature of social phenomena, CFPS investigates Chinese society at three levels: at the community level, where we collect macro and intermediate data on political environment, rural landscapes, infrastructures, populations, resources, transportations, health systems, and fiscal conditions; at the family level, where we collect information on family structures and relationships, living conditions, social networks, income and expenditures, and family assets; and at the individual level, where we collect personal information on education, occupation, income, physical and mental condition, and marriage. Thus, the subjects in our study are not isolated but connected. Researchers are able to study the relationship between individuals, families, and the society.

The continuity of social phenomena makes social systems more complicated and more difficult to study than many phenomena. Time is an important concept in studies of social issues and changes. From the perspective of methodology, time is the information that identifies the chronological order of events. For instance, individual behavior varies over time as personal experiences are gained.[7] A panel study tracks these dynamics and captures their variations on the temporal dimension, which is a highly effective approach to studying social trends. By performing a panel study, which involves measuring change over time for the units of analysis within the

[7] Xie (2012).

population, we are able to better understand the casual processes described, and thus to predict population trends. Thus, a panel survey is of great value in important areas of research such as population heterogeneity, causal inference, and status transitions in the social sciences.[8] Despite its high cost, complex designs and difficult implementation, the panel survey provides more information and more valuable materials than cross-sectional and trend studies, thus yielding more rewards for scientific studies. Because of this, CFPS decided to collect data on the target group at different times. This means that we will continue to keep track of the gene members as well as their biological or adopted children every other year.

To conclude, the data of CFPS is collected using multi-dimensional, multi-level and longitudinal methods and aims to provide the most thorough and reliable data for more valuable and scientific research as well as reliable evidence for state policy making.

1.2.4 Research Unit

The nature of social science research is heterogeneity.[9] Individuals, even in the same pool, differ from each other in characteristics. It is known that education level, income, life style, physical condition, and social network vary among the population in China. The existence of heterogeneity leads us to distinguish between individuals. An individual is a basic social unit of human society heterogeneity and a source of social phenomena, such as health, happiness, and work. Thus, in order to understand the society, we need to understand the differences in quality of life, financial status, and social roles between individuals. That is why individuals are the most basic yet important study units and targets of CFPS long-term follow-up surveys .

Genetics and one's background are not the only sources of heterogeneity in quality of life, financial status, and social roles in the population. Social environment and personal experiences also play an important part. Individual differences are affected by social structures, the most important one being the family.

First of all, the family is the principal environment for the socialization of children. A person's ascribed status is determined by his or her family, and the family plays a major role in its children's enculturation of social rules from the moment they are born. The family's influences on personal attitudes, actions, and expectations is long-lasting. Thus, it is essential to understand a person's family in order to study the individual.

Second, the family is a primary site for economic and social interactions in Chinese culture. Many important social interactions, such as financial activity,

[8] Ren & Xie (2011).
[9] Xie (2012).

housing, raising children, and supporting the elderly, occur at the family level. Thus, research on family-related topics is essential for studying Chinese society.

In addition, the family affects relations between generations. How does parents' social status affect the next generation? How are family resources allocated between children? How are resources from grown-up children transferred to parents? A clear and comprehensive understanding of the family and its members is key to answering these questions.

Furthermore, the family is crucial for research on marriage and gender. Couples come from different families, which commonly have different backgrounds, and form new families through marriage or cohabitation, where social status and resources are redistributed and recomposed. Marriage and the family also reflect gender differences in social capital and division of labor.

Last but not least, the family is central to Chinese culture. Chinese people worship their ancestors and observe filial piety. They strive for achievements to glorify their family name. There is a strong norm about forming marriages between families of comparable social status. The desire to continue the family lineage, along the male line, is also paramount. These traditional values indicate the importance of the family in Chinese culture. The family provides significant physical and emotional support for individuals, and in return, children are obligated to pay the family back. Despite the waning of traditional family values in recent decades, the strong significance of the family in Chinese culture remains intact. We can see this in parents' heavy investments in their children, extensive kinship networks, and reliance on informal transfers of financial resources among family members.

In summary, family is essential to understanding Chinese society, and thus it is an important research and investigation unit for CFPS. CFPS conducted an in-depth investigation of information on family relations and family members and built an accurate family network structure which clearly established relationships between family members. Detailed information on family financial and social activities was also collected in our research. We do hope CFPS can provide a broader view and enhance your understanding of Chinese society.

1.3 International Comparison

In the early stages of design, CFPS learned about research approaches and investigation tools from some advanced research programs, including the Panel Study of Income Dynamics (PSID), the National Longitudinal Surveys of Youth (NLSY), and the Health and Retirement Study (HRS), and came up with its own features and advantages which can meet various researcher demands from different academic backgrounds.

The PSID [10] was designed and initiated by the University of Michigan in 1968. It is the most authoritative panel study of family economy in the United States. Its original purpose was to study poverty and the effects of Lyndon Johnson's "War on Poverty" project on economics and the public welfare. Later on, the research topics gradually expanded to employment, income, wealth, housing, food expenditure, transfer payments, marriage, and fertility. The sample size was 5,000 households at the beginning, and only one adult from each household was chosen to be interviewed by telephone. Since 1997, the Child Development Supplement to the PSID (PSID-CDS) has been added to the former study.

The NLSY,[11] designed by the Ohio State University and launched by the National Opinion Research Center (NORC) of the University of Chicago, is an authoritative panel survey program studying American youth on the labor market. The longitudinal surveys involved two cohorts of young people. The 1979 cohort contains 12,868 respondents aged between 14 and 22 when they were first interviewed in 1979, and the 1997 cohort contained 9,000 young people aged between 12 and 16 when they were first interviewed in 1997. The research focused on human capital of the youth and their labor market activities. The survey included education, employment, career trainings, working hours, income and assets, attitude and behavior, health, and political involvement, etc.

The HRS began in 1992. Designed and launched by the University of Michigan, it was the most influential longitudinal program on aging in the United States. This survey was targeted at people above the age of 50 and the sample size was about 26,000. HRS concentrated on the demographic labor force participation and health changes, as well as information about income, work, assets, pension plans, health insurance, disability, physical health and functioning, cognitive functioning, and health care expenditures.[12]

The PSID, NLSY and HRS were all large, nationally representative panel research programs in the United States. Although these studies were monographic, their research was comprehensive and was an important data source for a variety of related scientific research. CFPS benefits from their experiences, and we hope to contribute to social research in China as well. Similar to the surveys above, CFPS consists of a large, nationally representative sample: members of 14,960 households were interviewed in the baseline survey, including 57,155 gene members for long-term follow up survey. As a multi-level, multi-dimensional, and longitudinal research program, CFPS's content is thorough and comprehensive, including various important life events through both childhood and adulthood, and some specifically designed for family relationships, family financials and communities.

[10] For more information on PSID, please refer to http://psidonline.isr.umich.edu/default.aspx.
[11] For more information on NLSY, please refer to http://www.bls.gov/nls/nlsy79.htm, and http://www.bls.gov/nls/nlsy97.htm.
[12] For more information on HRS, please refer to http://hrsonline.isr.umich.edu.

Based on the experiences of the surveys above, CFPS has made some improvements. We believe the multi-level social structure is a critical characteristic of Chinese society, in which family is of utmost importance. Therefore, CFPS conducted a more extensive and intensive survey on family members and their family relationships than the earlier panel studies did. Surveys on family relationships usually focus on one or two core members of a household, collecting information on their family backgrounds and other members of the family, which causes the data collected this way to be limited. In contrast, CFPS asks all family members who are eligible for the research, even children, to complete their own personal questionnaires either by themselves or with the help of others. Thus, the statistics provided by CFPS are more detailed, accurate and complete. Moreover, the unique design of CFPS[13] enables researchers to acquire information not only on the one-dimensional relationship between interviewees and other family members but also on the entire family; not only on immediate relationships but also on cross-generational and sibling relationships; and not only on members living in the household but also on members living apart. With the help of the CFPS individual coding system, researchers can locate family members and obtain an accurate family network. Thus, the provision of more valuable information on family structures and family members is an important advantage of CFPS.

1.4. Survey Technologies

As a national comprehensive survey program, CFPS consists of a large sample size, a wide survey coverage, and a complicated design, which make traditional paper and pencil surveys no longer feasible. CFPS adopted CAPI in 2010, and CATI in 2012. These computer assisted interviewing technologies were adopted to guarantee the efficiency and quality of the survey.

Computer assisted interviews rely on the interview management system, a professional interviewing software. The system can assist interviewers in completing the questionnaire and managing interview data information. Moreover, it makes it more convenient for interviewers to contact headquarters and receive feedback on problems. To be specific, the main features of the interview management system are as follows:

(1) Intricate designs of electronic questionnaires. Different styles of questioning, e.g. multiple choice, single selection, tables, loops, and intervals can be used. Customized questions can also be designed via complicated logical skips in accordance with the characteristics of different groups. With the setting of hard checks and soft checks, electronic questionnaires can give prompt instructions in response to illogical or unreasonable answers, which makes it possible for interviewers to communicate with respondents instantly and make corrections.

[13] It refers to the T table design. It is introduced in detail in the following part of the manual.

(2) Quick management of samples. The headquarters can send samples to interviewers remotely and modify the interviewing tasks according to the practical needs of the survey. The system can also record detailed information on each interviewee and respondents' family, including the address, interviewing method, and payment, so as to make it convenient for supervisors and interviewers.

(3) Real-time data transmission. With the help of the interviewing management system, interviewers can transmit and exchange data with the headquarters instantly. The headquarters can easily track the progress and exercise remote control by spotting the problems in real time and solving them with the interviewers in the field. In addition, the interviewing management system saves researchers from having to input data from traditional surveys using paper and pen, so that the data cleaning and analysis can be done on a contemporaneous basis.

(4) Real-time supervision of interviewer behavior and quality control. The system can supervise the computer operations of interviewers via recording, and the headquarters can inform the interviewers how to improve their non-standard interviews in real time.

(5) Paradata analysis. The interviewing management system is able to collect a set of paradata, such as pause time for every question, records of interviewers' modifications of the options, etc. Analyzing these data will provide scientific evidence for improving the survey research design in the future.

2. Sampling

2.1 Sampling Design

The sample of CFPS is drawn from 25 provinces/cities/autonomous regions in China excluding Hong Kong, Macao, Taiwan, Xinjiang, Xizang, Qinghai, Inner Mongolia, Ningxia, and Hainan. The population of these 25 provinces/cities/autonomous regions in China (excluding Hong Kong, Macao, and Taiwan) includes 95% of the Chinese total population. Thus, CPFS can be regarded as a nationally representative sample.

The original target sample size was 16,000 households. Half of the sample (8,000) was generated by oversampling with five independent sampling frames (called "large provinces") of Shanghai, Liaoning, Henan, Gansu, and Guangdong. Each of the subsamples had 1,600 households. The other 8,000 households were from an independent sampling frame composed of 20 provinces (called "small provinces") (Xie&Lu, 2015) (See Figure 4, Table 1). The "large provinces" were representative of the regional level, which could contribute to provincial population inferences and cross-region comparisons. With second-stage sampling, the five "large provinces," together with the "small provinces," made up the overall sampling frame representative of the national population.[14]

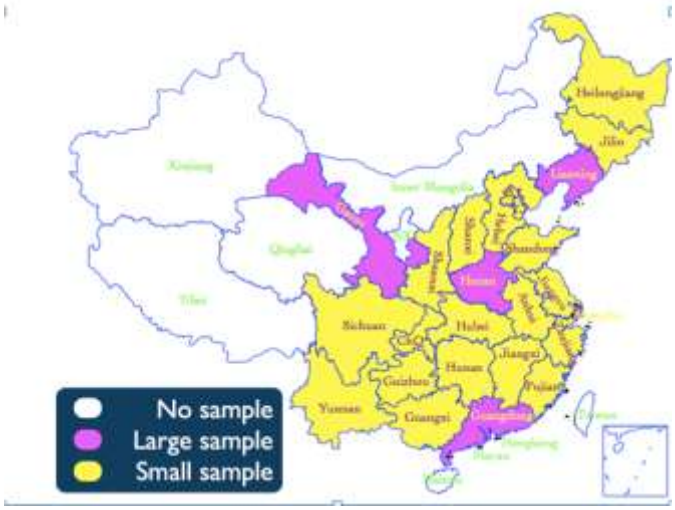


Figure 4. The Sources of CFPS Samples at the Provincial Level

Taking the regional differences in Chinese society and the reduction of survey processing costs into consideration, CFPS implemented Probability-Proportional-to-Size Sampling (PPS) with implicit stratification. Administrative units and socioeconomic status (SES) were used as the main stratification variables. Within the administrative unit, local GDP per capita was used as the ordering index for SES. If

[14] We call this sample "national re-sample" or "constructed sample."

the GDP per capita in the administrative unit is not available, the proportion of non-agricultural population or population density is used.

All the sub-sampling frames of CFPS were obtained through three stages: the Primary Sampling Unit (PSU) consisted of administrative districts/counties, the Second-stage Sampling Unit (SSU) consisted of administrative villages/neighborhood communities, and the third-stage (Ultimate) Sampling Unit (TSU) consisted of households.[15] In the first and second stages, CFPS used official administrative divisions for the sample selection. The third sampling stage was a systematic selection of housing units from street listing with a random starting point and equal probability method. For the consideration of response rate, the 2010 CFPS survey used the estimated response rates from 2008 and 2009 pilot studies as a reference and enlarged the sample size proportionately. A total of 19,986 households were selected according to systematic sampling principles, which ensured the expected sample size for the survey.

It is important to note that CFPS sampled the Chinese population as a whole instead of using traditional sampling methods which sampled urban and rural areas separately. The reason behind this is that the official rural-urban division can hardly reflect the reality of China's rapid urbanization. At the community level, we collected information regarding whether the sampled communities were urban neighborhoods or villages. At the household and individual level, we identified the *hukou* type and whether the household engaged in non-agricultural work or not. Users of such data may decide for themselves whether the community is rural or urban using such information rather than relying on administrative divisions.

For detailed information of sampling design and implementation, see *Sampling Design (CFPS-1).*

Table 1. CFPS Target Sample Size

Category	Provinces/Cities/Autonomous Regions	Target Households	Oversampling Rate
"Large Provinces"	Shanghai	1600	10.28
	Liaoning	1600	4.45
	Henan	1600	2.04
	Gansu	1600	7.30
	Guangdong	1600	2.02
"Small Provinces"	Jiangsu, Zhejiang, Fujian, Jiangxi, Anhui, Shandong, Hebei, Shanxi, Jilin, Heilongjiang, Guangxi, Hubei, Hunan, Sichuan, Guizhou, Yunnan, Tianjin, Beijing, Chongqing, Shaanxi.	8000	1.00

[15] Shanghai is different from other "large provinces," and therefore the sampling procedure was slightly different. More information can be found in the technical report, CFPS-1.

Table 2. Three Stages of CFPS Sampling

Stages	Guangdong, Gansu, Liaoning, Henan: 4 "Large Provinces"	Shanghai: "Large Province"	"Small Provinces"	Total
Primary	4×16 Counties=64 Counties	32 Streets (Towns)	80 Counties	144 Sampled Counties+32 Sampled Streets (Towns)
Second-stage	64×4 Communities =256 Communities	32×2 Communities =64 Communities	80×4 Communities =320 Communities	640 Communities
Third-stage	640×[28, 42] Households			19986 Households

（See Table 3.）

Table 3. Terminal Sample Size of CFPS 2010 Baseline Survey[16]

Region	Category	Expected Response Rate	Contacted Sample Size
Low Response Rate	Neighborhood committee (urban area and village in urban fringe[17])	60%	42
	Other village	70%	36
Intermediate Response Rate	Neighborhood committee (urban area and village in urban fringe)	70%	36
	Other village	80%	32
High Response Rate	Neighborhood committee	80%	32
	Village	90%	28

2.2 Terminal Sampling Frame

Due to high population mobility and separation of residents from registered households in China, we believe that the sampling based on the rosters of urban neighborhoods and villages would cause a large number of omissions. In order for us to get a complete sampling frame that covers all the residents so as to improve the accuracy of third-stage sampling, maps of local villages were drawn with paper and pen in the field. Several pilot testings were done in four villages in Beijing and Hebei from the beginning of 2009 to August, 2009. With detailed information on the features and coding system of buildings, the method of household list making, current map and the usability of households list, a preliminary plan for village sampling frame

[16] Technical report: CFPS-1.
[17] The division of main districts and urban fringe refers to the urban/rural code by the Department of Statistical Design & Management, National Bureau of Statistics of China.

was designed. Pre-tests were done in 4 urban neighborhoods/villages in Gansu and 4 in Zhejiang in November and December of 2009 to examine the practicability of the design. After that, the experience was summarized and improvements were made to the mapping and street listing design.

From December, 2009 to April, 2010, 23 group training exercises for mapping and street listing were conducted. Each group was trained for 3 days and there were 243 draftsmen trained in total. The training included the drawing and coding of the buildings, the collection of auxiliary materials, and the making of building lists and housing lists.

The mapping officially started in December, 2009 and finished in June, 2010, with the maps, statistical tables of residents' information, and households lists of 649 urban neighborhood/villages completed. Various methods were applied to audit the maps and other materials in order to ensure quality and credibility. For detailed map making methods and verification standards, see *Third-stage Sampling Frame Construction (CFPS-2)*.

After organizing these map materials and solving some specific problems such as multiple residences in one household, multiple households in one residence and some failures to confirm addresses,[18] the group started the third-stage sampling.

[18] See Technical Report CFPS-1 for more details.

3. Questionnaire Design[19]

3.1 Overview

Five major questionnaires were designed in the CFPS: the community questionnaire, the family roster questionnaire, the family questionnaire, the child questionnaire and the adult questionnaire. Surveys were conducted at three levels. At the community level, CFPS did an overall interview of the sampled villages/urban communities using the community questionnaire, mainly focusing on the infrastructure, population structure, policy implementation, economy, and social service, etc. At the family level, one member of each eligible household filled out two questionnaires, one on the family members' basic information and members' relationships, and the other on the basic information of the whole family. At the individual level, eligible individuals were surveyed, with children under 16 answering child questionnaires and family members older than 16 answering adult questionnaires. The child questionnaire was divided into two parts: proxy questionnaires answered by the child's guardian for children aged between 0 to 15, and a self report for those aged 10 to 15.

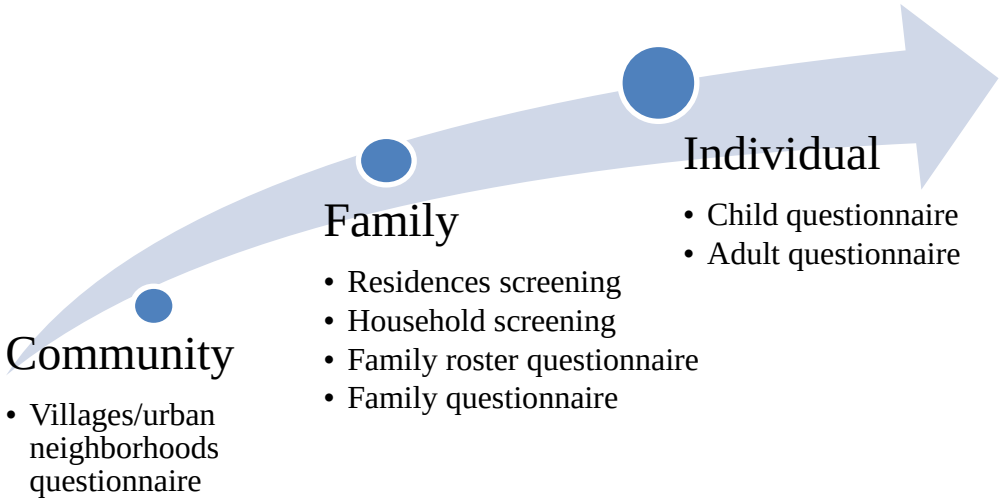


Figure 5. The Levels of the Major Questionnaires in CFPS

In addition to the five major questionnaires, the CFPS 2010 baseline survey designed household screening and household member screening questionnaires. The residence screening questionnaires were collected through the interviewers' observations and judgments in the field surveys; schematized questionnaires were not

[19] For more detailed information on questionnaire design, please refer to Sun et al. (2011).

needed. The household screening, similar to traditional questionnaires, was completed by interviewing the household members. See Figure 6 for the flow chart of all the questionnaires of the baseline survey 2010.

The family roster questionnaire lays a foundation for the individual questionnaire of the family members. As stated above, all baseline family members and their newly born and adopted children in subsequent waves are considered the gene members of CFPS; Non-gene immediate relatives (parents, spouse, children) of the gene members living in the household are defined as the core member of the household; and family members who are neither gene members nor core members are defined as non-core members. When the family roster questionnaire is completed, the system selects either an adult or a child questionnaire for every family member based on the person's age; for gene members and core members who are not present (i.e., a non-coresident family member), a proxy questionnaire is completed for each member, and the individual sample would be re-assigned to an interviewer who is in charge of the region that the non-coresident member is in. If no interviewers are available for that region, a telephone survey would be carried out.[20] Four types of non-coresident gene members and core members will not be interviewed as exceptions: those who are monks or nuns, those who are in prison, those who are serving in the army, and those who have moved abroad. However, they will be interviewed in subsequent waves if they move back to their original households. The non-core family members defined by the family roster questionnaire would not be administered individual questionnaires. Starting from the CFPS 2012 follow-up survey, CFPS began its interviews based on families and individuals defined from the previous waves, without doing any further household or household member screening. For newly-split families formed by the gene members, the system allocates each new unit a separate family roster questionnaire and constructs the family structure of the newly split family using the gene member as the starting point. For the deceased, CFPS collects data on exit questions from surviving family members and then stops tracking the member, but keeps the individual ID. See Figure 7 for the flow chart of the questionnaires of the follow-up survey.

CFPS applied multi-module designs for questionnaires. Each questionnaire was composed of different modules according to the specific situations of the families and individuals interviewed. The CAPI system made the setting up of personalized questionnaires in interviewing more convenient. For example, the school module was used for those still at school and the work module was used for those employed. This was also why we did not use two different questionnaires in rural and urban areas.

[20] In the CFPS 2010 field survey, we only conducted a local survey, that is, we finished the interviews with the households and family members in the sampled village/community as well as individuals who were not at the sampled household but still in the same district/county. For those family members who went outside the local district/county, we did not conduct individual interviews. However, we collected their information using proxy questions in the family roster questionnaire.

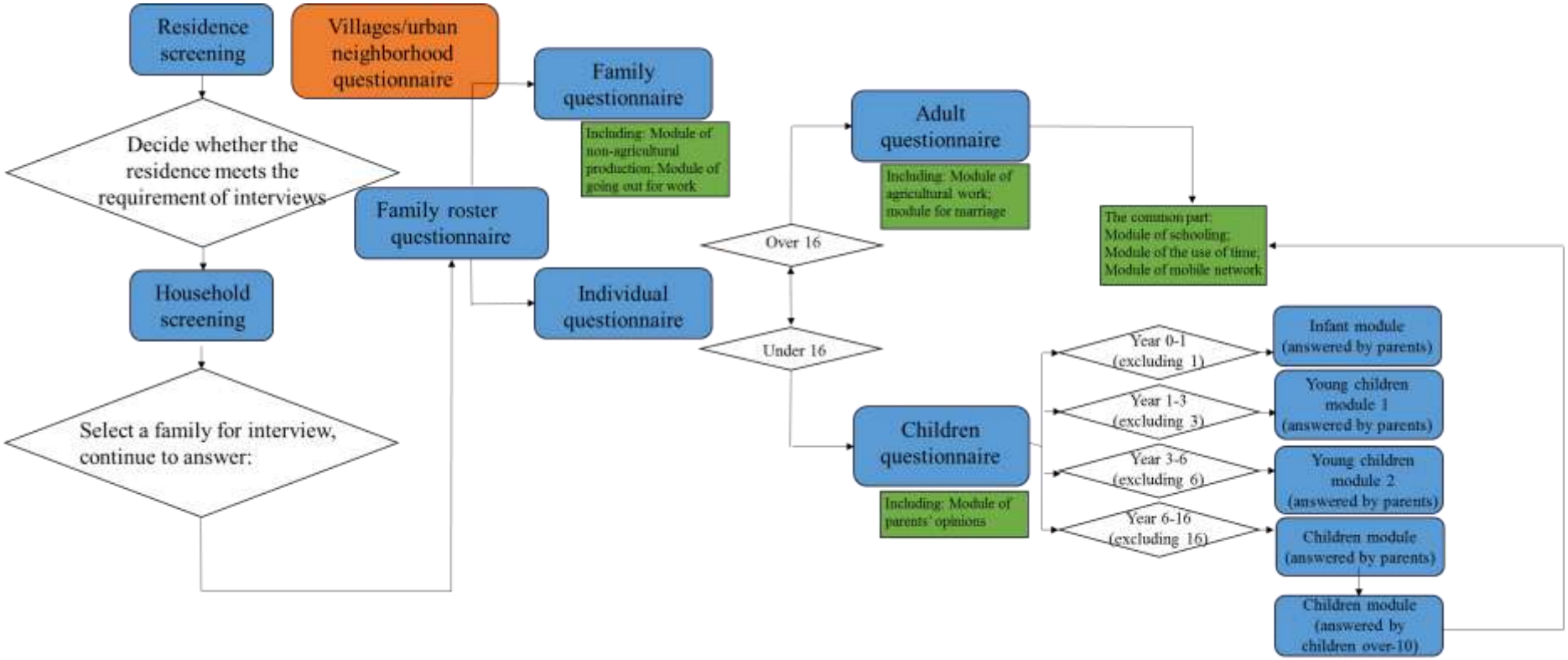


Figure 6. The Structure of Questionnaires of CFPS Baseline Survey 2010

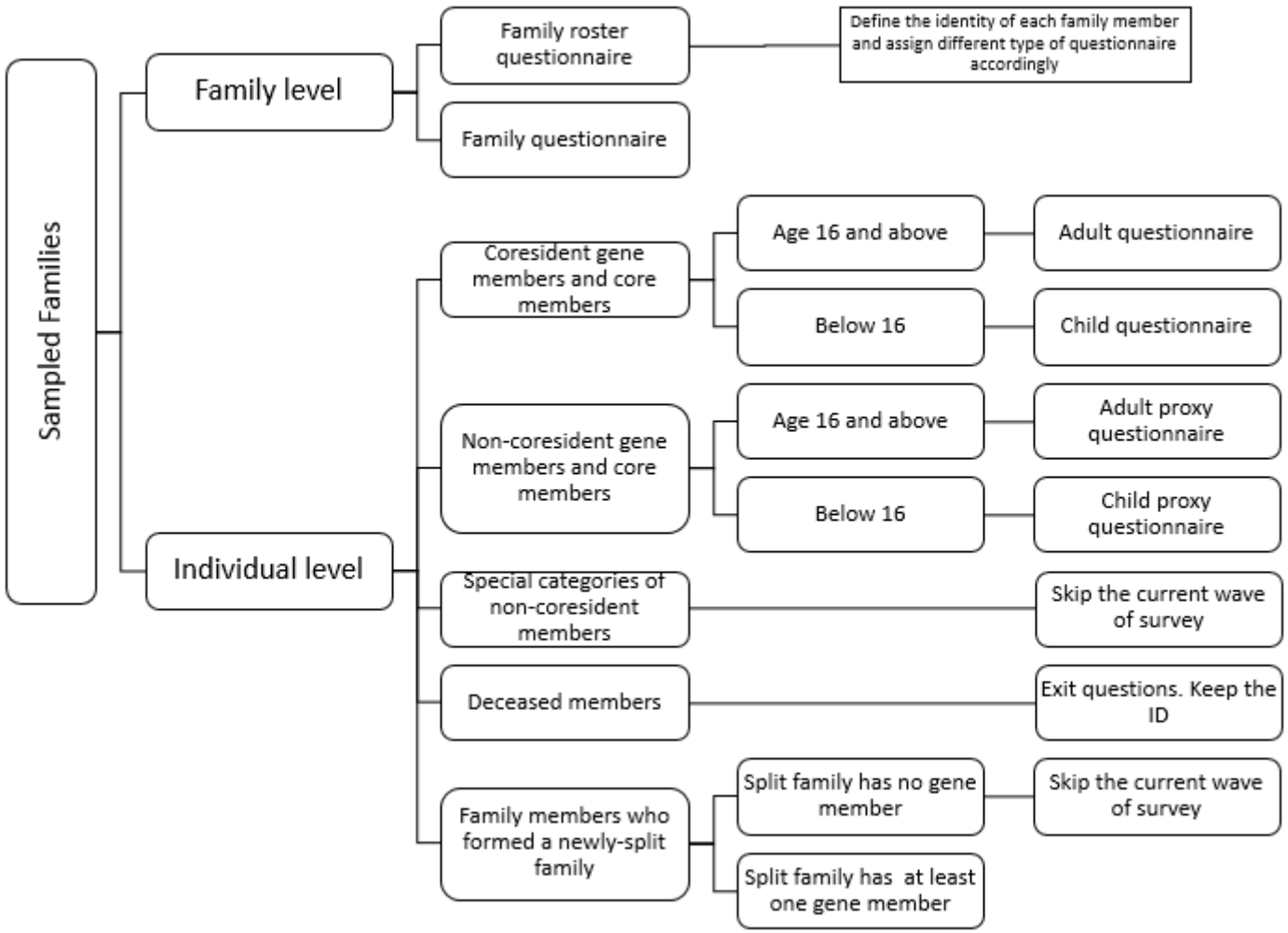


Figure 7. Flow chart of the questionnaires of the follow-up survey

3.2 Community questionnaire

The goal of the village/urban community questionnaire is to acquire information on the infrastructure, population, politics, economy, history, policies, and related information on the villages (i.e., rural communities) and urban neighborhoods (i.e., urban communities). For the process and content of the CFPS 2010 baseline survey, see Figure 8 and the second column in Table 4.

Figure 8. Flow Chart of the Community Questionnaire of CFPS 2010

Table 4. Main Content of Community questionnaire of CFPS 2010 and CFPS 2014

Module	Questionnaire Content
A. Infrastructure	type of village/urban neighborhood, respondent's position, facilities, bulletin, land borders, administrative area, water sources, fuels
B. Population Structure	total number of households, population, registered residents, permanent residents, migrant population, age structure, fertility and mortality, ethnic groups
C. Economy	basic living allowances, price level
D. Administration	type of administrative staff, working conditions, neighborhood transportation
E. History & Politics	historical changes, whether it is a tourist area, whether there are high-pollution enterprises, latest election of village/urban neighborhood's' committee
F. Housing Price	highest price in history, highest last month, average price last month
G. Environment & Transportation	transportation hours to the closest county, town, provincial capital, mineral resources, natural disasters, land resources
H. Labor Economy	labor force structure, gross output of agriculture & non-agriculture sectors, net personal income per capita, price of assisting workers, distribution of the "big family names"
J. Health Care & Fertility	area of medical station, number of medical staff, progress of rural cooperative medical services, one-child policy
K. Finance	Collectively-owned enterprises and their output, total fiscal

	revenue and its resources, total financial expenditure and the items
Y. Respondent's information	gender, age, political status, level of education of the respondent, information of the director of the village/urban neighborhood, names of professions of other respondents
Z. Interviewer' Observations	economic status, tidiness of roads, mental state of villagers/neighborhood, the homogeneity of villagers/neighborhood, architecture pattern, congestion, type of village/ neighborhood, topography, features of respondents

For the interviews using the village/urban community questionnaire, we asked interviewers to find those who know the communities well and have access to statistical data. The staff members of the village/neighborhood committee, especially the director in charge of daily administration, would be the best information source. Other staff members. such as the accountants in the village/neighborhoods, who are also familiar with the community because of their job or their long years of service, can also be potential respondents. In addition, the secretaries of the Party branches in the village/neighborhood can also be respondents if they have a comprehensive knowledge of the village/neighborhood. If the first type of respondents cannot answer all the questions, other respondents can be asked to fill out the questionnaire as well. But these are individual interviews in separate rooms rather than collective interviews.

"The end of the year" in the community questionnaire refers to the last day of a calendar year. For example, the end of the year of 2009 refers to December 31st, 2009.

CFPS 2012 did not adminster community questionnaires, but incorporated some of the questions (i.e. price levels) into the family questionnaires. CFPS 2014 conducted a follow-up survey on the original 649 communities. The basic framework of the community questionnaire in 2014 was largely consistent with that in the baseline survey. Statistics were collected on December 31st, 2013 from the sampled communities, and changes in the communities from January 1st, 2010 to December 31st, 2013 were recorded. See the third column in Table 4 for the main contents of the CFPS 2014 community questionnaire.

3.3. Residence Screening

Before the field interview, our draftsmen had already eliminated the vacant houses and the non-family households by field interviewing, consulting the neighbors and staff members of the village/neighborhood's committee during the construction of the ultimate sampling frame. In order to ensure the accuracy of the final stage of sampling, interviewers were still required to find the corresponding

samples selected, according to the map addresses and confirm that they matched with the actual residents before starting interviews. The confirmation of houses and residences is called the residence screening.

The interviewers check whether the sampled addresses are valid according to our map—that is, whether the addresses provided actually exist. After verifying the information, they decide on the type of the building on the address by consulting the residents or otherswho might know. Qualified residence samples then go through the screening. Invalid addresses, non-residences, and vacant houses were eliminated from our interviews. Multiple attempts may be required to gain accurate information for samples with addresses that are difficult to confirm.

3.4 Household Screening Questionnaire

After confirming the residence on the sampled addresses, the survey goes through the household screening process. The main purpose of the household screening questionnaire is to pick out those households that meet the requirements of interviews within certain residences. Below are the procedures of the screening questionnaire:

First, the number of independent economic units within the sampled residence is determined. For example, for parents and children living together, if they compose one economic community, we define them as one independent economic unit; if they are two economic communities, we define them as two units. The definition of the economic units is not limited to the house owners, but also includes those who have at least a part of the right of habitation of the sampled residences.

Second, the eligible family households are determined among the independent economic units. Two conditions below cause disqualification from our survey:[21]

1. Independent economic units composed of only one person who at the same time belongs to an economic community with more than two other family members elsewhere are not regarded as "households" in our survey. Instead, the individual will be seen as one of the family members of that household to which he or she belongs elsewhere.

2. We require that all households belong to mainland China, which means there must be at least one family member of Chinese nationality (excluding Hong Kong, Macao and Taiwan).

Finally, if there more than one household with one residence address meets the requirements set forth, the computer system randomly chooses one of the

[21] The original design of the screening of households required that at least one of the family members has lived in the sampled residence for 6 months. But this was abandoned in the executing process, for it could eliminate very few households.

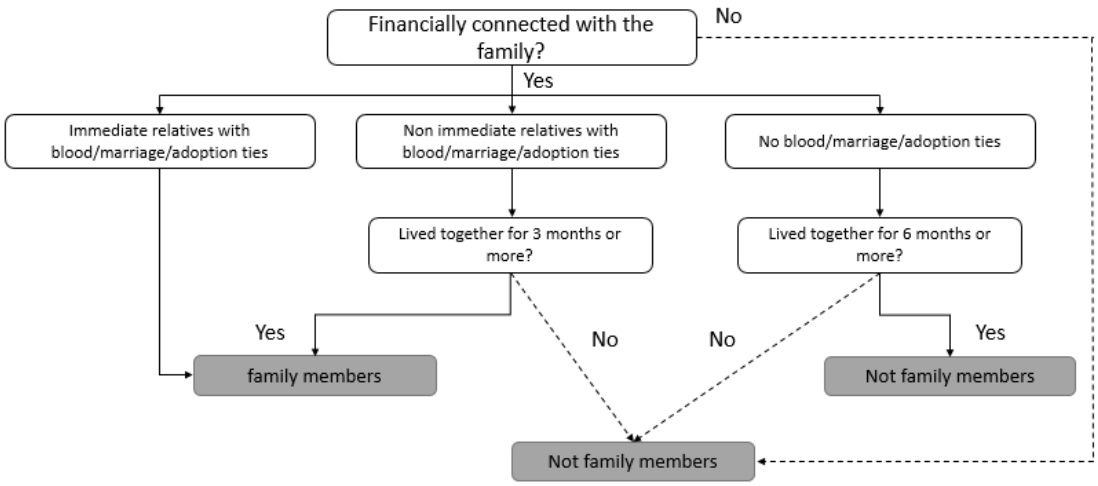
households to be interviewed. For all the households on the address, the questionnaire for the screening of households includes questions on the houses they own outside the local village/neighborhood and elsewhere in China.

3.5 Family roster questionnaire

3.5.1 The identification of family members at baseline

CFPS baseline survey places all members "living together"[22] in the same household into two categories.The first category includes all immediate relatives, and the second category includes non-immediate relatives who have been living in the family for at least 3 months before the interview. We define the above two categories as CFPS gene members, and their newly born or adopted children at subsequent waves are also considered CFPS gene members. Gene members are tracked in follow-up surveys. In addition to the family members aforementioned at baseline, the CFPS baseline survey identified non-family members who had been living in the family for at least six months, and some basic demographic information about these individuals is also collected. These individuals have no blood/marital/adoptive relationship with the family and thus are not our main focus, so the individual questionnaire is not administered. Once they leave the sampled families, we no longer include them in our follow-up survey.

The identification of family members is shown in the following flow chart (Figure 9).



[22] "Living together" means individuals are financially connected, including family members who have blood relationship/kinship, and other members who work in the household but do not have blood relationship/kinship, for example, domestic helper, diver, or distant relative working as domestic helper.

Figure 9. Flow Chart of the Identification of Family Members[23]

In the follow-up survey, considering the changing structure of the family, we adjusted the family roster questionnaire. The complex structure of Chinese families makes it more difficult to conduct the survey. We modified the family roster questionnaire in order to keep track of the changes in different families as thoroughly as possible, while keeping the survey manageable in practice. In subsequent surveys, we added a number of questions to help identify family members and their relationships to one another.

3.5.2 Content of the Family Member Questionnaire

The respondents of the family roster questionnaire must meet two requirements: they must be among the members who are living together and they must have blood/marital/adoptive relationships with other family members.

The family roster questionnaire is designed to collect information on the relations among different members of the family. It also collects important social demographic information on all family members and non-family member co-residents, including gender, age, educational level, occupation, hukou, and residence. Moreover, using different sources of information increases the completeness of social demographic information in CFPS.

3.5.3 The Design and Implementation of the T Tables

The design of the T tables of the CFPS family roster questionnaire is a CFPS innovation that aims to collect information on family relations. In past social surveys, information on family relations was collected through questionnaires at the individual level, which directly asked the respondent to provide information about their parents, spouses, and children. Such an approach has several drawbacks.

First, these surveys normally select only one respondent as the center of family relations and ask questions on his or her specific relations with relatives. It is thus presumed that there is only one center of family relations, i.e., the respondent himself/herself. Normally, a random assignment is given within the household selection. An alternative method is to select the household head. However, since family relationships have the structure of a tree network with multiple centers, every family member can be treated as the family center, so the selection of the center might not be meaningful. Previous methods build up a radial structure from a single

[23] This figure has been modified on the basis of Sun et al. (2011, p.126).

core (e.g., the house owner or the respondent), which is only a small part of the family tree network.

Second, since the traditional methods only collect radial structures based on singular cores which treat the respondents as the centers, researchers are restricted to learning about the relations between the respondents and other family members/relatives, whereas the relationships among family members apart from the respondents are unknown.

Third, previous surveys only ask questions about the "father," "mother," and "children" of the respondent in general, without asking their names or assigning codes to these relatives. Thus, even if more than one family member is interviewed, it is impossible to relate them in one family network through their names or codes.

Fourth, the information collected from previous surveys is mainly from the same generation (e.g., siblings and spouses) or two successive generations (e.g., parents and children). Without the information on the relations between relatives and their names, the information on the relations across generations is unknown.[24]

The T tables of the CFPS 2010 baseline survey solve the abovementioned problems. The T tables consist of three tables—T1, T2, and T3—which appear at the beginning of the family roster questionnaire (see Figure 10). Table T1 (on family members living together) and Table T3 (on immediate relatives not living together) record the basic social and demographic features of every family member and his or her immediate relatives (parents, children and spouses) who are not living with them. Table T2 (on relations) identifies the relations of all the family members and the corresponding relations between T1 and T3 members.

The Table on Family Members Living Together (T1)

Personal code	Name	DOB	Sex	Marital status	Highest degree	Main job	Admin./Managerial position	Info.on people outside home
101								
102								
…								
301								
302								
…								

The Table on the Relations of Immediate Relatives (T2)

[24] See Technical Report: CFPS-7.

Personal code	Name	Father	Mother	Spouse	Child 1	Child 2	…	Child 9	Child 10
101									
102									
…									

The Table on Linear Family Members Not Living Together (T3)

Personal Code	Name	DOB	Sex	Alive or not	Marital status	Highest degree	Main job	Admin./ Managerial position	Info. on residence and *hukou*
301									
302									
…									

Figure 10. The Design of the T Tables

All the information of the T tables is completed by the respondent of the family roster questionnaire. It does not require that every family member be involved in filling out the tables. In the field interviews, the interviewers complete the T tables using CAPI.

First, in the introduction to the family roster questionnaire above, information on the family members living together is already identified by a set of questions on the basis of economic relations. Those who are qualified as family members living together in this process will then fill in Table T1. The final T1 table includes the core family members with individual codes starting with the number "1" and the non-core family members with codes starting with "3."

Next, according to the information of the core family members in Table T1, the original name list of Table T2 is automatically generated. Table T2 is completed by "taking-turns," which means that everyone takes a turn to confirm his or her relationships with other immediate relatives (parents, children and spouses) as the center of family relations, which completes the table on the relations of immediate relatives (Table T2).

Finally, we collect the basic information on all the immediate relatives of the family members included in the name list on Table T2. Concerning the immediate relatives mentioned in T2, if they are already in Table T1 (i.e., they are family members living together), the CAPI system will automatically add their information and there will be no repetition of interviews. If the family members do not exist in T1, the system will generate a name list for Table T3. Again, the interviews will be done sequentially until the final table on immediate family members living together (Table T3) is completed. The individual codes of all the members of T3 will start with the number "2."

These three tables jointly present a thorough family network, where the relations within the same generation, immediate generations, and gapped generations are all connected, with detailed records of each member's personal information. The design of the T tables fixes the disadvantages of previous surveys in the inaccurate information on family relations, providing richer information and fuller pictures of family relations. It also provides valid information on all family members for in-depth studies on family issues. Meanwhile, the implementation of the T tables avoids repetition of interviews, which greatly increases the efficiency of the survey. A complete and comprehensive family relationship defined in the baseline survey also laid a good foundation for the CFPS follow-up surveys.

Despite the multiple advantages of the design of the T tables and maximal considerations of convenience and efficiency issues, the family roster questionnaire takes more time to complete and is more complicated in comparison with regular surveys due to the large amount of information collected. The behaviors and interview skills of the interviewers have direct effects on the cooperation of the respondents and the quality of the data in the T tables. Therefore, strict and comprehensive training is needed for the interviewers before they can complete field surveys.

3.5.4 The Family

The target samples at the individual level in the CFPS will remain constant over the years. Once a person is identified as a CFPS gene member (family members at baseline and their newly born and adopted children in subsequent waves), he/she becomes a permanent respondent of CFPS and will be tracked in follow-up interviews. However, the family in which the individual lives and its members may change overtime. First, the family structure may change for reasons such as deaths or births of family members. Second, the family may change as a unit. Some families dissolve when the last family member passes away, and new families emerge due to divorces or marriages. Information on family structure needs to be updated at each wave, as it is an important variable for individuals. Meanwhile, family structure itself can be a valuable research topic.

Although CFPS aims to follow up with individuals in subsequent waves, we start with the family roster questionnaire based on the information from the most recent wave in practice. This is done so as to keep track of the changes in family structure. To be specific, CFPS adopts the "addition and subtraction" method to collect information about family changes from the last wave. The first step is subtraction. For all family members defined in the most recent survey, CFPS collects information on whether each of them is still living with the other

members.[25] For those who are no longer living together, CFPS then separates them into two categories: those who need to be interviewed and those who do not. The latter group includes family members who are practically inaccessible (e.g., died, ran away from home), and those who moved away from family to live in institutions (e.g., become a monk or a nun, be in prison, serve in the army, move abroad or move to nursing homes). For the family members who need to be interviewed, we define their status based on whether they are financially connected with the current family: if yes, they are defined as members of a newly-formed family; otherwise, they are defined as non-coresident family members of the existing family. After excluding those who do not need to be interviewed and previous members who are considered to be in a newly-formed family, we have the "list of existing family members" of the household at hand.

The second step is addition. In this step, we focus on the new family members since the last survey. The definition of a family member is generally the same as that in the baseline survey, which includes: (1) gene member living at home and their non-gene member parents, children and spouse; (2) other immediate relatives of the gene member who have been living in the household for over 3 months and who are not economically independent of the household. We add the new family members to the "list of existing family members" to get the "list of current family members" of the household.

The third step is adjustment. The two steps stated above capture most of the changes in the family structure. However, the information collected includes changes in the family at only two points, namely the family structure at the time of the last and current surveys, and omits the family members who joined and left the family in between. For example, a child who was born to the family or adopted by the family after the last survey but left before the current survey should be considered a CFPS gene member, but he/she would not be captured by the two steps above. Although such cases are rare, we will make further adjustments for completeness. In this step, we collect information on any newborn child of all the gene members in the household who was born after the last survey but left the family before the current survey. For these non-coresident new gene members, we apply the same rule to define their family membership based on their economic independence.

If there are no gene members in the household after applying the three steps above, we will stop interviewing the household. For the newly-formed families, the starting point of our "addition and subtraction" method is the gene member who

[25] For family members who are temporarily not at home (short-term absence) at the time of the interview, we ask the respondent to define if he/she lives in the household based on whether he/she will return within 3 months and live in the household long-term.

split from the original family. Then we construct the family structure based on the same procedure mentioned above.

3.5.5 Economic Independence

In the family roster questionnaire, we need to define the status for two types of individuals: those who recently joined the family and those who left the family. For either type, financial independence is a critical criterion. However, in these two cases, there are different criteria for economic independence. For the newcomers, the criteria for economic independence in all the follow-up surveys are consistent with those in the baseline survey; that is, whether the new family member "lives together" with other family members in the household and whether the new member is financially connected with other members of the household. However, things are more complicated when determining whether the leaving family member is still financially connected with the original family or has started a new family, as the leaving family member may still maintain a financial connection with the original family in the form of financial or non-financial support, even though he/she no longer physically lives in the household. We have explored and tried several ways to figure out an objective criterion for defining the family membership for members who are not physically living in the original household.

In the 2011 sample maintenance survey, we tried to define the economic connection between the non-coresident and the original family using a unified objective criterion. Specifically, if the financial transfer went beyond 1000 Yuan, then the family member was defined as economically dependent, and otherwise as economically independent. In practice, we came to realize that the amount of financial transfer depended heavily on the income of the family and the local socioeconomic development level, and therefore this was not a good measure of economic dependence.

In the 2012 follow-up survey, we abandoned the reliance on objective amounts of money. Instead, we considered the individual's reasons for leaving the household. Some reasons strongly indicate a newly-split family, for example, getting married or divorced. We considered members leaving for such reasons as belonging to a newly-split family. Other reasons are less clear, for example, leaving for study or for work. We took into consideration whether an individual has a spouse or child at home, whether he or she is employed, and whether the person supports or is supported by the family. Theoretically, this is an effective way of differentiating family membership, but such a design is not perfect due to the complicated nature of family. As this method defines the status of each member one by one, it sometimes splits the family members living in the same unit. For example, we defined those who are employed, married, have no spouse or child in the original family, and neither support nor are supported by the original family, as members of a newly-split family.

However, in the case of a couple leaving their extended family and working in another city, if only the husband supports his original family, but the wife as a daughter-in-law does not directly support the family, our approach of identifying the individual's family status was to define the husband as a member of the original family and the wife as a member of a newly-split family, which goes against common sense and is very inconvenient for practical purposes.

In the 2014 follow-up survey, we made another attempt to deal with the problems mentioned above. First, we switched our basic unit from an individual to an address- based unit. We considered all the members living at the same address as a unit to see if the unit as a whole was economically independent from the original family. That solved the problem of splitting a family. In the 2012 follow-up survey, the family membership of the non-coresident was determined by the family respondent of the original household. However, in the 2014 survey, we introduced a dual-judgment method. That is, in addition to asking the family respondent of the original household about the status of a non-coresdient member when completing the survey at the original household, we also asked the non-coresident member himself/herself about his/her economic relation with the original family when interviewing the new address unit. Such a design helped to better create the family structure as well as to define the scope of family in later data collection about family economic conditions.

3.6 The Family Questionnaire

The main purpose of the family questionnaire is to collect information on daily life, social interactions, and economic activities of the sampled families. The main content of the CFPS 2010 baseline family questionnaire is shown in Table 5. The follow-up surveys adopted this as their main framework while also making adjustments.

Table 5. Main Content of the Family Questionnaire of CFPS 2010

Module	Content
A. Geography and transportation	Nearest public transportation, medical stations, high schools, commercial centers
B. Living conditions	Use of water, fuel resources, electricity, bathroom conditions, garbage disposal, employment of house maids
C. Social interactions	Spring Festival visits, gift giving, family-lineage, ancestor worship, neighborhood interactions, communications with relatives
D. Housing	House ownership, self-built houses or commercial apartments, houses for rent, house area, living time, market value and rent of the houses, the apartment structures, other estates, difficulty in housing
E. Management conditions[26]	U module for working outside (people who are outside, working address, time devoted, whether they go home during vacations, the transfer payment, whether their family has increased/decreased assisting workers because they are working outside), the government's support, reason for poverty, V module for non-agricultural management (type of non-agricultural industry, number of participants, total assets, shares held by family members, number of employees, turnover, after-tax profits), houses for rent, land and other means of production for rent, property for sell, demolition of the houses, land acquisition
F. Family income	Savings, financial products, pensions/social security/basic livingallowances, salaries/rewards/allowances/bonuses, non-salary/agricultural income, value of gifts
G. Family assets	Insurance indemnity, others' debts, value of collections, present value of other assets
H. Family expenditure	Highest expenditure, loans, daily expenditure (food, travel, communications, etc.), special expenditure (family appliances, medical care, education, commercial insurances), donations, total expenditure
J. Durable goods	Cars, motorcycles, tractors, televisions
K. Agricultural production	Land type, land areas, revenue and expenditure, types of crops, output, sales, income, types of domestic animals and fishing, output, sales, income, raising conditions of domestic animals
Z. Interviewer Observations	Respondent's housing situation, tidiness of the family, mental state of their family members, relations of family members, relations between elderly and the young, relations between genders, personal characteristics of the respondent

[26] The E module consists of U (working outside) and V (non-agricultural management). If there are family members going out for work or who work in non-agricultural industries, they shift to answering sub-module U or V. After finishing the questions in these two parts, they come back to the E module and continue the following questions.

The family questionnaire is best answered by the person who has a comprehensive, detailed knowledge of family life and the family's financial situation. It can be completed by one respondent or multiple respondents. For example, questions on agricultural activities may be answered by a family member who is in charge of finances related to agricultural activities, questions on a family business may be answered by a family member who is in charge of the business, and questions on family expenditure may be answered by a family member who is in charge of food purchases. Although the family questionnaire has been consistent across waves, we made several adjustments to the measurement in practice, most notably in the income and expenditure section. Tables 6 and 7 show the itemized design of family income and expenditure by wave. In general, CFPS 2010 adopted a more general scheme, while CFPS 2012 attempted a detailed itemization design to increase the completeness of data collection. CFPS 2014 combined some overly detailed items and optimized several questions on family income and expenditure.

In the family income section, CFPS 2012 made the following adjustments to the 2010 questionnaire: (1) We added some income items that were omitted in 2010. For example, in the 2010 survey, we did not ask about the market value of the agricultural products consumed by the family, revenue of self-employed business, wage income from agricultural employment, stipends/scholarships and income from internships or part-time jobs for students. CFPS 2012 added specific questions on these items. (2) We further refined the main income categories in 2010. Take income from agricultural production as an example. In CFPS 2010, we asked about the gross income and total cost of agricultural production, including all farming, forestry, pasturing, fishing, and sideline produce. In 2012, we instead asked about the income from selling agricultural products (including the crops cultivated, forestry products, poultry, livestock, fishery products, and other sideline products), market value of the agricultural products eaten or used by the family, and the cost of every production process. Another example is public transfer income. In the 2010 survey, we asked about total income from pension/social security/minimum living allowances (*Dibao*) in one question. In the 2012 survey, we first asked the respondent to list the items of public transfer income that the family had received, and then asked the amount of income item by item. In this way, the respondent would better understand and recall the relevant income to minimize the possiblity of omission. (3) For many types of income, we added unfolding brackets questions. When the respondent was unable or unwilling to give the specific amount of an important type of income, we asked the respondent to select the range of income. This design made the questions less sensitive and thus lowered the missing rate. The CFPS 2010 survey used the unfolding method only for gross wage income in the family income, while in the 2012 survey, this method was used in gross agricultural income, income from self-employed business, private enterprises, and individual wage income. We took the average of the upper and lower limits as an approximation of the otherwise missing item. (4) We adjusted the question on wage income. In the 2010 survey, we asked the family respondent to report the wage

income of each family member in turn and then give a total number or range of the total wage income of the family. In the 2012 survey, we instead asked about the wage income in the individual questionnaire. Therefore, the total wage income is the sum of wage incomes from all relevant individual questionnaires.

Although the design of the income section in the 2012 survey helped us to collect more information, it lengthened the interview and made the questions more difficult to answer. As a result, in the 2014 survey we combined some overly detailed items and adjusted the time range of some income indicators. For the data cleaning process and information regarding variables, please refer to section 7.4.

Table 6. Design of the family income section in CFPS

2010	2012	2014
Ⅰ.Business income 1. Agricultural income[27] – Net income from farming, forestry, pasturing, fishing and sideline production ■ Net income from farming and forestry products ■ Net income from livestock and fishery businesses 2. Net profit from each private enterprise	Ⅰ.Business income 1. Agricultural income (net income from selling agricultural products plus the value of self-consumed agricultural products) – Farming and forestry products – Livestock and fishing products 2. Net profit from each private enterprise	Ⅰ.Business income 1. Agricultural income (net income from selling agricultural products plus the value of self-consumed agricultural products) – Total value of agricultural products and sideline products 2. Total net profit from private enterprises
II. Wage income 1. Wage income (including wages, bonuses, subsidies and dividends) 2. Money sent or brought back by family member(s)	II. Wage income Individual questionnaire 1. Income from employment in agriculture related work (farming or other jobs) 2. Income from all employment in non-agricultural work	II. Wage income 1. Income from working for other farmers, including agricultural work and other tasks 2. Money sent or brought back by family

[27] CFPS 2010 did not collect information on the value of agricultural products consumed by the family. We estimated the value in data cleaning. Please refer to Section 7.4 for details.

who work(s) away from hometown[28]	post-taxation wage income, and non-cash welfare 3. Income from internships / part-time jobs / work-study programs while receiving formal education	member(s) who work(s) awayfrom hometown[29] 3. Total wage income from non-agricultural employed work (including wages, subsidies, bonuses, and non-cash welfare)
III. Public transfer income 1. Family income from pension/social security/minimum living allowance *(Dibao)* 2. Total income from government public transfers (cash and goods)	III. Public transfer income 1. Pension/retirement subsidies for retired workers (from individual questionnaire) 2. Government subsidies – Minimum living allowance *(Dibao)* – Reforestation subsidies – Agricultural subsidies (including direct grain subsidies and farming machinery subsidies) – Wubaohu subsidies (targeted at low-income, blind, disabled, elderly, and youth who cannot support themselves) – Tekunhu subsidies (targeted at extremely poor family) Work injury subsidies to immediate relatives – Emergency or disaster relief (including material goods)	III. Public transfer income 1. Total pension (retirement subsidies) in the family 2. Total government subsidies (cash and goods)

[28] This was not included in the calculation of total family income in 2010 data cleaning.
[29] This was not included in the calculation of total family income in 2014 data cleaning.

	– Other government subsidies 3. Donations or compensation Donations (cash and non-cash) – Financial compensation for land expropriation - Compensation for housing demolition/relocation 4. Scholarship and educational subsidies received as students	3. Donations or compensation – Social donations (cash and goods) – Financial compensation for land expropriation - Compensation for housing demolition/relocation (including cash and housing) –
IV. Asset income 1. Total income from renting out house(s)	IV. Asset income 1. Income from renting out house(s) – Income from renting out the house in which the family currently lives – Income from renting out other housing properties	IV. Asset income 1. Total income from renting out house(s)
2. Total income from renting out land or other means of production	2. Income from renting out land – From renting out the collectively distributed land – From subletting the land	2. Income from renting out land – From renting out the collectively distributed land – From subletting the land
3. Total income from renting out other goods	3. Renting out other family assets (such as equipment)	3. Renting out other family assets (such as equipment)
4. Income from selling assets (household goods)		4. Investment income[30]
V. Other income 1. Other income from cash/non-cash gifts	V. Other income 1. Private financial support or donation – Financial support or donations from non-coresident relatives	V. Other income 1. Private financial support or donation – Financial support or donations from non-coresident relatives

[30] Not included in the family income in data cleaning.

	– Financial support or donations from other people	(cash and non-cash) – Financial support from other people (cash and non-cash) – Gifts and cash due to banquets, ceremonies and social relations

Design of the expenditure section in the follow-up survey has undergone a similar process to that of the income section. CFPS 2010 used tables to summarize various kinds of family expenditure, including daily expenditure (converted from per month to annual) and special expenditure (annual). As the items in 2010 were overly aggregated, CFPS 2012 kept the main categories and broke them into more detailed items. It also listed the details to remind the respondents; for example, when asking about daily used commodities and necessities, we listed detergent, soap, toothpaste, toothbrush, etc. The CFPS 2014 questionnaire on family expenditure was basically consistent with that in CFPS 2012, but optimized some questions by combining some overly detailed items, adjusting the time range of several items, optimizing the range setting in soft checks, and adding supplementary questions on family expenditure in significant events that were omitted in the 2012 survey. In this way, the respondent could better recall the information and give more accurate answers.

Table 7. The design of family expenditure in CFPS

2010	2012		2014
I. Production and business costs 1. Total costs of farming, forestry, pasturing, fishing and sideline production	I. Production and business costs 1. Costs of farming and forestry production		I. Production and business costs 1. Costs of farming forestry production
	– Seeds, fertilizer and pesticides	– Seeds, fertilizer and pesticides	
	– Labor	– Labor	
	– Rental of machines and irrigation	– Rental of machines – Irrigation	
	– Other costs	– Other costs	
	2. Cost of Poultry, livestock and fishery production	2. Cost of Poultry, livestock and fishery production	
	– Fish or breeding stock	– Fish or breeding stock	

	– Labor – Forage – Other cost	– Labor – Rental of machines – Forage – Other cost
II. Food expenditure (last month) – Food expenditure	II. Food expenditure (last week) – Eating out (including treats) – Cigarettes and alcohol consumed by own family – Other foods self-consumed by the family – Value of agricultural foods self-consumed by the family	II. Food expenditure (monthly average in the last 12 months) – Total food consumption (including snacks, beverages, cigarettes, and alcohol self-consumed by the family) – Eating out
III. Living expenditure (last month) – Communication fee – Transportation costs (including vehicle maintenance) – Daily used commodities – House rent – Hiring domestic helper or hourly worker	III. Living expenditure (last month) – Postage and communication (including telephone, mobile phone, internet and postal costs) – Water and electricity – Fuel – Local transportation (including petroline) – Daily used commodities – House rent – Hiring domestic helper or hourly worker –	III. Living expenditure (monthly average in the last 12 months) – Postal and communication (including telephone, mobile phone, internet and postal costs) – Water – Electricity – Fuel – Local transportation (including public transportation and petroline) – Daily used commodities – House rent

	− Recreation and entertainment	
	− Lottery	
− Financial support to family members − Housing mortgage		
− Car loans		
− Other mortgages		
IV. Long-term living expenditure (last year)	IV. Long-term living expenditure (last year)	IV. Long-term living expenditure (last 12 months)
− Clothing	− Clothing and accessory	− Clothing and accessory
− Recreation and entertainment		− Recreation
	− Travel	− Travel
− Living expenditure (heating, property management fee)	− Concentrated heating	− Concentrated heating
	− Property management fees (including parking lot rentals)	− Property management fees (including parking and cleaning)
− Housing purchase/construc-tion (excluding mortgage)		− Housing mortgage
		− Housing maintenance and decoration
	− Automobiles purchase	− Purchase, maintenance and repair of automobiles
	− Purchase and repair of other communication, transportation tools and related accessories	− Purchase and repair of other communication, transportation tools and related accessories
	−	−
− Household appliance	− Appliances for work	− Purchase and maintenance of furniture, appliances and other durable goods
− Other household goods and service	− Furniture and other durable goods	
− Education	− Education	− Education
− Medical care	− Medical care	− Medical care
	− Fitness	− Nutritional supplements

	− Cosmetic services	− Cosmetic services and haircuts
− Commercial insurance purchase	− Commercial medical insurance purchase − Commercial asset insurance purchase	− Commercial insurance purchase
	− Financial support and donations to non-co-residing relatives	− Financial support and donations to non-co-residing relatives
	− Financial support and donation to other people	− Financial support and donation to other people
− Social donations in cash and in kind	− Social donations in cash and in kind	
	−	− Social donations in cash and in kind
	− Tax and fees	
	− Land rent	− Land rent
	− Rents of other household assets (e.g. equipment)	
− Other expenditure	− Other expenditure	− Other expenditure
V Important events − Marriages and deaths of family members − Gifts to relatives and friends		V Important events − Banquets and ceremonial spending − Gifts to relatives and friends
Confirm total cost of family expenditure of last year		Confirm total cost of family expenditure in the past 12 months

3.7 The Individual Questionnaire

3.7.1 Design principles

In CFPS, those younger than 16 are defined as children and those over 16 are defined as adults. CFPS has a child questionnaire and an adult questionnaire for the two groups respectively.

As the age of 16 is the dividing point for children and adults and the questionnaires for children have different modules for children of different ages, the calculation method for age is very important. The system only takes into account the

year of birth, and ignores the month. Specifically, the age equals the year of the survey minus the year of birth.[31] For instance, if a child was born in October 2000, and the survey was being performed in July 2010, we would regard the child's age as 10 in the CFPS, even if he or she was not yet ten years old. This rule is also applied in the other types of CFPS questionnaires and follow-up surveys.

Regarding the respondent of the individual questionnaires, adult self reports must be completed by the adults themselves.[32] The child questionnaire is divided into two sections: the child section completed by the children themselves and the parent section completed by the children's guardians. For children of all ages the parent section is to be completed by the target child's main guardian, who can be the child's parent or the primary caregiver who knows the child best. In addition, children between 10 and 16 (excluding 16 years old) need to complete the self report.

Starting from 2012, CFPS added individual level proxy questionnaires. The proxy questionnaire is used under two circumstances: (1) For non-coresident family members, we ask a resident family member to complete a proxy questionnaire. A self-report questionnaire will later be attempted with the non-coresident family member. (2) For those who are physically unable to participate in the survey (for reasons such as a language or mental disability), we ask the family member who knows him/her best to complete the proxy questionnaire. The proxy questionnaire is a simplified version of the self report. We fully consider the proxy nature of the questionnaire and delete the subjective attitude section, which cannot be answered by others. The addition of the proxy questionnaires helps us to collect as much key information as possible about the gene and core members.

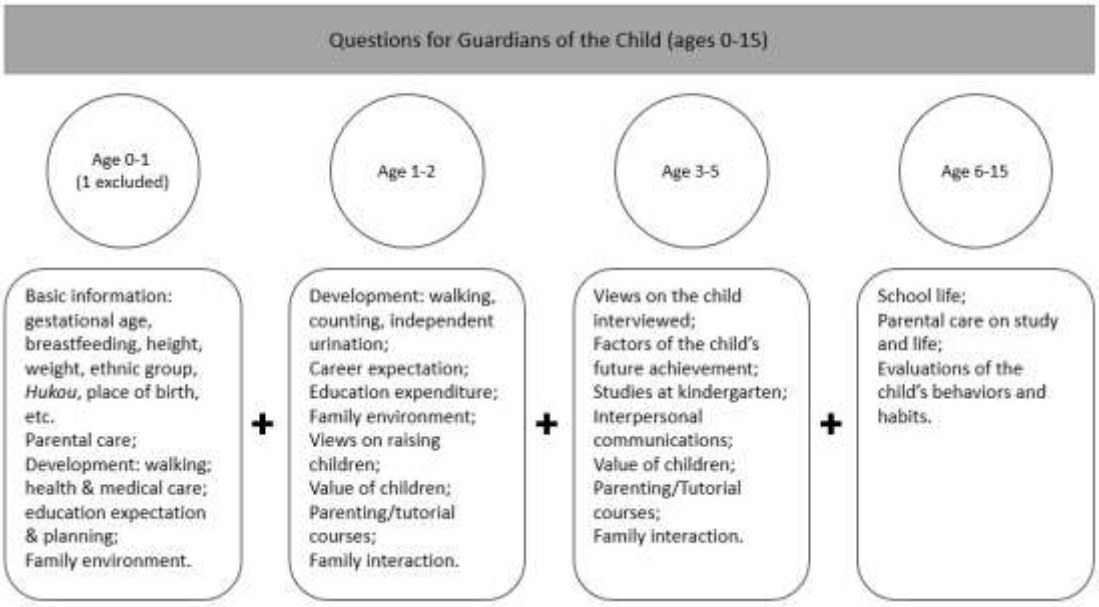
3.7.2 Content of the questionnaires

In the 2010 baseline survey, we designed different modules in the child questionnaire for children of different ages (i.e., those under 1 year old, those between 1 and 3 excluding 3, those between 3 and 6 excluding 6, and those between 6 and 16 excluding 16). The upper part of Figure 11 shows the structural process and content of the interviews. As is shown in the figure, the number of questions increases with the age of the child. Children between 0 and 1 have the fewest questions and children between 6 and 16 have the most. The latter group had to answer all the questions for children under 6 years old, plus questions on schooling and parental care. Meanwhile, we asked the parents to make some assessments about the child's

[31] As the individual questionnaire is generated based on the family roster questionnaire, the survey year used in age calculation is the year when the family roster questionnaire was finished, which could be different from the year when the individual questionnaire was finished.
[32] Notice that the proxy part in the child self-answer questionnaire is different from the proxy questionnaire for children mentioned below.

behaviors. In the 2012 follow-up survey, we combined these modules and used logical skips to screen out the questions that were unsuitable to certain ages or had been answered in a previous survey. In this way, we were able to keep the questionnaire concise and reduce redundancy.

For the detailed content of the child section in the child questionnaire in the baseline survey, see the lower part of Figure 11. In subsequent surveys, the general content remains the same.



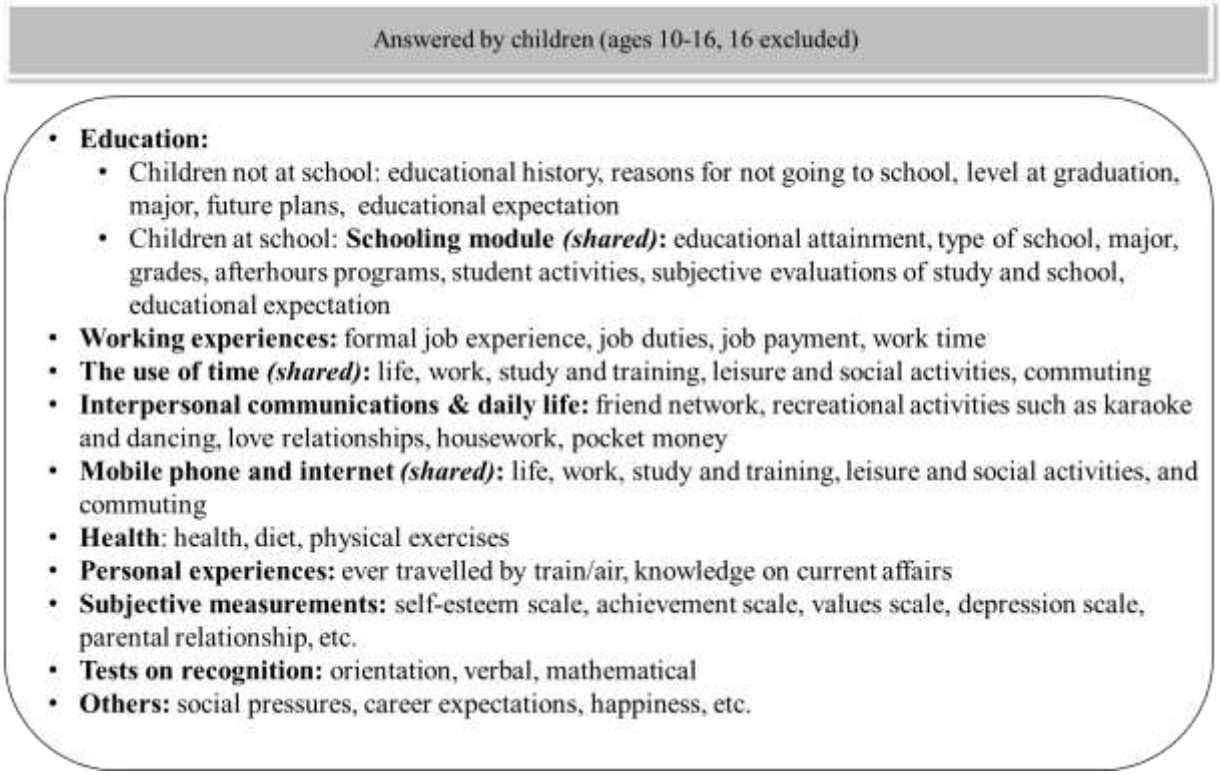


Figure 11. Structural Process and Content of Interviews Using the Children's Questionnaire of CFPS 2010[33]

The general content of the adult self report questionnaire has been stable. Users may refer to the second column in Table 8 for detailed contents in the baseline survey. However, in order to adapt to the data collection needs in the follow-up surveys, CFPS distinguishes four categories of questions used in the follow-up: baseline questions, core questions, rotation questions, and extension questions. Baseline questions are for first-time respondents. They are all objective recall questions and need to be answered only once. The core questions are repeatedly asked across different waves to measure the changes in particular variables over time. Rotation questions are rotated in different follow-up surveys or with different respondents based on certain rules. Questions are rotated either by wave, or within waves by respondent characteristics such as age group. The extension questions include questions from added modules which often reflect topical issues of the society. Such questions usually appear only in one wave. Users may refer to the last column in Table 8 for the main contents in the adult questionnaire. Moreover, we added several questions to the 2012 and 2014 follow-up surveys. Table 9 lists these additional questions and their types.

[33] This figure has been modified on the basis of Sun et al. (2011, p.151).

Table 8. Main Content of the Adult Questionnaire of CFPS 2010

Module	Content	Question type
Basic information	Date of birth, birth weight, birth place, residence, *hukou*, ethnicity, family category during the Cultural Revolution, political party and organizational affiliations, time spent living with parents before age 3 and between ages 4 and 12	Baseline questions
Siblings	Number of siblings, name, date of birth, alive or not, age at and cause of death, marital status, educational attainment, occupation, administrative/managerial positions, residence, living with parents, parents' cause of death	(used only in CFPS 2012)
Educational history	Educational attainment, type of schools at different educational levels, time spent at school, when did school end, name of schools, graduated or not, subject and major, educational expectation	Baseline questions
Language use	Importance of different languages, language spoken at home	Core questions
Schooling *(shared)*	Current level of education, type of school, major, grades, extracurricular tutoring, student activities, subjective evaluations of study and school, educational expectation	Core questions
Marriage	Current marital status (married/remarried/cohabitating/divorced/widowed), date of birth of the present/last/first spouse/cohabitation partner, time at marriage/cohabitation, pre-marital cohabitation, channel of initial contact, reason for the divorce ending the last/first marriage	Core questions
Relations with children	Evaluation of the relations with children by those aged 60 or older, intergenerational transfers	Core questions
Work	See Figure 13	Core questions
Personal income	Non-operating income, operating income, financial support from relatives and friends, government subsidies	Core questions
Time use *(shared module)*	Life, work, study and training, leisure and social activities, commuting	Core questions
Leisure	Leisure activities and frequencies, means of travel, overseas experiences	Core questions
Mobile phone and internet *(shared module)*	Use of mobile phones, QQ, MSN, e-mails, importance of internet, frequency and places of internet use	Core questions
Social relations	Help seeking, confiding troubles, self-reported social status	Core questions
Subjective measurements	Values, social attitudes, achievement scale, life satisfaction	Core questions

Politics	Experiences of thefts or robbery, unjust treatments, media interests, evaluation of government work	Core questions
Health	Height, weight, self-reported health, discomfort, chronic diseases, experiences of hospitalization, medical expenses, coping with diseases, satisfaction with health care, traditional Chinese medicine, physical exercises, diet, P-ADL, smoking and drinking, sleeping, memory, depression scale, main health care-giver	Core questions
Mental health	K6 Scale, CESD Scale	Rotation questions
Cognitive assessments	Literacy/vocabulary, math, word recall, number series	Rotation questions
Personal information and observations of the interviewers	Contact information, respondent, personal characteristics of respondent	Core questions

Table 9. Added questions in CFPS adult questionnaire in follow-up survey

Module	Content	Question type
Added in CFPS2012		
Information about deceased siblings	Education level and occupation of siblings who died before the baseline survey	(Only in CFPS 2012)
Pension insurance	Status of participation, fees and benefits of various kinds of pension insurances	Extension questions
Fertility intentions	Ideal number of children	Core questions
Trust	Trust towards different types of people	Core questions
Religion	Religious beliefs and participation in religious activities	Core questions
Anchoring vignettes for health assessment and social status	Assess the health condition and social status of hypothetical cases	Core questions
Added in CFPS 2014		
Parent information	Parents' birth year, parents' occupation and political affiliation when the respondent was age 14	Rotation questions
Law module	Law module	Extension questions
[EHC-RESI]	EHC Migration module	Core

		questions
[EHC-Marriage]	EHC Marriage module	Core questions
[EHC-Job]	EHC Employment module	Core questions
Family decision	Who has the final say about family issues	Core questions
Marriage satisfaction	Satisfaction level towards marriage/cohabitation and spouse/partner	Core questions
Political vote	Participation in political votes	Core questions
Social security	Attitudes towards local security and judicial fairness	Core questions
Reading	Reading in the last 12 months	Core questions
Traditional attitudes	Attitudes towards parent-child relationships and gender roles	Core questions

3.7.3 Improved Measurements

Although the main contents in the CFPS individual questionnaire basically remained stable in the follow-up surveys, we improved the measurement for some variables in order to meet the needs of data collection. Detailed introduction is provided below.

3.7.3.1 Work module

a. Original design in the 2010 baseline survey.

Figure 12 shows the flow chart and the contents of the work module of the adult questionnaire in CFPS 2010, which lays the foundation for later updates.

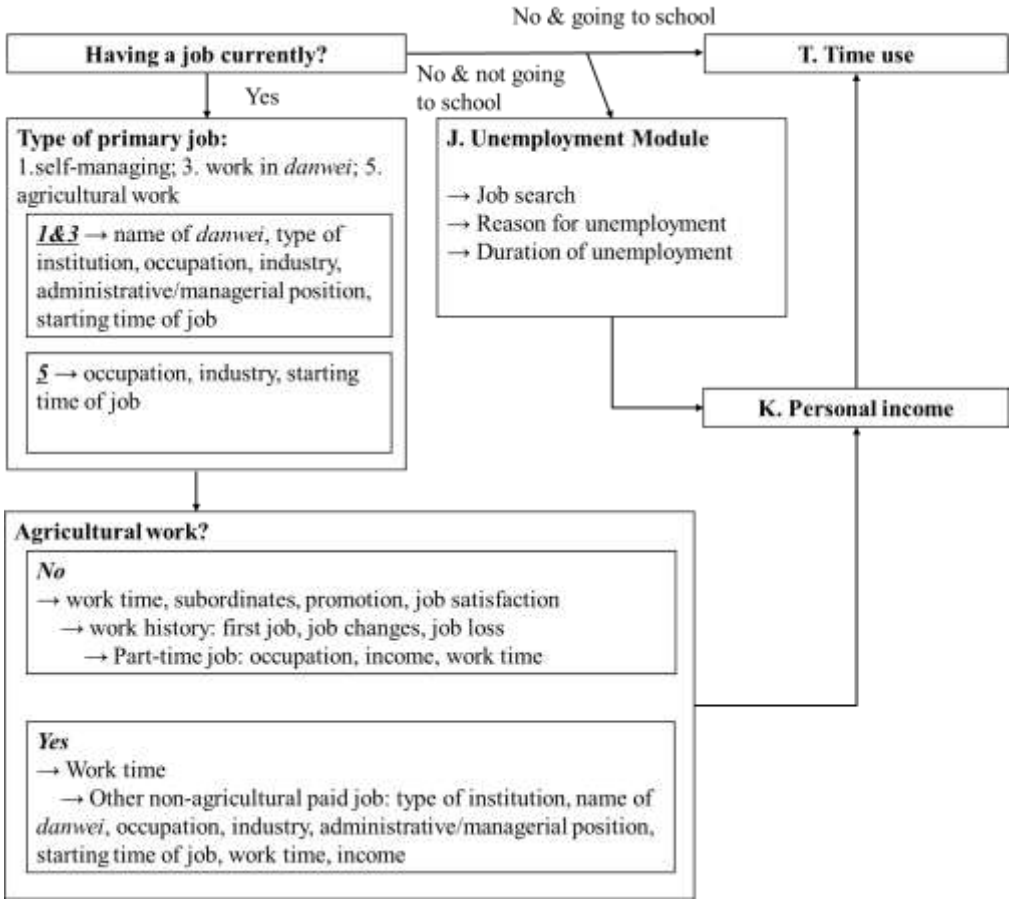


Figure 12. Flow Chart of the Module Structure of the Adult Questionnaire of CFPS 2010

b. Defining employment status

Usually there are two ways of defining employment status. The simple way is to define it subjectively, that is, directly ask the respondent if he/she is currently working, and have the respondent define his/her employment status. However, the more standard way is to measure it objectively by asking a series of questions. CFPS initially used subjective measurement, but switched to objective measurement in 2012. CFPS mainly followed the protocols from the International Labor Organization (ILO) and also took into account the designs in CPS (Current Population Survey), CULS (China Urban Labor Survey), HRS (Health and Retirement Survey), CHARLS (China Health and Retirement Longitudinal Study). Adjustments were also made to adapt to the CFPS population.

CFPS counts agricultural work for family production, agricultural and non-agricultural waged jobs, self-employment and private business as employed, but excludes housework and volunteer work. See Figure 13 for the flow chart and definition of current employment status.

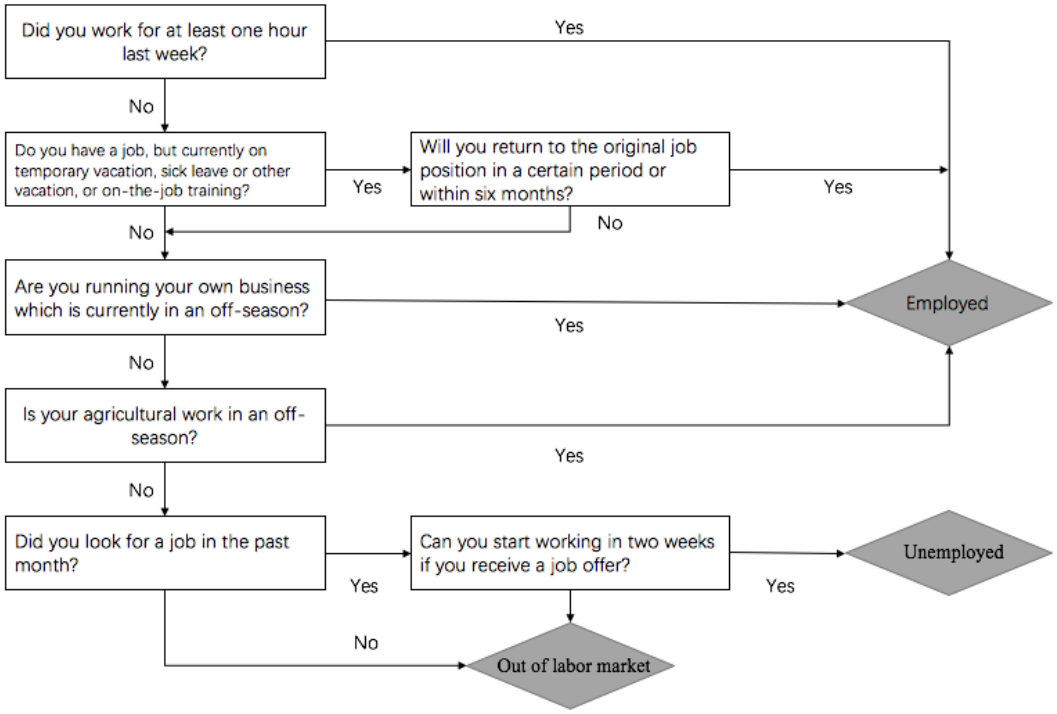


Figure 13. Flow chart of employment status in CFPS

c. Job class

Different questions are applicable depending on the types of jobs (i.e., agricultural, non-agricultural, employed and self-employed). For example, the compensation for agricultural work rarely contains insurance, non-cash benefits, bonuses, and pensions, which are important parts of the income structure of employed work. The income of agricultural production is usually undividable within the family, while wage incomes are for individuals. Therefore, type of job is a key variable in the collection of work information. A wrong classification of job class can cause problems in the interviewing process and a series of unsuitable questions, which would hamper the interview and lower the data quality.

As shown in Figure 12, CFPS 2010 only asked for information on the current main job and for simply classified jobs as agricultural or non-agricultural. Such classification was adequate when the information collected was rough and not customized.

Starting in 2012, CFPS aimed to collect information on all jobs, which requires a better classification of job class. CFPS 2012 used screening and asked the respondent to report if he/she was participating in the following 5 job classes: agricultural jobs for his/her own family, agricultural jobs employed by others, employed non-agricultural jobs, self-employed/private business, and work as a helper. Then for each self-reported job, we would ask a series of more detailed questions relevant to the particular job class. During the interviews, we became

aware of one drawback associated with this approach: a significant number of respondents did not correctly report their job class because they were not clear about the definitions, which resulted in duplications and missing data.

In CFPS 2014, we retained the job categories and their customized questions from 2012, but abandoned the self-reported job class as a screening question. Instead, we directly asked two objective questions for every reported job, after which the system defined the job class based on the answers and assigned further questions. Users may refer to Table 10 for definitions of each type of work.

Table 10. Definition of job class

Work for oneself/own family or employed by others/ other families/ organizations / institutions / companies?	Agricultural or non-agricultural?	
	Agricultural	Non-agricultural
Work for oneself/own family	Class 1: Family agricultural work	Class 2: Individual/ private business/other self-employment
Employed by others/ other families	Class 3: Agricultural work for other families	Class 5: Non-agricultural casual workers
Employed by organizations / institutions / company	Class 4: Agricultural employed	Class 4: Non-agricultural employed

In 2014, agricultural and non-agricultural jobs were defined based on the nature of the jobs and the employers. Such a design solved the problem of lacking a unified objective definition of agricultural and non-agricultural jobs. It is worth noting that the definition of agricultural and non-agricultural jobs was slightly different for employed and self-employed jobs. For employed jobs, the distinction between agricultural and non-agricultural jobs relies on the nature of the employers, for example, the employer of employed agricultural work must be a farmer, but the actual jobs could be participating in agricultural production or providing temporary minor assistance). On the other hand, the respondent who was doing non-agricultural employed work must be employed by a non-farmer individual or organization. Their actual jobs could be agricultural or non-agricultural. For self-employed work, the boundary between agricultural and non-agricultural was less clear. For example, agricultural production activities for one's own family was referred to as "all kinds of agricultural production activities and relevant business activities," which could be agricultural production such as growing apples or business activities such as selling family-produced apples. Another example is that

self-employed/private business could be “shoeshine stands on the streets” or “managing an apple orchard and selling apples by renting land and hiring local farmers.”

d. Extension of the work module

Starting in 2012, CFPS collected information on all jobs between two waves. As it was impractical to collect detailed information on every job due to time constraints, we compromised by prioritizing different jobs. In every wave, we collected detailed information on the main job since the last wave, but only brief information on other jobs. See Figure 14 for the differences.

CFPS modeled after PSID in defining the main job in the following way: (1) If the respondent has only one job at the time of the interview, then it is the main job; otherwise, the respondent determines which one is the main job. (2) If the respondent is currently out of jobs, his/her latest job is considered the main job. If multiple jobs end at the same time, the respondent determines which one is the main job.

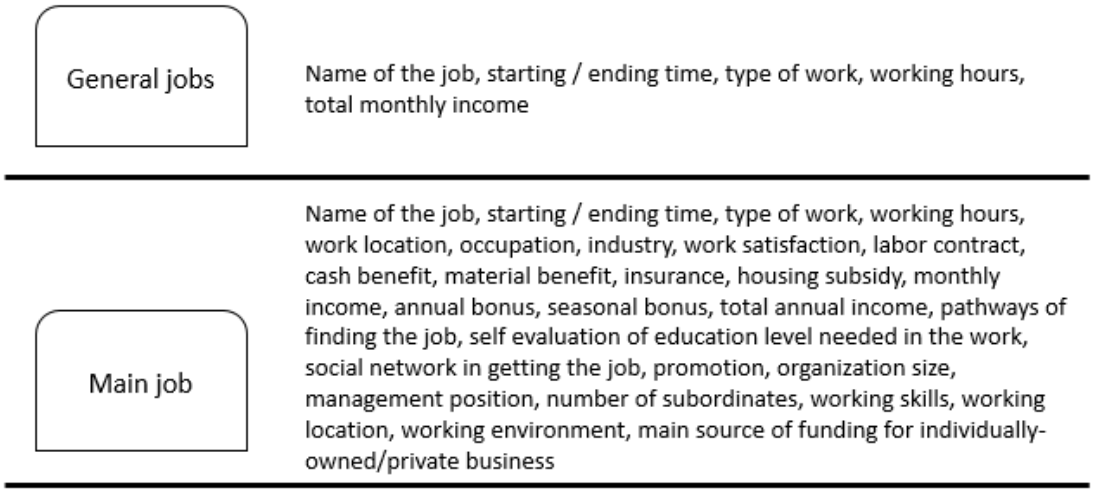


Figure 14. Information collected for main job and other jobs

3.7.3.2 Event History Calendar

Event History Calendar (EHC) is a useful tool for collecting complicated information by helping the respondent recall the timing of the events. As the survey is conducted every two years, we adopt EHC in three modules to increase the accuracy of the data: residence status, marriage, job module. Figure 15 illustrates the

display of EHC screens in CFPS. Before collecting data, EHC generates a time frame for recalling based on the information collected from previous interviews[34] and displays an empty calendar on the screen. In the figure, the recalling time frame for the respondent is from October 2012 to June 2014. During the interview, EHC automatically displays the answers on the screen in a calender format. As the figure shows, the respondent lived at address A from October 2012 to June 2013 and from January 2014 to June 2014. After his divorce from spouse A, he quit his job B and moved to address B. With the help of hints such as marriage, birth, moving and changes in jobs, the respondent was able to recall more information with higher accuracy. Based on this principle of memory, on one hand, EHC clearly presents the timeline to the user and helps the respondent to locate the timing of some events by referring to the timing of other events; on the other hand, EHC helps to increase the completeness and accuracy of the recall by helping the researcher to construct a timeline of events for the respondent. More detailed information about EHC could be found in the forthcoming *Technical Report: CFPS-35.*

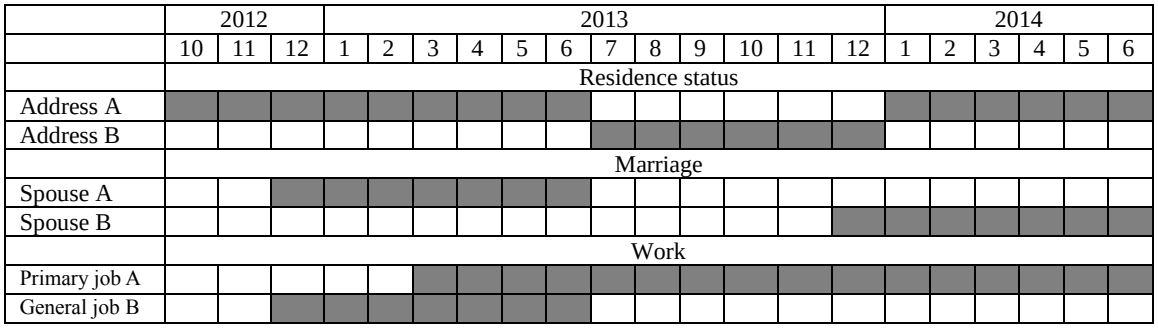


Figure 15. Illustration of Event History Calendar method in CFPS

3.7.3.3 Cognitive Assessments

There are four types of cognitive assessments in CFPS: the literacy test, the math test, word recall and the numerical series. Each type assesses a different aspect of cognition and the four types complement one another. CFPS divides the four assessments into two sets, which are alternated across waves. Set A contains the literacy test and the math test, which measure educational achievement. Set B includes word recall and numerical series tests, which reflect the fluid intelligence of the respondents. Set A was used in 2010 and 2014 and Set B in 2012 and 2016.

Both the literacy test and the math test have multiple equivalent forms. When a respondent takes the test for the first time, a form is randomly chosen for the respondents, and at subsequent waves, the computer loads a different form from the

[34] We set the time frame for recalling for follow-up respondent to be from the last survey month to this survey month, for first-time respondent to be from 1st January in the last survey year to this survey month.

one used in the last administration. Meanwhile, the system tries to assign different forms to members in the same household to minimize spillover effect. See Section 7.3 for data cleaning and a variable summary of cognitive tests.

3.7.4 Psychological Scales

The CFPS individual questionnaire not only collects abundant information on social demographics, behaviors and cognitive function, but also evaluates psychological conditions using psychological scales. The nationally-representative data on individual psychological factors collected by the CFPS psychological scale provide valuable information for relevant research. The main contents of the CFPS psychological scale include individual characteristics, parent-child relationships and subjective attitudes. We use validated psychological scales whenever possible, constructed both domestically and internationally, in order to ensure the reliability and validity of the responses. Meanwhile, we also adjusted some scales to adapt them to the Chinese context. CFPS used nearly 20 psychological scales in the first three waves. They were the Positive Behavioral Scale (PBS), Self-Discipline Scale, Nowicki-Strickland Locus of Control Scale for Children, Rosenberg Self-Esteem Scale (RSES), Kessler 6 Rating Scale (K6), Center for Epidemiologic Studies Depression (CES-D), Responsibility Scale, Parental Bonding Instrument (PBI), The Home Observation for Measurement of the Environment Inventory (HOME), The Value of Children, Parental Attitude Scale, Relationship with Parents Scale, Factors for Achievement Scale, Trust Scale, Inequality Scale, Family Value Scale, Job Satisfaction Scale, Importance Scale, and Marriage Satisfaction Scale. More detailed information about CFPS 2010 K6, CFPS 2012 CES-Dand RSES scale can be found in the Technical Reports "Composite Variable (2): Education and Depression (CFPS-12)" and "Psychological Scale CFPS-26." In addition, Chapter 14 in *China Report 2016* contains detailed information on the CFPS psychological scale. Table 11 shows the years of data collection for different scales.

Table 11. Psychological scales in CFPS 2010 –CFPS 2014

Name	Note	Questionnaire	Target respondent		
			2010	2012	2014
Individual characteristics					
Positive Behavior Scale (PBS) WE301-WE312	Measure positive behavior, 5-point-scale	Child questionnaire: Adult proxy report	Age 3、7、11、15	Age 3、7、11、15	Missing previous response or newly entered age group 3-15
Self-control Scale WM701-WM712	Measure self-control ability, 5-point-scale	Child questionnaire: child's self-report	×	Age 10-15	Missing previous response or newly entered age group 10-15
Nowicki-strickland Locus of Control Scale for Children (NLCS-C) QM4011-QM4019 QM40110-QM40111	Measure the locus of control, 5-point-scale	Child questionnaire: child's self-report; Adult questionnaire	Age 13、15	×	Missing previous response or newly entered age group 10-21
Rosenberge Self Esteem Scale (RSES) QM1011-QM1019 QM10110-QM10113	Measure self-esteem, 5-point scale	Child questionnaire: child's self-report; Adult questionnaire	Age 10	Age 10、12、14	Missing previous response or newly entered age group 10-21
K6 by Ron Kessler (K6) QQ601-QQ606	Measure mental health, 5-point scale	Child questionnaire: child's self-report; Adult questionnaire	Age 10 and above	×	Age 10 and above

Center for Epidemiological Survey, Depression Scale（CES-D） QQ6011-QQ6019 QQ60110-QQ60120	Measure mental health	Child questionnaire: child's self-report; Adult questionnaire	×	Age 10 and above	×
Responsibility Scale WF801-WF807	Measure responsibility, 5-point scale	Child questionnaire: Adult proxy report and child's self-report; Adult questionnaire	Proxy report: age 6—15 Self-report: age 10 and above and at school	Proxy report: at school, or not at school but age 4 and above Self-report: age 10 and above and at school	Proxy report: at school, or not at school but age 4 and above Self-report: age10 and above and at school
Parent-child relationship					
Parental Bonding Instrument (PBI) WM201-WM214	Measure perceived parental bonding, 5-point scale	Child questionnaire: child's self-report	Age 11	11,13,15	Missing previous response or newly entered age group 10-15
HOME Scale WG301-WG306 WG308	Measure the incentives and supports from family, 5-point scale	Child questionnaire: Adult proxy report	Age 1-5	Age 1-5	Age 1-5
The value of children to parents WE201-WE209	Measure childbearing motivation, 5-point scale	Child questionnaire: Adult proxy report	Age 2, 6, 10, 14	Age 2, 6, 10, 14	Missing previous response or newly entered age group 10-15
Parenting Attitude Scale WE101-WE108	Measure parenting attitudes, 5-point scale	Child questionnaire:	Age 1, 5, 9, 13	Age 1, 5, 9, 13	Missing previous

		Adult proxy report			response or newly entered age group 0-15
Parent-child Relationship Scale QM1001-QM1006	Measure subjective assessment of parent-child relationship, 5-point scale	Adult questionnaire	×	×	Age 16 and above
Subjective attitude					
Factors of Sense of Achievement Scale QM3011-QM3017	Measure the subjective importance of each factor affecting sense of achievement, 11-point scale	Child questionnaire: Adult proxy report and child's self-report; Adult questionnaire	Proxy report: Age 4, 8, 12 Self-report: Age 12, 14	Proxy report: Age 0, 4, 8, 12 Self-report: Age 10-15	Missing previous response or newly entered age group 21 and below
Inter-personal Trust Scale QN10021-QN10026	Measure inter-personal trust, 11-point scale	Child questionnaire: child's self-report; Adult questionnaire	×	Age 11, 13, 15; Age 16 and above；	Missing previous response or newly entered age group 10-15； Above age 16
Inequality Scale WV101-WV108	Measure perceived inequality in the society, 5-point scale	Child questionnaire: child's self-report; Adult questionnaire	Age 16 and above	Age 10, 12, 14	Missing previous response or newly entered age group 13-15
Family gender-role Scale QM1101-QM1104	Measure subjective attitudes towards gender	Adult questionnaire	×	×	Age 16 and above

	roles in family, 5-point scale				
Job Satisfaction Inventory QG501-QG506	Measure job satisfaction, 5-point scale	Adult questionnaire	Age 16 and above	×	×
Perceived Importance Scale QM501-QM510	Measure subjective importance of money, relationship, family etc. ,5-point scale	Adult questionnaire	Age 16 and above	×	×
Marital Satisfaction Inventory QM801-QM803	Measure satisfaction in marriage and cohabitation, 5-point scale	Adult questionnaire	×	×	Age 16 and above

4. Field operation

4.1 Pilot studies

A pilot survey is an important step preceding a baseline survey. Before the CFPS national baseline survey launched in 2010, we conducted two pilot studies in 2008 and 2009. Between May and September in 2008, a pilot survey was initiated in Beijing, Shanghai and Guangdong. This pen and pencil pilot survey focused on society, economy, education and healthcare. The pilot survey in 2008 consisted of 2,400 households, with 800 households in each province/municipality located in 8 counties/districts. Four villages/neighborhood communities were included in each county/district, and 25 households were included in each village/neighborhood community. CFPS adopted PPS sampling and interviewed 2,375 households, 7,214 individuals in 24 counties/districts, and 95 villages/neighborhood communities.[35]

Between May and September in 2009, the pilot follow-up survey was conducted of the sampled households in 2008. However, two types of samples were not included: households who had moved out of the village/neighborhood community, and respondents who had left their households. In the official survey, these two groups were both tracked. The sample size of the pilot study in 2009 was 1,995 households.

Unlike in 2008, Computer Assisted Personal Interviewing (CAPI) technology was introduced in 2009. During the study, we conducted comprehensive tests of CAPI technology, including the stability and reliability of its real-time interview management technologies, real-time technological support, and real-time data quality control.

These two pilot studies laid a solid foundation for the baseline survey in 2010. In addition, data collected in the two surveys from these three regions are now available for further research.

4.2 CFPS 2010 baseline interviewers

For better control of the implementation costs and time, we employed local interviewers mainly from the sampled neighborhood communities in the 2010 baseline survey. Each interviewer was in charge of 2 neighborhood communities, and in big cities, 0.2-0.5 additional interviewers were required.

Interviewers were mainly recruited online. After the resume screening, phone interviews, and face-to-face interviews, a total of 453 interviewers were selected.

[35] The Institute of Social Science Survey (ISSS), Peking University (2009).

They were divided into 14 groups to receive 6-day training at Peking University. The training courses were taught in small classes from February 22nd to August 13th in 2010, including tutorials, group practices, simulation surveys, and field interviews in real-life settings, etc. Finally, 438 interviewers passed the training and exams and became official interviewers.

Table 12 below shows the basic information of interviewers in 2010. More detailed information on the interviewer recruiting process and training for the 2010 baseline survey can be found in *The Implementation Report* (CFPS-3).

Table 12. Basic Characteristics of the Interviewers of the 2010 Baseline Survey[36]
(Total Number of Interviewers: 438)

Characteristics	Category	N	%
Gender	Male	294	67.1
	Female	144	32.9
Marriage	Unmarried	265	60.5
	Married	173	39.5
Age (year)	18-19	10	2.3
	20-29	306	69.9
	30-39	101	23.1
	40+	21	4.8
Educational attainment	Graduate	11	2.5
	4-year college	198	45.2
	Junior college	153	34.9
	High school and below	76	17.4
Occupation	Corporate employee	137	31.3
	Student	109	24.9
	Public professional services	50	11.4
	Unemployed	44	10.0
	Family planning agencies	39	8.9
	Teacher	29	6.8
	Self-employed	30	6.6

4.3 Overview of 2010 Survey Implementation

The 2010 baseline survey covered a wide range of 25 provinces/municipalities/autonomous regions, including 162 counties/districts[37] and 649 communities.[38] The implementation was divided into two parts: the extensive

[36] Technical Report: CFPS-3.
[37] The 32 streets/towns in Shanghai included 18 districts/counties.
[38] Combined villages/neighborhoods were calculated separately.

survey during the survey season and some supplemental ones based on the results of the former survey.

The survey season started in April and ended in September, 2010. During the season, we completed interviews in 600 communities,[39] including 14,852 household screening questionnaires, 14,326 family member questionnaires, 14,192 family questionnaires, 32,202 adult questionnaires, and 8,789 child questionnaires.[40] In these 600 communities, 4,224 cases failed to fill out the family members' questionnaires.[41] The main reasons were as follows:[42]

(1) Inaccuracy of the sampling frame: 1,690
(2) Refusal: 1,490
(3) Failure to contact after 6 attempts: 461
(4) Not eligible and screened out: 374

The supplemental interviews mainly involved the following situations. One situation was where there were communities in which the interviews were not completed or did not reach the expectation of 25 households. There were 324 communities in which this was the case, and 118 interviewers were called to complete this part of the supplemental interviews. Among these communities, those where interviews were not done included both the communities whose members refused to be interviewed and those where the time period was not congenial for interviews. For example, World Expo 2010 was held during the survey season in 2010, and therefore we postponed some interviews in Shanghai in order to ensure safety in these communities.

The second part of the implementation involved samples with interviewers' misconduct. We needed to do the re-sampling and start the interviews all over again for these cases. In 2010, we discovered 5 serious interviewer cheating cases in total. Detailed information on these cases and their resolutions can be found in *Implementation Report* (CFPS-3). In order to ensure the consistency and validity of the statistics, Computer-Assisted Personal Interviewing was applied in the supplemental interviews. The content of the questionnaires, as well as the interviewing system, was exactly the same as the first interviews in the survey season.

After the supplemental interviews, altogether we completed 635 community questionnaires, 15,717 household screening questionnaires, 14,960 family roster questionnaires, 14,798 family questionnaires, 33,600 adult questionnaires, and 8,990 child questionnaires in 2010. The sample maintenance survey in 2011 and the follow-

[39] Interviews are regarded as completed after the interviewers confirm that there were 6 contact attempts and 3 refusals.
[40] Technical report: CFPS-3.
[41] Technical report: CFPS-3.
[42] Technical report: CFPS-3.

up survey in 2012 confirmed that there were no wrong addresses and respondents, no substitute interviewers, or data fabrication cases in the 2010 baseline survey.

4.4 Refusal and solutions in the 2010 baseline survey

Refusal led to substantial sample losses. There were more than 1,000 households who refused to be interviewed in the 2010 CFPS baseline survey.[43] The analysis found that the response rate mainly depended on the type of community. In general, in communities where the majority of officials or community committee staff resided, upscale communities, elderly communities, and military communities, the respondents were more resistant to interviews and the refusal rate was higher. Moreover, as the village/neighborhood committees were involved in contacting village/neighborhood residents for our interviews, whether the committee staffs were willing to cooperate and the state of the relationship between the committees and communities also had a direct influence on the respondents' responses to the interviews. If the committee made less effort to cooperate with our survey or in their management of the community, or if it did not have actual administrative power over the community, the refusal rate was also higher.

Several solutions have been proposed to deal with refusals and other factors that lead to sample loss. First, according to the requirements in the implementation process, we required interviewers to pay more visits and ask the coordinators and other respondents who had completed the interviews to try to persuade stubborn respondents in order to change their minds. The case could only be suspended after 3 serious refusals, and the interviewers needed to inform their supervisors about the reasons for refusals and fill out forms explaining the cases. Second, we mailed persuasive letters, ISSS newsletters, and copies of *China Report*[44] to the respondents to gain their trust in our program. Finally, we arranged several trials by our best interviewers, group of supervisors, and staff members from the National Family Planning Commission to help tackle the task. In this way, we made some progress in solving the refusal problem.

4.5 Baseline survey final contact result of CFPS 2010

Tables 13-15 show the contact results at the household level in the 2010 baseline survey, and Tables 16 and 17 show the contact results at the individual level. Tables 8 and 9 show the implementation rates at the household and individual levels. For detailed statistical approaches and calculation methods, see Sample Contact (CFPS-5).

[43] This is process data.
[44] *China Report* (renamed *China Family Panel Studies*) is descriptive reports on numerous hot issues of Chinese society based on the latest CFPS data.

Table 13. Distribution of the Completion Status of CFPS Baseline Survey Sample Units[45]

Status		Neighborhood Community		Village		Overall	
		N	%	N	%	N	%
Eligi-ble	Interviewed (I)	5,081	63.09	9,879	82.79	14,960	74.85
	Refused (R)	348	4.32	143	1.20	491	2.46
	Incomplete due to other reasons (O)	14	0.17	25	0.21	39	0.20
	Not contacted (NC)	17	0.21	36	0.30	53	0.27
Not eligible (NE)		652	8.10	812	6.80	1,464	7.33
Unsure of eligibility (UE)		1,941	24.10	1,038	8.70	2,979	14.91
Total		8,053	100	11,933	100	19,986	100

Table 14. Distribution of Non-Eligible Sample Units of the CFPS Baseline Survey[46]

Type	Neighborhood Community		Village		Overall	
	N	%	N	%	N	%
Wrong address	62	9.51	23	2.83	85	5.81
Non-residence	97	14.88	37	4.56	134	9.15
Vacant house	412	63.19	628	77.34	1,040	71.04
Failure in household screening	81	12.42	124	15.27	205	14.00
Total	652	100	812	100	1,464	100

[45] Technical Report: CFPS-5.
[46] Technical Report: CFPS-5.

Table 15. Distribution of the Sample Units with Unsure Eligibility in CFPS Baseline Survey[47]

Type	Neighborhood Community		Village		Overall	
	N	%	N	%	N	%
Address not contacted	7	0.36	5	0.48	12	0.40
Address correct, unable to contact the residents	662	34.11	484	46.63	1,146	38.47
Refusal to be interviewed	1,194	61.51	430	41.43	1,624	54.51
Cannot conduct household screening due to other reasons	78	4.02	119	11.46	197	6.61
Total	1,941	100	1,038	100	2,979	100

Table 16. Distribution of the Completion Statues of CFPS Baseline Survey Individual Sample Units, by Urban and Rural[48]

Status	Neighborhood Community		Village		Overall	
	N	%	N	%	N	%
Not contacted	508	3.06	1,163	2.87	1,671	2.92
Interviewed	12,793	77.03	29,797	73.49	42,590	74.52
Refusal	1,752	10.55	2,533	6.25	4,285	7.50
Incomplete due to other reasons	539	3.25	1,535	3.79	2,074	3.63
Not eligible	1,015	6.11	5,520	13.61	6,535	11.43
Total	16,607	100	40,548	100	57,155	100

[47] Technical Report: CFPS-5.
[48] Technical Report: CFPS-5.

Table 17. Distribution of the Sample individual with Ineligibility in CFPS Baseline Survey [49]

Reasons of absence	Current Residence Within CFPS Region		Current Residence Outside CFPS Region		Overall	
	N	%	N	%	N	%
Out for school	126	1.80	972	13.89	1,098	15.69
Out for work	294	4.20	5,135	73.36	5,429	77.56
Monk	17	0.24	59	0.84	76	1.09
Visiting families and friends	28	0.40	246	3.51	274	3.91
Serving sentence					19	0.27
Serving in the army					75	1.07
Abroad					29	0.41
Total	465	6.64	5,931	84.73	7,000	100

Table 18. Implementation rate of the CFPS baseline survey sample unit (%)[50]

Type	Formula[51]	Resident Committee	Village Committee	Overall
Response rate	RR3=I/(I+R+NC+O+UE)	69.35	89.16	81.25
Cumulative response rate	$RRcum=RR_{\text{resident screening}}*RR_{\text{family member}}$	69.35	89.16	81.25
Cooperation rate	COOP1=I/(I+R+O)	93.35	98.33	96.58
Contact rate	CON2=(I+R+O)/(I+R+NC+O+eUE)	74.29	90.68	84.13
Refusal rate	REF2=R/(I+R+NC+O+eUE)	4.75	1.29	2.67

[49] Technical Report: CFPS-5.
[50] Technical report: CFPS-5.
[51] "e" represents the percentage of total eligible samples in the eligibility screening

Table 19. Implementation rate of the CFPS baseline survey individual unit (%)[52]

Implementation Type	Community Type		Age Range		Overall
	Resident Committee	Village Committee	Adults	Children	
RR5=I/(I+R+NC+O)	82.05	85.07	82.52	90.67	84.14
COOP1=I/(I+R+O)	84.81	87.99	85.69	92.34	87.01
CON3=(I+R+O)/(I+R+NC+O)	96.74	96.68	96.31	98.29	96.70
REF3=R/(I+R+NC+O)	11.24	7.23	9.05	6.08	8.47

4.6 Baseline sample maintenance

CFPS made detailed sample maintenance plans to prevent sample losses and ensure effective long-term follow-up surveys. The maintenance work took several steps. The first round of maintenance started in September 2010, right after the survey season, targeting all the sampled households and neighborhood communities, including both the interviewed and the non-interviewed. We sent out mail, including thank-you letters to interviewed households, letters to those who refused to be interviewed and letters to village/neighborhood communities. The second round of maintenance was implemented during the Spring Festival, targeted at the 14,767 households interviewed in the 2010 baseline survey. During this period, the Institute conducted the supplemental interviews of CFPS and surveys on the satisfaction of health care reform with some of these households. For those households, we did "free-ride" on-site maintenance.[53] For the remaining 97 communities, we made phone calls to those who had telephones and paid visits to those who did not have telephones or for whom the phone maintenance did not work. On this basis, we sent Spring Festival greeting cards and annual newsletters to the households with valid addresses by post. For more information on the maintenance plan and the process of the sample maintenance, or the results of statistical analysis of the maintenance outcome, see *Sample Maintenance* (CFPS-18).

4.7 Follow-up strategies

After the baseline survey, CFPS follows the gene members defined in the baseline and subsequent waves and their families every two years. Prior to the completion of this manual, we have conducted the 2012, 2014, 2016 follow-up surveys. In addition, CFPS conducted a sample maintenance survey in 2011. The follow-up surveys aim to track gene members defined in the baseline survey and

[52] Technical report: CFPS-5.

[53] As it turned out later, a total of 1,226 households from 59 villages/neighborhoods were not successfully maintained in the "free-ride" way. They were combined with the 97 independent villages/neighborhoods to receive telephone/on-site sample maintenance.

follow a certain rule to determine the target respondents and families in each subsequent wave to ensure the representativeness of the sample. CFPS adopts the following follow-up strategies:

(1) Gene members are always interviewed in follow-up surveys.
(2) Core members are interviewed only when their relationship with the gene members is maintained.
(3) All members will be contacted in subsequent waves regardless of their interview status in the previous wave, except those who were reported as deceased in previous waves.

CFPS used CAPI in the baseline survey and began to add CATI as an assisting tool in 2012 in response to the increasing migrant samples in CFPS. CATI was introduced mainly for non-co-residing members and households moving out of the initially sampled areas. Those respondents were hard to reach for face-to-face interviews. We constructed the CATI questionnaires based on the core contents of the face-to-face questionnaires. Also, starting from the 2012 follow-up survey, CFPS introduced proxy reports to reduce information loss due to failure to track individual respondents. If a family had a non-coresiding member who was not present at the time of the inteview, we invited the resident family member who was the most familiar with the non-co-residing member to complete a proxy questionnaire that collected some basic information on the non-coresiding member. Attempts were also made to track the non-co-residing member for a self report. If we succeeded in tracking the respondent, we invited the respondent to finish the self-answered questionnaire, which could be done in either a face-to-face or a telephone interview depending on the preference of the respondent. If we failed to track him/her down, we would still have the information in the proxy questionnaire. Therefore, starting from 2012, there were 3 forms of individual questionnaires: face-to-face self report, telephone self report and proxy report.

4.8 Field operation of follow-up survey

Before every full-sample follow-up survey, CFPS conducted a pre-survey or pilot survey to ensure good implementation of the full-sample follow-up survey. For the 2012 and 2014 pre-surveys, we conducted small-scale surveys in one county in Guangdong which had a large migrant population, one county in Gansu, and eight counties in Beijing. The pre-survey was intended to test the survey system, evaluate the difficulty in tracking samples and help optimize the process in practice. We started to integrate CAPI and CATI questionnaires in 2016, so we conducted a pilot study to test the questionnaires with a convenience sample and a pre-survey using a real sample. Through the former, we evaluated the time length of telephone interviews, and improved the questions based on the feedback from interviewers and interviewees.

The test sample of the pre-survey included around 500 households which were scattered at the county level and considered hard to reach for face-to-face interviews. We further tested the telephone interview system and estimated response rates from the pre-survey.

CFPS employs about 400 to 500 interviewers per wave. With the integration of the CATI and CAPI, the number of telephone interviewers has been on a steady increase and in CFPS 2016, we had over 70 telephone interviewers. Most of the training sessions were done at Peking University, and a few were done in other collaborating universities/institutions. In the 2010 baseline survey, most interviewers were local people, and fewer than a quarter of the interviewers were students. The proportion of student interviewers steadily increased in each follow-up survey, growing from one half in 2012 to over three quarters in 2016. The CAPI interviewers were mainly from the local areas, while CATI interviewers were mainly college students in Beijing and professional telephone interviewers from cooperating institutions.

As the target respondents became geographically more scattered in the follow-up survey than in the baseline survey, we implemented the follow-up survey in several stages. In 2012, there were three stages of the field operation: in the first stage, which started on July 20, 1012 and ended On November 30, 2012, we revisted the original addresses from the last wave and conducted face-to-face follow-up interviews; in the second stage, which started on September 19, 2012 and ended on January 18, 2013, we tried to track the non-residing family members in their new places; in the third stage, which started on Feburary 1, 2013 and ended on March 4, 2013, we conducted supplementary interviews during the Chinese Spring Festival, when many migrant workers return to their hometowns. We made a great effort to track non-coresiding members, who fell into four categories: non-coresident family members, members of split families, whole families temporarily leaving their original addresses, and relocated households. The follow-up strategy for these four types consisted of 3 steps: (1) reassign the samples to the nearest interviewers for face-to-face interviews in the new place; (2) dispatch a specialized follow-up team to follow the respondents who lived far from any local interviewers; The specialized team included both hired interviewers and survey supervisors from ISSS. (3) telephone interviews. CFPS conducted telephone interviews if respondents chose to be interviewed. On average, one supervisor was in charge of about 16 counties and 40 interviewers. The field operation team in 2012 consisted of a survey manager; 4 production managers, each of whom was in charge of preparation, training, implementation and telephone interviews; and 10 production managers for different provinces. In 2012, we also did a paper questionnaire to improve some dubious household information from the baseline survey.

The follow-up survey in 2014 also was done in three stages: during the first stage , which started on July 4, 2014 and ended on June 7, 2015, we revisited the

original addresses and conducted local follow-up interviews. During the second stage which started on August 8, 2014 and ended on May 18, 2015, we conducted telephone interviews. During the third stage, which started on February 7, 2015 and ended on March 22, 2015, we conducted supplementary interviews during the Chinese Spring Festival. In 2014, CFPS developed a tracking system to meet the researcher's needs when following a large scattered sample. The main functions of the system include scheduling and conducting CATI interviews, allocating newly-split family samples, and supervising the interview process. We improved our sample allocation process from manually organizing spread sheets to using the tracking system to allocate, assign and manage the follow-up samples. The tracking system had the following main features: First, it displayed all the telephone numbers collected from previous waves, screened the numbers and marked the invalid ones. This improved efficiency in contacting respondents. Second, the system streamlined the scheduling and interviewing processes for CATI. Third, it kept track of the progress of interviews and result codes of the samples. The tracking system was shown to successfully increase the efficiency of the interviewing processes, improve the accuracy of the sample allocations, and thus significantly increase the response rates of the survey. For those who were unwilling to do face-to-face interviews during the initial household visit, the sample would be available in the tracking system for re-assignment for CATI interviews. Then the telephone interviewer in our center would call to arrange an interview and decide whether to conduct a face-to-face interview or a telephone interview acording to the respondent's choice and the distribution of non-local follow-up interviewers (face-to-face interviews could be arranged when there were 3 or more respondents in a city). In addition, for respondents who moved out of the 25 baseline provinces, the system would arrange telephone interviews.

The 2016 follow-up survey was conducted in two stages: During the first stage, which started on June 28, 2016 and ended on April 30, 2017, we revisited the original addresses and conducted local interviews. During the second stage,which started on May 13, 2016 and ended on April 30, 2017, we conducted telephone interviews. Some CAPI interviewers also became CATI interviewers in this round. They were allowed to conduct CATI interviews outside the center. As CATI questionnaires in 2014 were longer than the questionnaires used in previous waves, this task became more difficult than formerly. We made three adjustments in the field operation: First, we simplified the process of initiating telephone interviews to increase efficiency and response rate. If the respondent chose to be interviewed by telephone at the time when the appointment was scheduled, the system would immediately generate a telephone questionniare and proceed directly to the interview. Second, we encouraged the experienced CAPI interviewers to participate in telephone interviews. They could be assigned to conduct telephone interviews from a distance via internet calls. Third, we contracted some CATI samples to a telehpone interview company in order to complete the telephone interviews within a shorter time period.

In addition to the full-sample survey every two years, CFPS conducted a sample maintenance survey in 2011. The sample maintenance survey was implemented in two stages: local interviews and follow-up interviews outside the local areas. The local interviews during the first stage were done in CAPI. The follow-up interview targeting at migrant teenagers tried a mixed interview method, which combined telephone interviews, internet interviews and mailed paper questionnaires. The face-to-face interviews started on July 21, 2011 and ended on November 23, 2011. The mixed interviews started on December 29, 2011 and ended on January 9, 2012. In addition, following the earlier sample maintenance strategies, we regularly maintain the sample every year.

4.9 Interview results at the household level

Tables 20 and 21 show the results of the CFPS 2012 and CFPS 2014 follow-up surveys respectively. The cross-sectional response rate was 79.4% and 77.9% respectively for 2012 and 2014. We divide the sample into completed sample and incomplete sample based on the interview status in the previous wave. In 2012, the cross wave retention rate of the baseline families[54] was 85.3%. The cross-sectional response rate of newly-split families was 35.9%. In 2014, the adjacent-wave retention rate among the completed samples from 2012 was 89.7%. The cross-sectional response rate among the incomplete sample in 2012 was 44.6%.

Table 20. Interview status at household level in CFPS 2012

	All households		2010 sample		New in 2012	
	Count	Percentage (%)	Count	Percentage (%)	Count	Percentage (%)
Samples from last wave	14960		14960		0	
Newly-split family	2031		0		2031	
Exit Families (all deceased members)	37		35		2	
Total	**16954**	**100**	**14925**	**100**	**2029**	**100**
Failure to contact	1847	10.9	845	5.7	1002	49.4
Refusal	887	5.2	744	5.0	143	7.0
Lost contact information due to moving	469	2.8	410	2.7	59	2.9
Respondents unable to complete	298	1.8	202	1.4	96	4.7
Complete	13453	**79.4**	12724	**85.3**	729	**35.9**

[54] Adjacent wave retention rate was the proportion of completed cases among those completed from the last wave, excluding the exit families with only deceased members.

Table 21. Interview status at household level in CFPS 2014

	All households		Completed 2012		Incomplete in 2012	
	Count	Percentage (%)	Count	Percentage (%)	Count	Percentage (%)
Samples from previous waves	14925		12724		2201	
Newly-split families	3286		729		2557	
Family died out	60		36		24	
Total	**18151**		**13417**		**4734**	
Failure to contact	1938	10.7	624	4.7	1314	27.8
Refusal	1215	6.7	426	3.2	789	16.7
Lost contact information due to moving	589	3.2	209	1.6	380	8.0
Respondents unable to complete	265	1.5	124	0.9	141	3.0
Complete	14144	**77.9**	12034	**89.7**	2110	**44.6**

More information about the follow-up at the household level could be found in Chapter 12 of *China Report 2016*.

4.10 Interview results at the individual level

Tables 22 and 23 show the results of the follow-up survey at the individual level. The cross-sectional response rates[55] in 2012 and 2014 were 74.1% and 72.8% respectively, and the adjecent wave rention rates were 80.6% and 83.8%. Among the 6 sampling frames, the adjecent wave rention rates[56] and recovering rates for the incomplete samples in Shanghai and Guangdong were relatively low, resulting in significant loss of samples (see Table 23). Compared with that in 2012, the adjacent wave retention rate in 2014 increased, but the recovering of incomplete samples from previous waves became increasingly difficult. Table 24 shows the forms of individual questionnaires for completed samples. The proportion of CATI interviews increased as the sample became more geographically scattered.

Table 25 shows the number of gene members in each wave and their interview status. After two waves of follow-up, the number of gene members increased by 2,004.

[55] Deceased members and members who do not need to be followed up are deducted from the denominator when calculating the response rate.
[56] The completion rate in this round of the incomplete sample in the last round after deducting the unqualified sample.

Table 26 shows the interview status in all three waves, where strictly completed means the respondent finished the self-answered questionnaire, and loosely completed means the respondent finished the individual questionnaire. Under the loose criteria, 8.9% of the gene members never finished any individual questionnaire, and 45.6% completed all three rounds of the follow-up survey. Under strict criteria, 87.9% of the gene member finished at least one interview.

Table 22. Interview status of individual samples by status of last wave

	CFPS 2010					
CFPS 2012	Complete	Incomplete	No need to follow-up	Out of county non-coresidents	New sample	Total
Complete	33956	3637	43	3974	2729	44339
Incomplete	8185	4170	38	2377	748	15518
Deceased	42	18	41	35	4	140
No need to follow-up	407	205	1	26	0	639
	CFPS 2012					
CFPS 2014	Complete	Incomplete	No need to follow-up		New sample	Total
Complete	36856	6075	52		2722	45705
Incomplete	7103	9122	43		774	17042
Died	97	75	39		1	212
No need to follow-up	283	246	2		0	531

Table 23. Retention rate of individual sample by sampling frames by last wave status

	Complete in the previous wave		**Incomplete in the previous wave**		**New sample**		**Total**	
	n	**retention rate**	**n**	**retention rate**	**n**	**retention rate**	**N**	**retention rate**
				CFPS 2012				
Shanghai	3522	65.7%	807	31.2%	185	72.4%	4514	59.9%
Liaoning	3658	80.8%	994	43.2%	279	77.0%	4931	73.1%
Henan	5005	86.9%	1486	67.9%	475	85.1%	6966	82.8%
Gansu	4917	87.1%	1957	66.8%	419	76.6%	7293	81.1%
Guangdong	4185	76.3%	2238	51.6%	336	71.1%	6759	67.9%
Others	21303	80.8%	7083	51.9%	1787	79.4%	30173	74.0%
National	**42590**	**80.6%**	**14565**	**53.8%**	**3481**	**78.5%**	**60636**	**74.1%**
				CFPS 2014				
Shanghai	2677	75.5%	1799	20.4%	195	62.1%	4671	53.9%
Liaoning	3552	87.2%	1319	35.6%	200	75.5%	5071	73.5%

Henan	5690	89.8%	1195	48.4%	569	86.3%	7454	83.0%
Gansu	5833	86.2%	1380	53.4%	416	83.7%	7629	80.3%
Guangdong	4537	74.5%	2160	29.5%	379	66.0%	7076	60.4%
Others	22050	84.1%	7801	44.8%	1738	78.4%	31589	74.1%
National	**44339**	**83.8%**	**15654**	**40.1%**	**3497**	**77.9%**	**63490**	**72.8%**

Table 24. Forms of questionnaires for individual questionnaires

	Completed interview	**Total**	**Self report**		**Proxy report**	
			CAPI	**CATI**	**CAPI**	**CATI**
CFPS 2012	n	44339	40504	1027	2806	2
	Percentage	100.0%	91.4%	2.3%	6.3%	0.0%
CFPS 2014	n	45705	39450	2238	2829	1188
	Percentage	100.0%	86.3%	4.9%	6.2%	2.6%

Note. The contents of CAPI and CATI versions of the proxy report are the same.

Table 25. Number of gene members and their interview status

	Complete	Incomplete	No need to follow-up	Deceased	Out of county non-coresidents	Total	Response rate
2010	42590	8030	123	0	6412	57155	84.1%
2012	42964	14971	136	639	0	58710	74.2%
2014	43043	15918	198	528	0	59687	73.0%

Table 26. Interview status of gene members

Completed waves	Loose criterion		Strict criterion	
	Count	Percent (%)	Count	Percent (%)
0	5729	8.9	7729	12.1
1	13417	20.9	14164	22.1
2	15744	24.5	15075	23.5
3	29243	45.6	27165	42.4
Total	64133	100.0	64133	100.0

More information about the follow-up results at the individual level could be found in Chapter 13 of *China Report 2016*

5. Quality Control

5.1 Quality Control Measures and Technologies

5.1.1 Baseline survey

Strict quality control measures were applied in the 2010 CFPS baseline survey to ensure the quality of the data. For the factors that might affect the data quality, such as improper designs of the questionnaires, inaccurate terminal-stage sampling, irregular behaviors of the interviewers, mistakes in data collection, and compilation processes, we applied different methods of monitoring and intervention, including telephone checks, field checks, audio record checks, interview reviews, statistical analyses, and so on.

For example, if interviewers used substitutes for the sampled households/individuals or interviewed the wrong households/individuals, the representativeness of the sample would be affected and thus cause trouble for data users. Therefore, the quality control team applied different measures before and after the survey was done. To prevent arbitrary substitutions and interviewing the wrong households, the sampling staff or the cadres of the neighborhood communities would deliver the letters for the households to the correct addresses before the interviews. Then the residents in these households would call the Institute to report the required information. The quality control team compared this information with information collected by the interviewers to ensure consistency. After launching the interviews, they also confirmed the accuracy of household sampling by field check: the field checking team documented the number of eligible households at the address and provided feedback. The field checking team was also responsible for documenting interview refusals and non-contacts. In the case of arbitrary substitutions and interviewing the wrong individuals, the team mainly used field checks, telephone checks, and audio record checks to perform post-hoc quality control. In addition, if interviews had been done with wrong households or individuals, re-interviews were conducted with the correct ones. We had put much emphasis on the behavioral code of the interviewers during their training sessions.

Moreover, regarding systematic biases caused by improper designs of questionnaires, we did statistical analyses of the questionnaire data and Para-data weekly to identify the sources of errors. We then revised the questionnaires accordingly and updated the interviewing system to control the quality. For inaccuracies in the terminal-stage sampling, field checks were used. For irregular behaviors of the interviewers, such as the use of leading questions, or cutting corners during the interviewing process, we used telephone checks, audio records and paradata analysis to help ensure the quality of the data.

With the CAPI system, the likelihood of systematic biases in data collection and compilation was much lower than in the traditional surveys with paper and pencil. However, biases could not be completely avoided during the coding process of open-ended questions. Three coders were asked to code the same questions using systematic coding tables. There were also errors in the data compilation process, which could be controlled via cyclic inspections by multiple people.

5.1.2 Follow-up survey

Starting from 2012, since the data quality is no longer affected by precision in the ultimate sampling frame, we no longer use the supervision strategy targeted at the problems in the ultimate sampling frame as we did in 2010. We have continued to use statistical data checks, audio record checks, and phone call checks in the follow-up surveys. In 2012 and 2014, we conducted no on-site field checks due to cost constraints. Meanwhile, the software recall technique was replaced by data check in 2012 due to its cost and low efficiency.

In 2012, we further delineated the problematic interviewing behaviors and set objective criteria for judgements of misconduct. For data checking, we added item-level time-length checks. Before starting the interview, we set a minimum interview time length for each question. After collecting the data, we compared the actual time length with the minimum to get a preliminary assessment of the quality of the interview. We further defined 12 types of misconduct in the interview, including "imagined answer," "short-cut skip," "typo," "inaccurate reading of questions," "insufficient probing," and "illegitimate proxy report." We set objective criteria for each type of misconduct. We changed the audio record check from replaying full-sample recording to item specific recording, which increased efficiency and pertinence.

5.2 Quality Control Strategies

Our quality control is comprehensive. First, all the variables in each questionnaire and Para-data should be checked using statistical methods regularly (every 7 days). Second, all the interviewers should go through all the checking methods. Third, for each interviewer, data from each questionnaire need to be checked. Fourth, each type of contact results by each interviewer also needs to be checked.

Since 2012, statistical data check has covered 100% of the complete interviews. The checking cycle has been shortened from every 7 days to every day. When problematic interviewing behaviors are detected, we increase the level of monitoring for that inteviewer. We apply alternative checking strategies for samples that are ineligible for certain quality checks (e.g., samples with no audio recording) to ensure the coverage. In addition, since 2014, we have adjusted the percentage of samples for quality checks for interviewers based on the quality control results of his/her existing complete interviews.

Our quality control has substantial coverage. First, 60% of the addresses of incomplete interviews have received on-site field checking. Second, the households whose questionnaires for families or individuals were completed are to be checked via audio record, phone calls, on-site, or interview reviews by the percentages of 15%, 25%, 15% and 5%, respectively.

We made significant changes to the rules for selecting samples to be checked in 2012. In the baseline survey, the samples to be checked were selected randomly. Since 2012, the samples to be checked have been a combination of targeted samples and randomly selected samples. First, all the samples that have failed the statistical data check are monitored via audio recordings. Then, we randomly select 10% of the samples that pass the statistical data check and divide them into two groups, one for phone call check and one for audio record check. After that, we check the first three cases of each type of questionnaire for each interviewer to see if the interviewer has acquired the interviewing techniques. Finally, when misconduct is detected by one method, we use another method to double check.

Regarding the procedure for quality control, we always use statistical analysis of data for overall data verification. When the sample size of each type of questionnaire reaches 30, the data is merged every 7 days to check for systematic errors. Audio record check is the preferred measure to be used in the overall check, which starts on the second day after receiving the collected data from interviewers and continues to the third day after the end of the survey season. The phone check is preferred for the cases without audio records or cases that failed the audio record checks, which starts on the second day after receiving collected data. The field check is mainly used to examine the accuracy of the terminal-stage sampling, which starts within 20 days after receiving the collected data. The interview review is mainly used for the problems found in previous forms of checks, such as too short or too long interview time. The whole process of quality control emphasizes timely responses.

Since 2012, we have set the statistical data check as the first step in quality control, preceding other monitorings. We conduct daily statistical analyses on all the completed interviews using SAS programs to identify suspicious cases more efficiently. After the data check, audio recording check is preferred for the samples that are completed in the earlier stages of the survey that fail the statistical data check and that are randomly selected but without valid phone numbers. A phone-call check is preferred for randomly selected samples.

5.3 Proportions and Results of Quality Check[57]

5.3.1 Baseline survey

A total of 28% of all the households interviewed received audio record checks successfully, covering 16% of all the completed questionnaires.

A total of 19% of all the households interviewed received telephone checks successfully.

During the survey season, a total of 25% of all the households received field checks. In addition, during the periods between December, 2010 and February, 2011, and July to November,2011, the Institute conducted another two rounds of field checks on part of the sample and the entire sample, respectively.

For interview reviews, 3% of the valid cases were involved.[58]

The statistical analyses covered the interview time, non-response rate, outliers, internal consistency reliabilities of the attitude scales, and so on. Reports were submitted to the quality supervision department and the department of survey implementation on a weekly basis.

The results of quality control show that there were no interviews of wrong households among all the households checked via telephone, in field and record. A total of 81 adult questionnaires were not answered by the respondents themselves, with 21 interviewers involved. Among them, 59 questionnaires were affected by the misconduct of one single interviewer; 7 questionnaires were answered by other family members when the respondents were not at home;

[57] Results come from the Technical Report: CFPS-4.
[58] In the mid-late period of the survey, the method of interview reviews was replaced by the statistical analysis of the time used in the interviews due to its high cost and low efficiency.

for the remaining 15 questionnaires, either the respondents were unable to answer questions or others cut in to answer the questions. For the child questionnaires, 22 questionnaires that should be answered independently by children themselves were completed by others, with 20 interviewers involved. Among those cases, 7 questionnaires were answered by parents. In addition to these cases, we also discovered several serious cases of misconduct in field checks, which were documented in Chapter 4 of this manual.

Most interviewers strictly complied with the behavioral code of interviews. The quality control suggested that the short interview durations in some interviews were due to interviewers' misbehaviors such as speeding up and skipping questions. We constructed a "sample of short interviews" that included the interviews with very short interview time and the interviews where the questions answered in a very short time accounted for more than 50% of all the questions monitored. This sample contained 1,051 cases in total (including questionnaires for families and individuals) and involved the work of 50 interviewers. Among all questionnaires that were randomly checked, 7,914 had at least 1 question omitted.[59] We took specific actions on this problem during the survey and made significant improvement.

The quality control of the 2010 CFPS not only was effective in ensuring data quality but also yielded much experience in the measures and strategies of quality control. The detailed design and the quality control process can be found in *Quality Supervision Report* (CFPS-4).

5.3.2 Follow-up survey

The statistical data check in 2012 covered 52,545 completed samples, among which 4896 failed the check, accounting for 9.32% of the total sample. About 35% of the completed household cases (n=18,515) also went through audio recording checks. Among the checked samples, 1,030 failed, accounting for 5.56%. We conducted phone call checks on 3,062 families (26.54% of the total sample) that had completed the interviews. Among them, 566 failed, accounting for 18.48%. Excluding 352 samples that failed the check only for remuneration problems, the adjusted failure rate in phone call checks was 6.99% of the failing samples.

The statistical data check in 2014 covered 67,482 completed individual observations, among which 2,531 failed the check, resulting in a failure rate of 3.75%. We checked 15,928 (20% of the total sample) observations via audio records. Among them, we succeeded in checking 15,484. Among the samples checked, 579 failed to pass, accounting for 3.7%. We selected 4,904 families for call back checks. Among them, we succeeded in reaching 3,745 families and the success rate was 76%. Among them, 207 failed to pass the check, accounting for 5.57%.

[59] The quality of the interview could not be judged simply by the problem of leaving out questions. Among the questions omitted, despite some that were left out on purpose, the vast majority were obvious according to the observations or the information that the interviewers already acquired, e.g., when the interviewer saw the respondent calling on the mobile phone, he would directly answer yes to the question "Do you have a mobile phone" without asking.

More information about the quality control implementation and results can be found in forthcoming technical reports.

6. Data Sets and Data Processing

6.1 General Introduction to Data Sets

The CFPS 2010 baseline data consists of data sets based on the community questionnaires, the family roster questionnaires, the family questionnaires, the adult questionnaires, and the child questionnaires. The general structure of the data sets in CFPS follow-up surveys are similar to that in the baseline survey. However, as CFPS did not administer community questionnaires, only the data sets in CFPS 2012 and 2014 contain responses from community questionnaires. Since CFPS2012, we have constructed a cross-year person-level file to keep track of the interview status and basic demographic information on all individuals ever entering the CFPS study.

As introduced in the chapter on sampling, there are 6 sampling frames representing 6 subpopulations. In the data set released, we marked different subpopulations with the indicator variable "subpopulation." The values from 1 to 6 of the "subpopulation" represent Shanghai, Liaoning, Henan, Gansu, Guangdong, and other provinces/municipalities respectively. Also, we added a dichotomous indicator "subsample." The value 1 of the "subsample" refers to the resampled national sample.

The complete national sample includes all the CFPS data, composed of the 6 subsamples representing the 6 subpopulations. After weighting, the complete national sample represents the national population. The resampled national sample was constructed by resampling the five "big provinces" proportionally to the "small provinces." The resampled national sample is directly representative on the national level (i.e., without weighting). For details on weights, please refer to Chapter 9.

Table 27. Number of Variables and Sample Size of CFPS 2010 Baseline Data Sets

	# of Vars	Sample Size						
		Full sample	Resampled sample	Shanghai	Henan	Gansu	Liaoning	Guangdong
Community	221	635	417	58	64	65	63	64
Family roster	355	57,155	36,964	4,329	6,491	6,874	4,652	6,423
Family	624	14,798	9,661	1,405	1,506	1,537	1,478	1,394
Adult	1,493	33,600	21,812	3,162	3,732	3,704	3,129	3,070
Child	968	8,990	5,944	360	1,273	1,213	529	1,115

Note: See Codebook for detailed information of the variables.

In addition the public use data sets abovementioned, CFPS also has a county data set for restricted use.[60] The county data set contains a series of macro variables at the county level in the CFPS sample (county GDP, GDP per capita, population, employment rate, average years of schooling, proportion of population in the working age, proportion of old population, sex ratio between age 10 to 19, ratio of non-agricultural population). This county data set does not contain actual county codes or names, but users can link to the public CFPS data set (e.g., family data set and individual data set) via a sequential county code. The macro variables are mainly from 2010 national statistical yearbook. In order to protect the data confidentiality, the data values in the CFPS county data set have been blurred to minimize the possiblity of reverse identification, but their statistical properties have been maintained. Detailed information about the masking process can be found in *Technical Report CFPS-23.*

6.2 Data Cleaning

6.2.1 Cleaning of the Family Relations Data set

The main content of the CFPS 2010 family relations data set comes from the three tables T1, T2, and T3. As mentioned above, these three tables were used in the questionnaires to collect social-demographic information on family members and their relations, which would provide important information and help in the data cleaning process later.

The data cleaning of the entire 2010 family relations data set was divided into several stages. In the first stage, the basic cleaning focused on correcting errors in the implementation process, such as dealing with wrong information input by interviewers and invalid questionnaires (e.g., cheating questionnaires, and repeated questionnaires), correcting the wrong sample codes of the households, etc. This part of data cleaning was mainly based on the information feedback from the implementation team in the field.

In the second stage, the basic match cleaning focused on the correspondence between the individuals in the family relations data set and the individuals in the adult and child data sets. According to the designs of CFPS, the family members with their individual codes starting with the number "1" in the family relations data set need to answer the individual questionnaires. Therefore, if the contact result showed that they had completed their questionnaires, we needed to find the corresponding individual questionnaires in the individual data sets at this stage. At the same time, we needed to confirm that the questionnaire found in the individual data sets matched the family member himself/herself—there had to be strict correspondences. We checked the correspondence through some key variables such as names and birth dates and corrected mistakes in the confirmation process.

[60] Restricted-use data, as opposed to public use data, is accessible with further application. For use public use CFPS data, users only need to file an online application and get approval. Then users could download the data from the internet. For restricted-use data, users should file a special application, stating the research purpose and specific use of the restricted-use data.

The third stage was the deep cleaning of the family relations data set. After the first two stages, we had already done the matching work in the following two aspects:[61] First, we matched all the family members and the immediate relatives in the T1 and T3 tables according to the relation index in the T2 Table. We found logic errors or questionable facts in some samples, such as female fathers, male mothers, situations where the husband and wife did not know each other, or age gaps between parents and children that were too small or too large, etc. Second, we matched the family relations data set and the individual data sets and also found some logic errors or questionable facts, such as situations where the total number of children the respondent's spouse reported was different from the number the respondent reported, or the spouse was no longer a living member of the family based on family relations data but the marital status in the individual data sets was "married" or "cohabitation," etc.

For the data cleaning in the third stage, we first migrated to a new system version and solved some errors in the codes of family members caused by the computer program. All other errors were corrected manually based on reliable information sources such as the 2010 family relations data set and the 2010 individual data sets, the 2011 family relations data set and the 2011 individual data sets. If there were no reliable references, we normally defined the absolute logic errors as "missing," e.g., wrong gender for father, same gender for spouses. For other questionable details that were not absolutely wrong, e.g., a small age gap between parents and children, we left them unchanged. For the detailed process and methods of manual data cleaning, see *Data Cleaning of the Family Relations Data Set* (CFPS-7).

We used three T tables at baseline to construct a family tree network depicting the relationships among all family members within a household. In subsequent waves, the family roster questionnaire focused on capturing changes over time, such as family splitting, members leaving households, and new family members and their relationships with existing family members. During data processing, we try to update the family network based on the information on changes. For reasons such as adult children gaining financial independence and divorces, families may split. When this happens, we designate one family network as the original family keeping their original family ID,[62] and the other network as the new family with a new family ID.

The constructed family foster dataset at each subsequent wave only contained families that we successfully interviewed at that wave, without families from previous waves but lost during that wave.[63] The construction of a family roster dataset consists of two main steps. The first step is to update the family roster dataset with new family members and their relationships with existing family members. The second step is to add newly split families. Similar to the baseline family roster dataset, we performed a series of consistency checks related to the gender, age and marital status.

[61] Se e detailed matching method in the technical report: CFPS-6.
[62] In general, we set the family with an earlier family interview as the original family (usually the unit at the original address), and the other split unit as the new family. The distinction between the original family and the new family was sometimes arbitrary, especially when both moved away from their original addresses.
[63] There were a few families in which all members died. Such families are not included in the family roster dataset, but are present in the cross-wave individual status dataset.

The family roster dataset in subsequent waves has had a few distinct features compared with those of the baseline.

First, members from split families may have multiple records in the family roster dataset, one in the original family and one in the new family, in order to reflect the dynamic changes of family members. Users can identify the family ID of the individual respondent by restricting to samples with the variable co_aXX_p with XX representing the survey year (i.e., 12, 14 etc). The corresponding family ID (fidXX), linked with co_aXX_p=1, is the actual famliy ID of that respondent for that wave, and the fidXX linked with co_aXX_p=0 indicate the original family of the respondent. If we restrict our samples to only those with co_aXX_p=1, then each person's ID (pid) has only one observation without any duplicates, which means that a pid has only one valid fid for a certain wave.

Second, the family roster dataset in subsequent waves contains the family ID for the current wave as well as for previous waves. Variable names suggest that corresponding year of family ID (e.g., fid10, fid12, fid14).

Third, we designate different types of gene members depending on whether they belong to splitting familes, whether they physically live in the household, and whether they are deceased. We use the variable genetype to indicate the individual's membership (e.g, resident gene member, new gene member, non-coresident gene member, deceased gene member, etc.). Users may refer to the technical report CFPS-33 for more detailed information.

6.2.2 Cleaning of Other Data Sets

The cleaning of the other data sets was also divided into several stages.

Data cleaning in the first stage mainly focused on data sets of individual questionnaires, which was done at the same time, with the deep cleaning of the family relations data set for cross-validation. On the one hand, while cleaning the family relations data set and checking the correspondence between the information in these two data sets, we found some mistakes in the individual data sets, e.g., a big gap between the ages of marriage claimed by the spouses, a spouse who was no longer alive in the family relations data set whereas the marital status in the individual data sets was "married" or "cohabitation," etc. If the problem was with the individual data sets after careful inspections, we corrected the errors in this data set. On the other hand, there were some logic errors and questionable details in the individual data sets themselves, e.g., the date of divorce was prior to the date of marriage, the date of marriage was prior to the date of birth, the age of marriage was less than 16, etc. In these cases, we relied on the information from the family relations data set to make corrections. The data cleaning principle of the individual data sets is consistent with that of the family data set. We dealt with such mistakes if there were reliable information sources. If not, we defined those absolute errors as "missing" and left those uncertain situations unchanged. Moreover, for important variables, instead of directly correcting the original values, we constructed a group of new "best variables" that summarized the most reasonable information after data cleaning. See the detailed introduction to "best variables" in the section on "Composite variables" below.

In the second stage, we did the cleaning in all data sets other than the family relations data set, from the end of the first stage to the final release of the data. Making careful checks over all variables, the data team deleted redundant variables, supplemented some of the missing variables, and corrected wrong variable names, labels and values, as well as some obviously wrong values.[64] However, although we paid great attention to some extreme values that were likely to be unreasonable, in most cases we accepted the original values unless there was particularly reliable evidence for correction. For example, during the data cleaning process, we found some outliers in asset and income variables. The CFPS data processing team tried to verfiy those data via audio recordings whenever possible, and corrected the values if there was a recording error. Most mistakes resulted from recording the wrong unit. More detailed information about data cleaning for asset variables can be found in *Technical Report: CFPS-29.* Starting in 2015, we conducted audio recording checks simultaneously at the time of the field interviews, focusing on a number of other variables besides income and assets.

Meanwhile, we also tried to harmonize data from different forms of questionaires. As mentioned before, CFPS added telephone questionnaires and proxy questionnaires to face-to-face self-report questionnaires. The original data from these two types of questionnaires are separate data sets. To make it easier for users to minimize the possiblity of missing valid observations, we integrated these data according to the following principles:

(1) We harmonized the variable names from both the CAPI and the CATI surveys. When CAPI and CATI had different measures,[65] the integrated dataset kept both variables and set the values of the variables in the other form as missing. Most respondents had either a CAPI inteview or a CATI interview, but when there were duplicate cases across a CAPI and a CATI, we retained the CAPI observation, dropped the CATI observation, and set the mode of interview (as indicated by the variable Iwmode) as CAPI.

(2) We also harmonized self reports and proxy reports. If a self report and proxy report used different measures, we kept both variable names and set the values of the variable in the other form as missing. When a variable had values from both a self report and a proxy report, we used the data from the self report.

6.2.3 Field Work Collection and Data Editing

Because of the efforts of the data team, most of the problems in the data sets had already been solved. However, there was still some data from a small proportion of households/individuals which could not be corrected due to a lack of reliable information or evidence. Post-survey field work collection is an important method of data verification. Considering the high cost, we used this method only for households or individuals with missing or wrong values in key variables which could not be fixed via regular ways. Until now, most post-survey field work has been targeted at key variables in the 2010 baseline survey that significantly affected the later follow-up survey.

[64] For example, one household chose the value "-8" in the variable fe3 "Do you participate in or manage any non-agricultural industries," but there was valid data in the non-agricultural module indicating the answer to this question. Therefore we changed the value of the variable fe3 to "1" (yes).

For those data, we tried to supplement or confirm some key information, undertaking the following efforts.

(1) Return Visits

For households with logic errors in the CFPS 2010 family relation data set, which we were unable to clean due to lack of information, we conducted return visits in 2012. These visits mainly aimed to fix the following three types of errors. The first type of error was due to lack of information. For example, concerning the unreasonable age gap between parents and children (less than 15 or more than 50), the reasons for this "error" could be: (1) it was indeed the fact and thus not an error; (2) the ages of the parents were wrong; (3) the ages of the children were wrong; (4) the children were not the parents' biological children; (5) other family members' information was mistakenly entered instead of the children's. If we could not acquire accurate information in the data cleaning process, we put these families into our list for return visits. The second type of error concerned the blood relationship between parents and children. According to the original design of the CFPS, the parental relationship should refer to the "blood relationship," a term which includes adoptions herebut excludes foster-parenting). But in field operations, this rule wasn't strictly obeyed, resulting in different standards being used to define parental relationships in the survey. Therefore, we listed all the reconstructed families in our return visits and screened the types of their children so that the blood relatives and fostered children could be accurately separated. The third type of error involved individuals who had no family relations. They either made up a household composed of themselves alone, which was rather rare even if they had no immediate relatives; or they had no relations with other family members although they were permanent residents of the households. These individuals' information was marked as questionable, for they might not be eligible as family members or some of their family relations may have been omitted in the survey. To regain their family relations and clarify the definition of family members in the baseline survey, we decided to visit households with these types of individuals again in the year 2012.

Data problems varied from family to family, which required interviewers' active probing. Therefore, a standard questionnaire and standard interviewing procedures could not be used in the return visits. Instead, we conducted paper and pencil interviews with case-by-case questionnaire designs, each of which included specific questions and the specific ways of asking and answering questions for specific families in order to confirm or retrieve information. The return visits took place simultaneously with our 2012 survey. For the households on the list of return visits, the interviewers needed to complete the case-by-case questionnaires that were specially designed for the specific families after finishing all the interviews in the 2012 survey.

At the end of the 2012 survey, the information we collected in the return visits was used to correct and update the 2010 data. For the first and third types of errors, we supplemented and corrected the information in the data sets. For the second type of errors, we created a new variable (bio_cN). This variable represents whether the person, who answered the questionnaire and the corresponding child were blood related or not. In such cases, the value "1" represents a blood relationship between the household head and the child, "0" represents not blood related, '-

8" represents not applicable, and "-9" represents data missing.[66] The bio_cN variable can help researchers understand the blood and adoption relationships in reconstructed families, which also explain why there might be reasons for parents and children not recognizing each other and for the age gaps between the parents and the children being small.

For more information on the return visits, please refer to *Data Cleaning of the Family Relations Data Set* (CFPS-7)

(2) Repeated Questions, Confirmations, and Supplemental Questions

In addition to revisits, we included questions in the 2012 survey to reconfirm information gained in the 2010 survey. The reconfirmation mainly focused on important variables, which were hard to correct due to lack of information, and key variables, which were influenced by improper questionnaire design or problematic implementations in 2010. Detailed information is listed below:

(1) Gender and date of birth of the respondent.

(2) Marital status in 2010 and important time information in his/her marital history, e.g., date of the marriage, spouse's date of birth, date of divorce, etc.

(3) For respondents who were divorced or widowed in the interviews in 2010, we asked about the education level of their last spouse since the T table design in 2010 did not collect this information. We re-collected it in the 2012 questionnaires, believing that it would be useful for researchers doing family and marriage studies.

(4) Information on education history. The designs of the 2010 and 2011 CFPSs only allowed interviewers to collect information on the education experiences of adults over 16. Children under 16were asked only about their highest educational level (for those who were not at school) or the current educational level (for those who were at school). Their detailed education information was not acquired. Therefore, we supplemented this part of the information on education by interviewing the respondents who did not answer questions on their education experiences in our surveys in 2010 and 2011. Moreover, we reconfirmed the current state and level of education and those in 2010 in the CFPS 2012 questionnaires.

(5) Questions on the education level concerning when respondents left school for those whohad done so. The question on educational level in the CFPS 2010 questionnaire was "What is the highest level education you have completed?" This question actually did not cover the last educational level for those who had quit school without graduating, which led to underestimations of their educational levels. Thus, the question was modified to "What educational level were you in when you left school?" in our 2012 questionnaire for all respondents who had left school before their graduation, in order to make an accurate estimation of their education levels.

(6) The information on the respondents' parents, including their dates of birth, their occupations, education levels and political statuses. In the CFPS 2010 field survey, we did not collect such information on parents who had died. Their information was collected in 2012 to make up for the data in Table T3.

[66] For the person who has left a marriage and has had no remarriages, no values were given to the corresponding child in terms of the child's bio variables.

After finishing the 2012 survey, we updated the data from CFPS 2010 based on information collected above. We thus remind data users to refer to the updated data of CFPS 2010 and the 2012 survey data to adjust and supplement data sets if needed.

7. Composite Variables

7.1 Education Level (2010)

Education level is a commonly used variable in social science research. In order to minimize missing data, CFPS collects information on education level from different sources: (1) self-reported education levels; (2) education level reported by other family members; For example, in the family roster questionnaire we ask the respondent to report the education level of all the family members; in the child questionnaire we ask the guardian to report the education for any child under age 10; in the individual questionnaire we ask the respondent to report the education level of his/her spouse; and we ask family members to report for those who cannot be interviewed. (3) changes in education levels in subsequent waves, which could be used to further interpolate and impute the missing values in previous surveys; We combine the information from the above resources and then evaluate, impute and adjust the education levels of the respondents. Based on our results, we generate a composite education variable for the convenience of the data users.

For the education levels of individuals completing baseline individual questionnaires, we provide three composite variables: Highest educational degree in CFPS2010, stage of education in CFPS2010[67] and years of education in CFPS2010 (see Table. 28). To be specific, these three variables combine information from proxy reports by family members in the 2010 individual questionnaire, self-reports from 2010 individual questionnaires, and backward imputed education levels from self-reported education levels in 2012 individual questionnaires (referred to as "backward imputed values"). Among these, the self-reported value in 2010 and backward imputed value in 2012 are composite variables based on information from various sources. Figure 16 shows the detailed information used when generating the composite variable of education level in 2010.

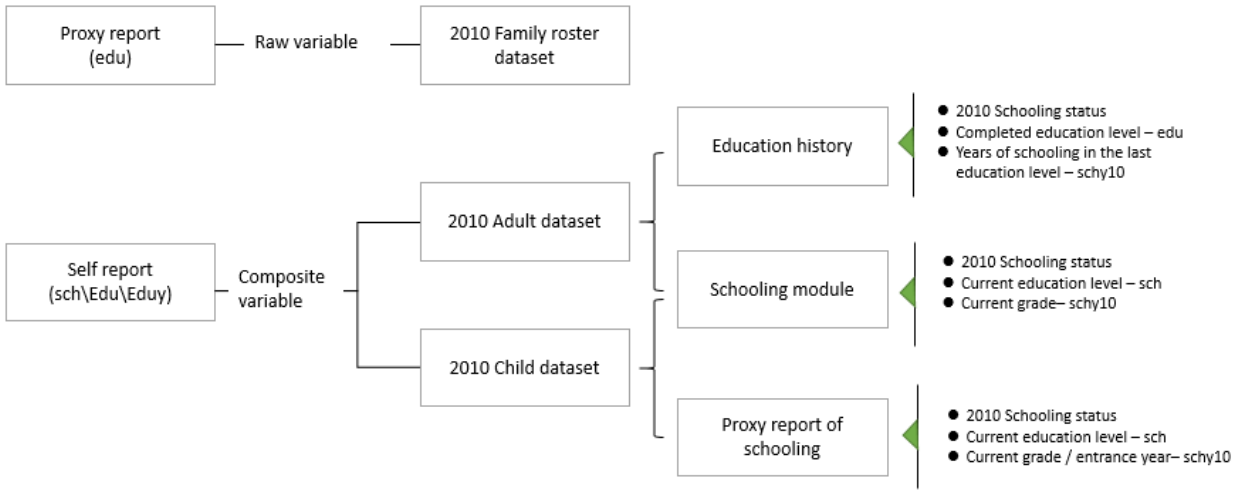
[67] This variable records the current education levels for the respondents who were then at school. For those who had left school, this variable records the education levels when they left school.

Table 28. List of best variables of education level in CFPS 2010

Variable name	Variable label	Type of variable
cfps2010edu_best	Highest educational degree in CFPS2010	Categorical
cfps2010sch_best	Stage of education in CFPS2010	Categorical
cfps2010eduy_best	Years of education in CFPS2010	Continuous

Of these three composite variables, the variable of stage of education in CFPS2010 counts the last incomplete education level for those who quit/dropped out in the middle of an education level, which is taken into account when computing years of schooling. For the respondent who was currently in school/quit in the middle of an education level/dropped out of school, we add the years of schooling in the last education level to the years of education calculated from the highest education degree.

More information about the generation process of the three composite education variables, preliminary statistical results, and data evaluation can be found in *Technical Report: CFPS-21*.



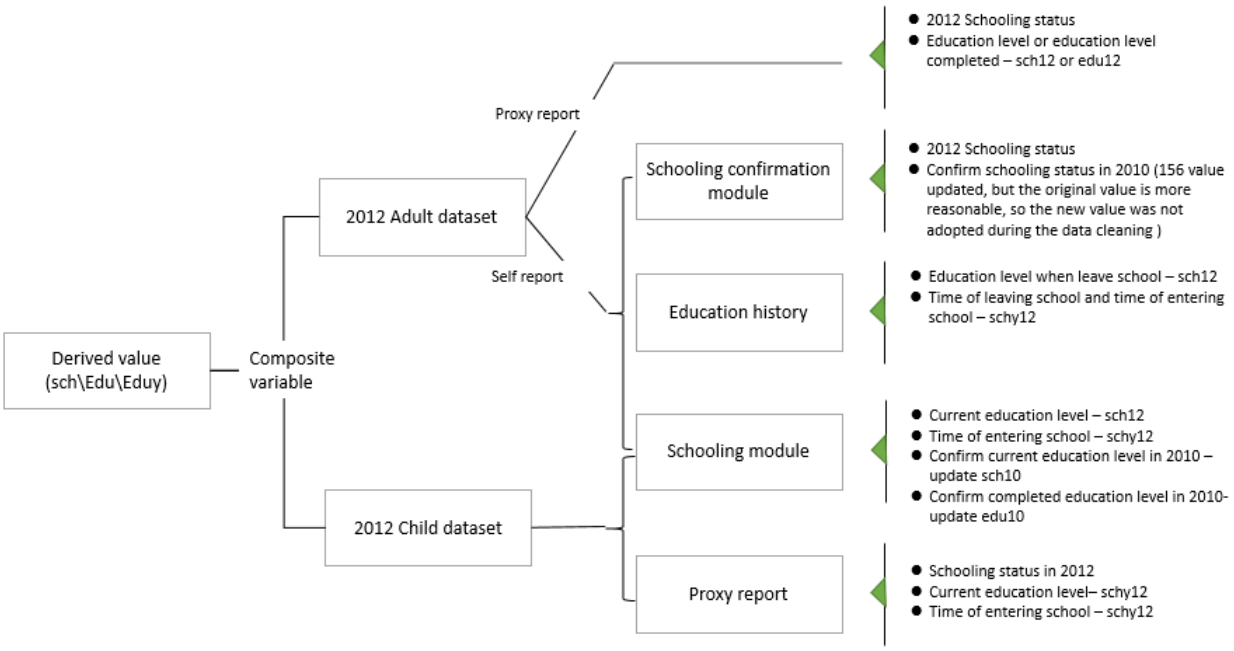


Figure 16. Sources of information for composite education variables

Table 29. Conversion table for years of schooling

	Highest education level attained	Years of schooling
1	Illiterate/ semi-illiterate	0
2	Primary school	6
3	Middle school	9
4	High school	12
5	2 or 3 year college	15
6	Bachelor's degree	16
7	Master's degree	19
8	Doctoral degree	22

Similarly, we created variables on the education levels and the years of education for fathers, mothers, and spouses. See Table 19 for the variable names and labels. For the detailed methods of creating the composite variables on educational level and preliminary results of statistical analysis, see *Composite Variables (II): Educational Level and Depression Scale* (CFPS-12), *Evaluation and Analysis of Differences in Self-reporting and Proxy on Education Level* (CFPS-21), and *Composite Variables (IV): Parents' Social Status* (CFPS-22)

Table 19. Education Level and Years of Education for Father, Mother, and Spouse

Name	Label	Name	Label	Name	Label
fedu	Father's education level (crude)	medu	Mother's education level (crude)	sedu	Spouse's education level (crude)
feduL	Father's years of education (crude)	meduL	Mother's years of education (crude)	seduL	Spouse's years of education (crude)
feduc	Father's education level (detailed)	meduc	Mother's education level (detailed)	seduc	Spouse's education level (detailed)
feduy	Father's years of education (detailed)	meduy	Mother's years of education (detailed)	seduy	Spouse's years of education (detailed)

7.2 Depression (2010)

CFPS 2010 applied the same scale to measure the mental state of individuals in the adult and child questionnaires (Table 30). We performed factor analysis and used the factor score to indicate the level of depression, which is named "fdepression." The variable "depressionf" is the additive index score based on the 6 questions. For detailed information on these two composite variables, see *Composite Variables (II): Educational Level and Depression Scale* (CFPS-12).

Table 30. The Depression Scale of CFPS 2010

Adults Question No.	Children Question No.	Items
Q6	N4	Please select the following statements about the mental state according to your situation in the recent month:[68] 1. Almost daily 2. Often 3. Half of the time 4. Sometimes 5. Never
Q601	N401	Feel depressed and cannot cheer up
Q602	N402	Feel nervous
Q603	N403	Feel agitated or upset and cannot remain calm
Q604	N404	Feel hopeless about the future
Q605	N405	Feel that everything is difficult
Q606	N406	Think life is meaningless

7.3 Cognitive Ability

The 2010 CFPS baseline survey applied the word test and math test to assess (Set A as introduced in a previous section) and evaluate the cognitive ability of all respondents who needed to complete the individual questionnaires by themselves (i.e., children aged 10-15 and all adults).

The verbal problems (X2 in the self-answered child questionnaires, X1 in the adult questionnaires) had 8 groups of problems of approximately the same difficulty level. In each group, there were 34 Chinese characters, listed from the easiest to the hardest. In the interviews, the computer would randomly select one group for each respondent to answer. According to different education levels in the T1 table, the respondents started to answer the verbal questions at different starting points. Those with "1-2" (primary school and below) started from the first character; those with "3" (junior middle school) started from the ninth; and those with "4-8" (senior middle school and above) started from the 21st. The characters were presented in cards, and the respondents were asked to read them aloud. If they could not read 3 characters in a row or the questions went to the 34th character, the test ended.

The mathematical problems included addition, subtraction, multiplication, division, exponents, logarithms, trigonometric functions, sequence, permutation and combination, etc. All the problems were divided into four groups at a similar difficulty level. The computer would randomly select one group for the respondents to answer. Each group had 24 problems, also listed from easiest to hardest. As with the verbal problems, the starting points were based on the education level in the T1 table: those with "1-2" (primary school and below) started from the first; those with "3" (junior middle school) started from the 13th; and those with "4-8" (senior middle school and above) started from the 19th.

[68] Repsonses options in the 2010 child questionnaire are slightly different. They are "1. Almost every day 2. 2-3 Times a week 3. 2-3 Times a month 4. Once a month 5. Never"".

Considering that these two groups of problems might be re-used in follow-up surveys, we did not release the problems to the public for consideration of the performance for future tests. We calculated the scores for these two groups of problems, designated as "wordtest" and "mathtest" in CFPS 2010 baseline data, respectively. We assigned the scores according to the question number of the most difficult problem that the respondent had answered correctly. If the respondent did not give any correct answers, the question number of the problem prior to where he/she started would be the final score. For the detailed methods of creating these variables and the results of preliminary statistical analysis of these variables, see *Composite Variables (I): Verbal and Mathematical Tests* (CFPS-11)

CFPS 2014 continued to use the word test and math test, but adjusted the three-stage fixed starting point strategy. The starting point was formerly fixed at a certain point depending on the education level at baseline. In CFPS 2014, if the answer to the first question was incorrect, we adjusted the starting point to a lower level until we reached the lowest level. Accordingly, we constructed two types of composite variables in the CFPS 2014 individual data set: wordtest 14 and mathtest 14 assumed fixed starting points and thus were comparable with the 2010 variables; wordtest14_sc2 and mathtest14_sc2 do not assume fixed starting points. When using these two variables, note that they are not standardized and not adjusted to age or education level. Users may need to further process these variables according to their specific research needs.

CFPS used cognitive Set B (word recall and number series test) in 2012 and 2016. Both measures came from the HRS in the US. During the word recall, interviewers read 10 words (e.g. mountain, rice, river, etc.) out to the interviewees and asked them to recall the words immediately after the reading (i.e., immediate recall) and also again after a few minutes (i.e., delayed recall). Interviewees would be given a second chance to recall if they failed to recall any word. Word recall is scored as the total number of correctly recalled words, regardless of the sequence. In CFPS 2012, IWR1, IWR2, and IWR represent the scores from the first attempt, second attempt and the total in the immediate word recall, and DWR represents the scores from delayed word recall.

Number series is a two-stage adaptive test. In the first stage of the routing test, respondents are presented with three items and are scored from 0 to 3 based on number of correctly answered items. In the second stage, the system assigns a set of items among four possible forms based on their scores from the first stage. The four possible forms vary in their difficulty levels, and those with better performances receive more difficult forms . The adaptive design is based on modern test theory, with the aim of obtaining a more accurate measure of the respondent's cognitive ability within a shorter amount of time. The traditional observed scores obtained by counting the number of correctly answered items are not applicable here because respondents are taking different forms of varying difficulyt. Individuals taking adaptive tests are often scored based on Item Response Theory models. Under IRT, every item has its own item characteristics such as item difficulty and item discrimination, and such item parameters are invariant to the particular sample that take the test. Every respondent has his or her own ability scores, independent of the actual items that he/she takes. An important application of IRT is that respondents' scores are unaffected by the actual form taken by the respondent, thus making it suitable for our purposes. Due to the complexity of this scoring method, we provide scores based on the Rasch models (one type of IRT models) in our public use datasets. The variable name for the Rasch-based score of

the number series item is NS_W. We also provide the standard error of this score with the variable name NS_WSE.

During the data cleaning process, we observed a large proportion of nonresponse in the number series test. This was mainly because the test would not start unless the respondent acknowledged that he or she understood the two examples of number series problems. We also discovered a typo in one item. Since this item was the last item in the second stage of the self-adaptive test, we imputed the score before computing the composite score. In the 2016 survey, we corrected the typo. For more information see *Technical Report: CFPS-31*.

7.4 Income

We have already introduced the designs for income data collection in CFPS in Section 3.6 of this manual. In the 2010 CFPS baseline survey, with rural families, we asked only about the income from selling agricultural products in farming, forestation, herding, side-line production, and fishing; we did not ask about the values of the products that were consumed by the rural families themselves. As the rural families in China would consume a considerable portion of agricultural products themselves, the data collected in CFPS 2010 cannot give us an accurate estimation of the total income from these rural families' agricultural production. The omission of the value of self-consumed products causes underestimation of the actual income from agricultural production, especially in less commercialized regions and rural families in poverty.

To correct for this, we designed an adjustment scheme for agricultural income of rural families based on information obtained in the questionnaire, and generated two variables, "inc_agri" and "net_agri," which represent the adjusted total income and the adjusted net income from the families' agricultural production, respectively. The adjustments are mainly based on information about the families' production output, sales volume, sales income, and net income (K6 and K7 in the CFPS 2010 family questionnaires). We calculated the number of products for family consumption by comparing the total production output and the total sales volume. We then converted this portion to income according to market prices. The total and net incomes of the family's agricultural production were imputed by adding this converted amount of income to the total and net sales incomes. For detailed information on the adjustments of agricultural production incomes, see *Adjustments of Rural Family Income* (CFPS-14).

Table 31 compares the means, standard deviations, and medians of the net income of agricultural production before and after the adjustment.[69] Adjustments were done for all families working on agricultural production in 2009. It is notable that the mean of the net income per household increased by 2,469.7 RMB and the median increased by 2,275 RMB after modification.

[69] Technical report: CFPS-14.

Table 31. Net Income of Family Agricultural Production, before and after Adjustment

	Mean (RMB)	Standard Deviation	Median	N of Household
Original	4,889.0	12,796.6	2,105.0	7,586[70]
Modified	7,358.7	14,044.3	4,380.0	7,586

Due to the complexity of economic conditions, there were many questions on personal and family income in the 2010 CFPS baseline survey. For data users' convenience, we created a group of composite variables based on personal and family incomes. The relevant variables and labels are listed below in Table 32. The calculations of the variables listed in Table 23 are as follows:

- Personal income: We first used the personal income from the questionnaires as the values; if the information was missing, we used the mean of the income zone instead; if this mean was still missing or less than 100, we replaced it with the total income calculated by all the items.[71]
- Total income: the income with the cost of agricultural production included.
- Net income: the income after deducting the cost of agricultural production.
- Total income or net income per capita: the average income with the total income or net income divided by the family size. Here the family size is the number of family members living together, as shown in the T1 table.
- Total/net family income: the total calculation of the five parts of income, i.e., wage income, total/net business income, property income, transfer income, and other income.
- Wage income: wages, awards, allowances, income from working out of town, and bonuses for the individual.
- Business income: agricultural income (income from farming, forestation, herding, side-line production and fishing) and non-agricultural income (for other non-agricultural production).
- Property income: rent of land or other means of production, house rent, other rent income, income from selling properties. As for financial property income such as interest on savings, stocks, funds, and bonds, since the CFPS only asked respondents about the savings the families held and their corresponding market values at the end of 2009 but the income might be consumed long before the surveys, we excluded this part of income from property income for the current year.
- Transfer income: government allowances, pensions, and other economic aids from the government, such as basic living allowances.
- Other income: gifts from relatives and friends, and other income claimed by the respondent's family.

[70] There were 7,798 households that reported they had worked in agricultural production in CFPS 2010 family samples. Here the 7,586 households were the cases that had no missing values both before and after adjustments. This sample includes urban families working in agricultural production.

[71] The items calculated in the total income included wages, awards, annual bonus, welfare items, income from second jobs or part-time jobs, income from other work, pensions, self-employed income, economic aid from relatives and friends, economic aid from villagers'/neighborhood committees, allowances from the government or workplaces.

Table 32. Composite Variables on Personal and Family Incomes

Name	Label	Name	Label
Income	Personal income	faminc_net_old	Unadjusted net family income
firm	Non-agricultural business income	faminc_old	Unadjusted total family income
finc	Wage income	faminc	Adjusted total family income
fproperty	Property income	faminc_net	Adjusted net family income
welfare	Transfer income	indinc	Adjusted total family income per capita
felse	Other income	indinc_net	Adjusted net family income per capita
foperate	Adjusted total business income (including agricultural and non-agr business)	foperate_net	Adjusted net business income (including agricultural and non-agr business)

In Section 3.6, we mentioned four adjustments in the CFPS 2012 income module, including adding income items that were omitted in 2010, refining the main categories, adding unfolding brackets questions to compensate for the missing values, and moving the questions on individual income from the family questionnaire to the individual questionnaire. The first three adjustments undoubtedly improved the quality of data on income in CFPS 2012. However, the last adjustment had some drawbacks. Usually, self-reported wage income would be more accurate than the number reported by a proxy, which is why we made this adjustment. However, only if the presumption that every family member who received wage income answered the individual questionnaire and reported his/her wage income is true can we get accurate family wage income. In practice, it is hard to avoid non-response and missing values. When we were cleaning the data on CFPS 2012 family wage income, we found that some family members were employed by an organization but did not report their wage income. We also found that some non-coresident family members missed the individual questionnaire. All these non-responses cause underestimation problems in family wage income. As wage income is the most important component of household income both in urban and rural areas, these missing values would greatly underestimate the total income of the household. We thus adjusted and imputed the wage income in 2012. More detailed information can be found in *Technical Report: CFPS-27*.

CFPS 2012 classified 5 main income categories by income source: wage income, business income, transfer income, property income and other income. As mentioned above, CFPS 2012 had better coverage of income items (see Table 6 in Section 3.6). We updated the definition of these five income categories based on the income items in 2012. Wage income includes the post-taxation wage, bonus and non-cash benefits from agricultural or non-agricultural employed work. Business income is the net income from family production in farming, forestry, pasturing, fishing and sideline, the value of agricultural production consumed by the family, and net profits from self-employment or operating private enterprises. Transfer income includes various kinds of transfer payments from the government (e.g., pensions, subsidies and alms) and social donations.

Property income is the income from renting land, real estate and other means of production. Other income includes financial support from families or friends as well as gifts and cash gifts.

Meanwhile, the difference in the coverage of income items between the 2012 and 2010 questionnaires raised a problem of comparability between the two waves. Given the importance of having comparable data across waves, we compared the income items in these two questionnaires in detail and generated a set of income variables in 2012 that were comparable to those in 2010. To be specific, we (1) eliminated the items that were not covered in 2010 or were not comparable due to differences in wording from the household income section in 2012. These items included wage income from agriculture-related employment, income from internships and part-time work during study, fellowships and scholarships, and compensation or housing demolition/relocation; (2) eliminated the non-agricultural self-operated income from 2012, because in 2010 we asked only about private enterprise but not about self-employed business, while in 2012 these two items were integrated into one question. We can see this by comparing the income items in 2010 and 2012 in Table 6. Therefore, there are two versions of the composite household income variable in 2012, one being the total income from all the items in the 2012 family questionnaire, and the other being the total income variable that is comparable with that in 2010. We recommend the former when using the 2012 cross-sectional data, and the latter for longitudinal analysis.

See Table 33 for the names and labels of the composite variables of individual[72] and household income in CFPS 2012. The variables with the suffix "_1" are calculated using complete household income information, that is, with all the income items in Table 6. The variables with the suffix "_2" are household income variables that are comparable to those in 2010, containing only the income items that are consistent with 2010. In addition, variables with the suffix "_adj" are adjusted household income. We also kept the income variables from the original data. Users are welcome to choose whichever variables they want to use. See the following notes for detailed adjustment for each variable.

(1) Household wage income wage_1_adj, wage_2_adj: impute the values for respondents with employed jobs but the zero value of reported wage income and non-coresident family members who failed to complete an individual questionnaire.
(2) Net family income fincome1_adj, fincome2_adj: impute the values for respondents with employed job but the zero value of reported wage income, and non-coresident family members who failed to complete an individual questionnaire. Substitute the net family income with family expenditure in cases of 0 or missing family net income.
(3) Family income per capita fincome1_per_adj, fincome2_per_adj: adjusted net family income divided by family size.
(4) Percentile of family income per capita fincperadj_p: calculated based on adjusted family net income per capita.

[72] Individual income was calculated in the same way in CFPS2012 as in CFPS 2010. We first used self-reported individual income as the value of individual income. If that was missing, we replaced the value with the average of income ranges from the unfolding brackets. If it was still missing, we replaced the value with the sum of all income items.

Table 33. Composite income variable names and labels in CFPS 2012 and 2014

Variable name	Variable label	Source
Income	Individual income	2012/2014
income_adj	Individual income (adjusted)	2012/2014
wage_1	Wage income	2012
wage_2	Wage income (comparable to 2010)	2012
fwage_1	Wage income	2014
fwage_2	Wage income (comparable to 2010)	2014
wage_1_adj	Wage income - adjusted	2012
wage_2_adj	Wage income - adjusted (comparable to 2010)	2012
foperate_1	Business income	2012/2014
foperate_2	Business income (comparable to 2010)	2012/2014
ftransfer_1	Transfer income	2012/2014
ftransfer_2	Transfer income (comparable to 2010)	2012/2014
fproperty_1	Asset income	2012/2014
fproperty_2	Asset income (comparable to 2010)	2012/2014
felse_1	Other income	2012/2014
felse_2	Other income (comparable to 2010)	2012/2014
fincome1	Net total family income	2012/2014
fincome2	Net family income (comparable to 2010)	2012/2014
fincome1_adj	Net total family income - adjusted	2012
fincome2_adj	Net family income – adjusted (comparable to 2010)	2012
fincome1_per	Net family income per capita	2012/2014
fincome2_per	Net family income per capita (comparable to 2010)	2012/2014
fincome1_per_adj	2011-2012 Net family income per capita - adjusted	2012
fincome2_per_adj	2011-2012 Net family income per capita – adjusted (comparable to 2010)	2012
fincper_p	Family income per capita percentile	2012
fincome1_per_p	Net family income per capita percentile	2014
fincperadj_p	Family income per capita percentile - adjusted	2012

CFPS 2014 adopted the structure of the 2012 questionnaire in the modules of business income, transfer income, property income and other income, so that these income variables are comparable between these two waves. As for family wage income, however, we know from the above that in 2012, the family wage income was underestimated because it did not cover the wage income from family members who did not complete the individual questionnaire or the specific question due to unavoidable non-responses. To deal with this problem, we put the question on family wage income back into the family questionnaire, which should be answered by the respondent of the family questionnaire. Notice that for rural samples, the family wage income mainly came from family members who did outside work, and thus the respondents of

the family questionnaire might not be fully aware of the exact income. They might give an estimation based on the money sent back from the worker, which resulted in potential underestimation of wage income. Therefore, in data cleaning, we imputed the total wage income from individual questionnaires based on the reported total family wage income from the family questionnaire in cases of missing values, 0 values or when the total wage income from the individual questionnaires was greater than the reported total family wage income from the family questionnaire.

The income items of 2012 and 2014 are comparable with one another, but not with the income items in 2010. Considering this, we generated a set of 2014 family income variables that are comparable with those in 2010. Similarly, we suggest using the complete family income variables when using the 2014 cross-sectional data or the 2012 and 2014 data, and using the comparable variables when the user needs to do a comparison analysis with income in 2010. See Table 33 for the names and labels of the composite income in CFPS 2014.

7.5 Family Expenditure

CFPS collects information on family expenditure in the family questionnaire. We introduced the detailed expenditure items in Section 3.6 of this manual (See Table 7 in Section 3.6). In general, there were four types of expenditures: (1) consumption expenditure (nonproductive expenditures): daily expenditure on food, clothing, housing, household appliance and daily used commodities and necessities, transportation and communication, entertainment and education, medical care and other consumption expenditures; (2) transfer expenditure: financial support to family and non-co-residing family members or friends, social donations and giftss and cash gifts in major family events; (3) insurance expenditure: expenditure on commercial insurance; (4) housing expenditure, including mortgage payment. Detailed expenditure items of each type are listed in Table 34.

Table 34. Family expenditure items in CFPS 2010/2012/2014

Family expenditure items[73]	**CFPS 2010**	**CFPS 2012**	**CFPS 2014**
Expenditure on Consumption			
1.Food	Expenditure on food	Expenditure on cigarettes, beverage and alcohol self-consumed by own family	
	Value of food self-consumed by own	Food self-consumed by own family	Food self-consumed by own family

[73] Cost of production and business activities in Table 7 do not count ias family expenditure.

	family (imputed)		
		Value of agricultural products self-consumed by own family	Value of agricultural products self-consumed by own family
2.Clothing	Expenditure on clothing	Expenditure on clothing	Expenditure on clothing
3.Housing	Housing rent (exclude housing mortgage)	Housing rent	Housing rent
	Expenditure on housing (e.g. property management fee, expenditure on heating etc. exclude housing mortgage and rent)	Expenditure on water and electricity	Expenditure on water and electricity
		Expenditure on fuels	Expenditure on fuels
		Payment for heating system	Payment for heating system
		Payment for estate services (including parking)	Payment for estate services (including parking)
4.Household appliance and commodities		Pasyment for automobile	Payment for automobile
		Expenditure on purchase and maintenance of vehicles except car	Expenditure on purchase and maintenance of vehicles except car
		Expenditure on electrical appliances for work	
	Expenditure on household appliances	Expenditure on furniture and other durable goods	Expenditure on furniture and other durable goods
	Expenditure on daily used commodities	Expenditure on daily used commodities and necessities	Expenditure on daily used commodities and necessities
	Expenditure on other goods and services	Expenditure on hiring domestic helper or hourly worker	

5.Communication and transportation	Expenditure on communication	Expenditure on communication	Expenditure on communication
	Expenditure on transportation (including vehicle maintenance)	Expenditure on local transportation (including petrol fee)	Expenditure on local transportation (including petrol fee)
6.Cultural recreation and entertainment	Expenditure on education	Expenditure on education	Expenditure on education
	Expenditure on culture/entertainment/ leisure activities	Expenditure on culture/entertainment/ leisure activities	Expenditure on culture/ entertainment/leisure activities
		Travel expenditures	Travel expenditures
7.Medical care	Expenditure on medical care	Direct medical expenditure	Direct medical expenditure
		Expenditure on health care goods	Expenditure on health care goods
8.Others	Expenditure on marriages and funerals of family members		
	Other expenditure	Other living expenditure	Other living expenditure
Transfer expenditure	Total value of donation to institution or person in cash and in kind	Tax and fees paid to the government	
		Donations (in cash and in kind)	Social donations (in cash and in kind)
		Financial support to non-co-residing family members	Financial support to non-co-residing family members
			Financial support to other family memberss
			Cash and non-cash gift for important events (e.g. marriage, birth, getting into higher education)
Insurance expenditure	Expenditureon commercial insurance	Expenditure on commercial medical insurance	Expenditure on commercial medical insurance
		Commercial asset	Commercial asset

		insurance expenditure (including car insurance) Payment for various kinds of pension	insurance expenditure
Housing purchase and construction mortgage	Housing mortgage	Housing mortgage (imputed)	Housing mortgage

We generated a composite variable for family expenditure for our users. The variable is the sum of the four types of expenditure. If there was no expenditure of one type, we recorded it as zero. CFPS 2014 asked the respondent to report the range of total family expenditure in the last 12 months besides asking for the expenditure of each type. When we generated the composite variable of total family expenditure, we mainly relied on the sum of the expenditure on each type, and only imputed the value using the reported total family expenditure when the sum of each expenditure was less than 100 or missing (all the relevant answers were do not apply/ refuse to answer/ do not know).

CFPS used a different recalling time range to collect the expenditure information, as different types of expenditures occurred in different frequencies. There were three types of recall time range: past week, past month and past 12 months. The total expenditure was measured in 12 months. If one type of expenditure was reported by week, we converted it to expenditure in 52 weeks[74] (weekly expenditure times 52 weeks). If the expenditure was measured monthly, we converted it to expenditure in 12 months (monthly expenditure times 12 months). A few respondents used the wrong time range by mistake when they switched back and forth from one time range to the other; for example, they reported the expenditure of the past 12 months when the time range should be 1 month.For this reason, when constructing the composite variables for expenditure by type, we screened the outliers in expenditure by comparing the percentile of income with the percentile of expenditure and adjusted the outliers.[75]

Although the changes in the specific items and the forms of questions in the CFPS expenditure module across different rounds affected the precision of measurement to some extent, the main categories of expenditure remained consistent in general. The total consumption expenditure and total expenditure are generally comparable across rounds. See Table 35 for the names and labels of composite variables related to expenditure in CFPS 2010/2012/2014.

Table 35. Names and labels of the composite variable on expenditure in CFPS 2010/2012/2014

Variable name	Variable label
Pce	Residents' consumption expenditure: Sum

[74] One year has 52 weeks.

[75] For example, when the total expenditure in the last month on each type of item reported by a family exceeded 12 times the average expenditure of the families in the same income percentile, we infer that the family reported its income in the time frame of last 12 months. We adjust this type of expenditure by dividing it by 12.

Food	Expenditure on food: Adjusted(*yuan*)
Dress	Expenditure on clothing
House	Expenditure on housing: Adjusted(*yuan*)
Daily	Expenditure on family equipment and daily necessities: Adjusted
Med	Medical and fitness expenditure
Trco	Expenditure on communication and transportation: Adjusted
Eec	Expenditure on education and entertainment(*yuan*)
Other	Other expenditure on consumption
Eptran	Transfer expenditure(*yuan*)
Epwelf	Welfare expenditure(*yuan*)
Mortage	Mortgage on housing
Expense	Total family expenditure(*yuan*)

7.6 Family Assets

In the CFPS 2010 and 2012 family questionnaires, the variable "total_asset" indicates the net family asset value, which was the difference between family total assets and total liabilities. Family assets include land, housing, financial assets, productive fixed assets and durable goods. Family liabilities include housing liabilities and non-housing liabilities. The value of land was estimated, for example, assuming that 25% of the family agricultural income comes from land and the return rate of land is 8%, and we could estimate the value of land (Mckinley, 1993). The housing property includes current residence and other housing. When calculating the value of house property, we counted a house with partial property rights as full property rights since we were not informed of the proportion of the property rights and a household has perpetuity. Financial assets include deposits, stocks, funds, bonds, financial derivatives, other financial products and borrowings. The data in 2010 did not contain the value of bonds, financial derivatives and other financial products. Productive fixed assets include productive firm assets, agricultural machinery and so on. Durable goods include automobiles, televisions, computers, refrigerators and other common household appliances. Housing liabilities is the number reported when answering the question about "housing debt with interest." Non-housing liabilities counts debt from education or medical care. Table 36 shows the published variables about assets. More information about the variables and data cleaning could be found in Technical Report: CFPS-29.

Table 36. Composite variables on assets and their labels in CFPS 2010/2012

Variable name	**Variable label**	**Source**
land_asset	Value of lands (*yuan*)	2012/2010
houseasset_gross	Gross house asset(mortgage not deducted) (*yuan*)	2012
resivalue_new	Market value of your current house:Final (*yuan*)	2012/2010
houseprice1_best	Current market value of your house:Best	2012
otherhousevalue	Total value of all other houses:Final (*yuan*)	2012/2010

houseprice2_a_1_best~ houseprice2_a_6_best	Total current market value of the Nth close to family: Best (*yuan*)	2012
house_debts	Total amount of mortgage for all houses (*yuan*)	2012/2010
house1_debts	Total amount of mortgage for current residence (*yuan*)	2012
houseother_debts	Total amount of mortgage for other houses (*yuan*)	2012
fixed_asset	Assets for production (*yuan*)	2012
company	Company assets (*yuan*)	2012/2010
agrimachine	Total current value of farm machineries (*yuan*)	2012
finance_asset	Finance asset (*yuan*)	2012
savings	Total amount of cash & deposits:Final (*yuan*)	2012/2010
govbond	Government bonds (*yuan*)	2012
stock	Stock (*yuan*)	2012/2010
funds	Funds (*yuan*)	2012/2010
derivative	Financial derivatives (*yuan*)	2012
otherfinance	Other financial products (*yuan*)	2012
debit_other	Money lent out to others (*yuan*)	2012/2010
nonhousing_debts	Financial debt(except house mortgage) (*yuan*)	2012/2010
bank_debts	Total amount of bank loans(except house mortgage) (*yuan*)	2012
ind_debts	Total amount of loans in debt to other parties than financial institutions (*yuan*)	2012
durables_asset	Expenditure on durable goods (*yuan*)	2012
total_asset	Net family assets (*yuan*)	2012
valuable	Market value of valuable collections (*yuan*)	2010
otherasset	Other assets (*yuan*)	2010

7.7 Occupation Codes

Two schemes were applied in the 2010 CFPS baseline survey to code the respondents' occupations: field coding by interviewers and post-hoc coding by the data team. Starting from 2012, CFPS no longer use field coding by interviewers. All occupation coding is done by the data team using post-hoc coding based on the information from the interview.[76] Table 37 summarizes all the occupation variables in the occupation coding.

In the baseline survey, the field coding was used in questions G307, G308, H405, and H406. With the help of the CAPI system, the interviewers directly coded the occupations and industries

[76] We compared the results of field coding and the post-hoc coding and found that the latter got higher quality. Therefore, we adopted the post-hoc coding only starting from CFPS 2012. Please refer to CFPS-8 for details.

of the respondents according to the coding dictionary of CFPS occupations and industries.[77] The interviewers went from the broad categories to detailed categories (4-digit codes) of occupations and industries. The interface of the coding system is shown in Figure 17.

Apart from the four questions above, questions on occupations and industries were all open-ended, being coded manually by trained data workers after the survey. In the coding process, we controlled the data quality with Two-way Independent Verification with Adjudication. In the first round, two data workers coded the occupational information of each respondent separately. If their results matched, it was then confirmed as the final code; if inconsistent, a third experienced data worker would code these items in the second round. The second code would be accepted if it was consistent with either of the two in the first round. If not, we would ask expert researchers to make decisions based on the codes by data workers, the field codes, and other auxiliary information.

Table 37. Questions and Auxiliary Items on Occupations and Industries of CFPS 2010

Questionnaire	Question No. (variable name)	Item	Type of Question
Family roster	B5 (tb5)	What is the main occupation of "family member"?	Open-ended
Family roster	D6 (td6)	What is the main occupation of "immediate relatives living separately"?	Open-ended
Adult	B309 (qb309）	What is the main occupation of "sibling"?	Open-ended
Adult	G303 (qg303)	What institution are you currently working at?	Multiple choice
Adult	G304 (qg304)	What is the name of your working place?	Open-ended
Adult	G305 (qg305)	The institution you are working at belongs to?	Multiple choice
Adult	G306 (qg306)	Your occupation?	Open-ended
Adult	G307 (qg307)	Which category does your occupation belong to?	CAPI-assisted field coding
Adult	G308 (qg308)	Which industry does your occupation belong to?	CAPI-assisted field coding
Adult	G601 (qg601)	What is your first occupation? and children?	Open-ended
Adult	G701（qg701）	What is your second occupation?	Open-ended
Adult	H404（qh404）	What is your non-agricultural occupation?	Open-ended
Adult	H405（qh405）	What is the category of your non-agricultural occupation?	CAPI-assisted field coding

[77] The CAPI method is still in the trial stage. Given the complexity of occupation classifications in China, it is difficult to confirm the category of occupations. Therefore, interviewers also recorded detailed occupational information (G306).

Adult	H406（qh406）	What industry does your non-agricultural belong to?	CAPI-assisted field coding
Children	J2（wj2）	What is the primary duty (the one that takes most time) in your formal work?	Open-ended

Figure 17. Computer Interface of the Field Coding System of CFPS 2010[78]

Due to substantial differences between the field codes and the post-hoc codes, we adopted the latter for questions G307, G308, H405, and H406.

The latest version of the national standard classification of occupations GB/T 6565-2009 was not released in the design stage of the baseline survey questionnaires. Therefore, the 2010 survey used a modified version of GB/T 6565-1999, drawing on the classification standards of occupations and industries in the Chinese General Social Survey (CGSS). This scheme classifies occupations into 8 categories, including 595 occupational codes. The classification of industries was adopted from the standards of the National Bureau of Statistics, which contains 20 categories.

In the post-hoc coding, we applied the latest version of national standard classification of occupations, i.e., GB/T 6565-2009. The categorization and ordering were entirely adopted, yet the codes were re-labeled.

For detailed technical methods and quality evaluation of occupation codes, see *Occupation and Industry Codes of CFPS 2010* (CFPS-8).

[78] Technical report: CFPS-18.

As previously stated, CFPS has been collecting information about all the jobs between waves since 2012. In CFPS 2012, if the respondent had participated in multiple types of jobs such as agricultural, non-agricultural and self-employed/private business in the past year, the user had to identify the respondent's main job. Given that the process in the work module was very complicated, the CFPS data processing team generated a composite variable for current main job in 2012. More detailed information could be found in *Technical Report: CFPS-30*.

CFPS 2014 applied different data collection strategies depending on whether the respondent had a full-time job since the last interview. For those without a full time job, CFPS 2014 collected information on possible internships and part-time jobs, and provided industry codes and occupation codes for the main internship/part-time job. For those with a full-time job, CFPS collected information on all the jobs since the last interview, and designed a specific module to collect information on the main job, for which the industry code and occupation code was generated. Information on other jobs besides the main job was collected in the EHC-Job module and no industrial and occupational information was collected for the other jobs.

7.8 Conversion of Occupational Codes

The coding system (GB/T 6565-2009) of the Chinese Standard Classification of Occupations (COSO) was applied in the 2010 CFPS baseline survey. For data users' convenience, we converted the occupational codes to several other schemes and created some composite variables related to occupations.

(1) The corresponding coding system of occupations in CPFS is International Standard Classification of Occupation (ISCO-88). These variables have "_isco" after the original variable names.
(2) There were two sets of socioeconomic indices based on ISCO-88, namely the International Socio-Economic Index of Occupational Status (ISEI) and Treiman's Standard International Occupational Prestige Scale (Treiman's SIOPS). The naming rule of these two sets of variables is to add "_isei" or "_siops" after the original names.
(3) The Erikson and Goldthorpe's Class Categories were applied to the adult data on occupations. The naming rule is to add "egp" after the original names.

Table 38. Examples of the Names of the Occupation Variables

	Father's primary occupation (B5 in questionnaires for family members)	Sibling 1's primary occupation (B309 in adult questionnaire)	Post-hoc codes for respondent's occupation (G307 in adult questionnaire)
CFPS occupation variable	tb5_code_a_f	qb309_occu_1	qg307code
ISCO-88 code	tb5_isco_a_f	qb309_isco_1	qg307isco
ISEI score	tb5_isei_a_f	qb309_isei_1	qg307isei
SIOPS score	tb5_siops_a_f	qb309_siops_1	qg307siops

EGP category	——	——	qg307egp

We converted not only the occupational codes for the respondents but also those for other family members in the family questionnaire and siblings in the adult questionnaire. For detailed conversion rules, see Table 38. In addition, we provided the Stata commands of the conversions from CSCO codes to ISCO88, ISEI, Treiman's SIOPS, and EGP.

For detailed construction methods and content of the composite variables on occupation, see *Conversion of Occupational Codes and Construction of Socioeconomic Indices* (CFPS-10).

7.9 Dialect code

The CFPS baseline and follow-up surveys collected data on the dialects used by the respondents based on the respondents' self reports and interviewer observations of respondents' dialects during interviews. The CFPS six-digit dialect code is based mainly on the *Chinese dialect atlas.* The code reflects the language family, language, supergroup, group and subgroup. The codes here mainly cover the language distribution of the Han ethnic group; therefore, the first two digits are always 11 (the Han language branch in the Sino-Tibetan family). In the four remaining digits, the first digit represents the supergroup (Madarin District, Jin district, Wu district, etc.); the second digit represent the district or group (e.g., Northeastern Madarin, or Binzhou sub-district under the Jin district), the third and the fourth digits represent the sub-districts under the Mandarin district(e.g., the Ji-shen sub district in the Northeastern Mandarin). The coder combined the writing information from the respondent, the district/county of the respondent, and the *Chinese dialect atlas* to code. The specific rules in coding the dialects can be found in the technical report *CFPS dialect codes (CFPS-28).* Table 39 includes questions relating to dialects and their variable names in the 2010, 2012, and 2014 questionnaires. Dialect coding is part of the restricted use data in CFPS, and interested users may apply for its use through further applications.

Table 39. Questions about dialect code and variable names

Year	Questionnaire	Question No. (variable name)	Item	Type of question
2010	Adult	D2（QD2）	What is the primary language used in your daily communication with your family at home?	Choice
2010	Child	K2（WK2）	What is the primary language used in your daily communication with your family at home?	Choice
2010	Common module	S3（KS3）	What is the primary language used in your daily communication with your classmates at school?	Choice
2012	Family roster	Z103	What was the main language used in the interview?	Choice
2012	Family roster	Z104	What was the dialect used in the interview?	Open-ended
2012	Family	Z103（KZ103）	What was the main language used in the interview?	Choice
2012	Family	Z104	What was the dialect used in the interview?	Open-ended
2012		D201（QD201）	What is the primary language used in your daily communication with your family at home?	Choice
2012	Adult	Z103（QZ103）	What was the main language used in the interview?	Choice
2012	Adult	Z104	What was the dialect used in the interview?	Open-ended
2012	Child	Z103（KZ103_B_1）	What was the main language used in the interview?	Choice

2012	Child	Z104	What was the dialect used in the interview?	Open-ended
2012	Child	K2（WK2）	What is the primary language used in your daily communication with your family at home?	Choice
2012	Child	Z103（KZ103_B_2）	What was the main language used in the interview of the child?	Choice
2012	Child	Z104	What was the dialect used in the interview?	Open-ended
2012	Common	S3M（KS3M）	What is the primary language used in your daily communication with your classmates at school?	Choice
2014	Family roster	Z103（KZ103）	What was the main language used in the interview?	Choice
2014	Family roster	Z104（KZ104）	What was the dialect used in the interview?	Open-ended
2014	Family	Z103（FZ103）	What was the main language used in the interview?	Choice
2014	Family	Z104（FZ104）	What was the dialect used in the interview?	Open-ended
2014	Adult	Z103（QZ103）	What was the main language used in the interview?	Choice
2014	Adult	Z104（QZ104）	What was the dialect used in the interview?	Open-ended
2014	Child	Z103（KZ103_B_1）	What was the main language used in the interview?	Choice
2014	Child	Z104（KZ104）	What was the dialect used in the interview?	Open-ended
2014	Child	Z103（KZ103_B_2）	What was the main language used in the interview of the child?	Choice

2014	Child	Z104（KZ104）	What was the dialect used in the interview?	Open-ended
2014	Common module	S3M（KS3M）	What is the primary language used in your daily communication with your classmates at school?	Choice

7.10 Best Variables

In the process of data cleaning, for some variables that needed to be corrected, we created new variables based on the original ones instead of correcting them directly. Both kinds of variables were made available to data users. As the newly created variables have the most reasonable values which take into consideration various information sources, we call them "best variables." The rule in naming these variables was to put "_best" at the end of the names of the original variables. For data users, please note that "best variables," which are the most adequate values based on answers provided, multiple information sources, and logic relations, are not necessarily the correct variables. Researchers can use either the original answers or the "best variables."

In every version of the dataset there may be several "best variables." Apart from the best variables for education and wealth, we will introduce some "best variables" of the survey in CFPS 2010: qa1y_best, qe605y_best, qe606y_best, and qe1_best. These four "best variables" are all in the adult data set.

The variable "qa1y" describes the year of birth of the adults. In the 2010 CFPS baseline survey, we could acquire this information for members in Table T1 from three sources (see Table 40): (1) the information in the questionnaires on family members answered by others; (2) the information provided by the respondents themselves; and (3) spouses' answers after spousal matching, which assumes that the current spouse (or first spouse) answered the individual questionnaire as well.

Table 40. Information Sources of the Year of Birth

	Questionnaire	Question No.	Question
Source 1	Questionnaires for family members	B1	What is the date of birth of "family members?"
Source 2	Adult questionnaires	A1	What is your date of birth?
Source 3	Adult questionnaires	E606/E302/ E211	What is the year of birth of your first spouse/current partner/current spouse?

Through statistical analysis, we found that the information from these three sources was not always consistent, which might lead to inconvenience for the data users. Therefore, we modified this variable both manually and with the help of a computer program. After checking the logic relations between the year of birth and several other life events and referring to the age information in the 2011 survey, we eliminated the unreasonable values and gave the best values of the year of birth, named "qa1y_best." We did not delete the original values with the "best" ones for two reasons. First, although they were the most reasonable values after our careful consideration, we were not sure if they were 100% correct. Second, the age of the respondent was directly related to the type of the questionnaires he/she had answered. Either their answers were correct or they were not; the questionnaires for individuals were automatically created according to the years of birth they claimed. The original values were the evidence for the questionnaires, which is why we must keep them in the data set. Otherwise, the data users might become confused about the rules used in creating the questionnaires.

The variable "qe1" in the questionnaire for adults is about current marital status. We also acquired this information in the questionnaires for family members and there were some inconsistencies between these two information sources. Moreover, we found some logic errors during the data cleaning of the family relations data set mentioned above, e.g., the marital status was "unmarried" while the respondent had a spouse in the T2 table; or the status was "divorced" or "widowed" while there was a living spouse in the T2 table, etc. Some of these errors resulted from wrong answers regarding marital status. In this case, we created the variable "qe1_best" as the most reasonable marital status for data users. Again, we did not directly change the original values, for the marital status would affect the question orders in the marriage module. If the values of the variable "qe1" were changed, it would lead to unnecessary confusion for data users.

In the marriage history module in the CFPS questionnaires, we used a retrospective method to collect respondents' information on the important events in their marriages, enabling researchers to study marriages and their relations to other events. However, this retrospective method as well as certain sensitive questions might pose a challenge to the accuracy of the data collected. As the respondents might not remember things precisely the way they were might deliberately hide some facts, in the data cleaning process, we found many inconsistencies (e.g., a second marriage preceded divorce from the first marriage); inconsistencies between different information sources (e.g., the couple gave different answers on the date of marriage or on the spouse's year of birth); answers that violate common sense (e.g., the age at first marriage was under 16); etc. We chose the most important two variables—"qu605y," year of the first marriage and "qe606y," birth year of first spouse to create a corresponding best variable. After eliminating the errors in the family relations data set in data cleaning, we referred to different information sources, e.g., age interval of proper marriage, birth year of first child, etc., and kept the most reasonable values. Thus, we created the "best variables" "qe605y_best" and "qe606y_best." We preserved the original values with the "best" ones for two reasons. First, we could not guarantee that the "best" values were 100% correct—they were just the most reasonable values, in our view. Data users can select variables based on their own judgments. Second, we believe that some inconsistencies are worth studying, e.g., the different answers on date of marriage given by the spouses, different memories of spouse's year of birth, etc.

For detailed information on the method of creating these four "best variables,"see *Composite Variables (III): "Best Variables" for Age and Marriage* (CFPS-13).

7.11 Confidentiality Issues

Personal information is strictly protected in CFPS. We established the following confidentiality policies:

(1) The respondent's name is not released. Instead, we provide personal IDs for data users.
(2) The exact date of birth is not released. Year and month of birth are available.
(3) We only release the address information at the province level. For the address information at the district/county and village/neighborhood levels, we only offer numerical codes for nominal differentiation.
(4) In question G1 in the adult questionnaire, options 5 and 6 were merged as "others." The question G103 is not released.
(5) The question M706 in the adult questionnaire is not released.
(6) Delayed release of the data in the add-on modules from collaborating organizations.

7.12 Miscellaneous

Table 41 provides a brief summary of some composite variables in the use of data that require special attention but are not covered in the introductions above.

Table 41. Special Instructions on the Use of Selected Variables of CFPS 2010

Content	Questionnaire	Question No. (Var Name)	Brief Instructions
Interactions and relationships between respondents over 60 and their children	Adult	F1-F3(qf1-qf3)	Children's information based on T tables. May contain errors due to matching of family relations. Corrections made in the 2012 survey. Users, see the 2012 data.[79]
Addresses information, e.g., places of birth and *hukou*	Adult	A102、A201 (qa102acode, qa201acode)	Address information at the provincial level available. For the address information at the

[79] Note that although we asked the questions based on the information for the children in the 2010 survey, the questions we asked again were about the specific situations in 2012 rather than recall of the 2010 information.

registration			district/county and village/neighborhood levels, numerical codes provided for nominal differentiation.
Career expectation	Adult	S8, S801	No information collected due to an error in the CAPI.
Career expectation	Child	M601(wm601), D101(wd101)	Coded. See CFPS-9: *Career Expectation Codes*
Language in daily communication	Adult & Child	Adult D2 (qd2), Child K2 (wk2), Share S3 (ks3)	Coded. See CFPS-20: *Dialect Codes*.
Metric, e.g., How far is the nearest senior middle school from your home? ____m/*li*/km.	Family	A6(fa6)	Standardized units available in variable labels.
Coding of multiple choices, e.g., What food have you eaten in the recent month?	Child	L5(wl5)	Instead of 0/1 coding for each option, a set of variables are created (i.e., w15_s_1 to w15_s_9). The value of w15_s_1 records the first option chosen by the respondent, the value of w15_2_2 records the second option chosen, and so on. If the respondent chose 3 options, the values from ws15_s_4 on would all be -8, as with other multiple-choice questions
"Others" options, e.g., What is the	Adult	J101(qj101)	In semi-open-ended questions, verbal information

reason that you are currently unemployed? 77. Others [Please specify]____			specified in option "77" is not released, as with other, semi-open-ended questions.

Composite variable

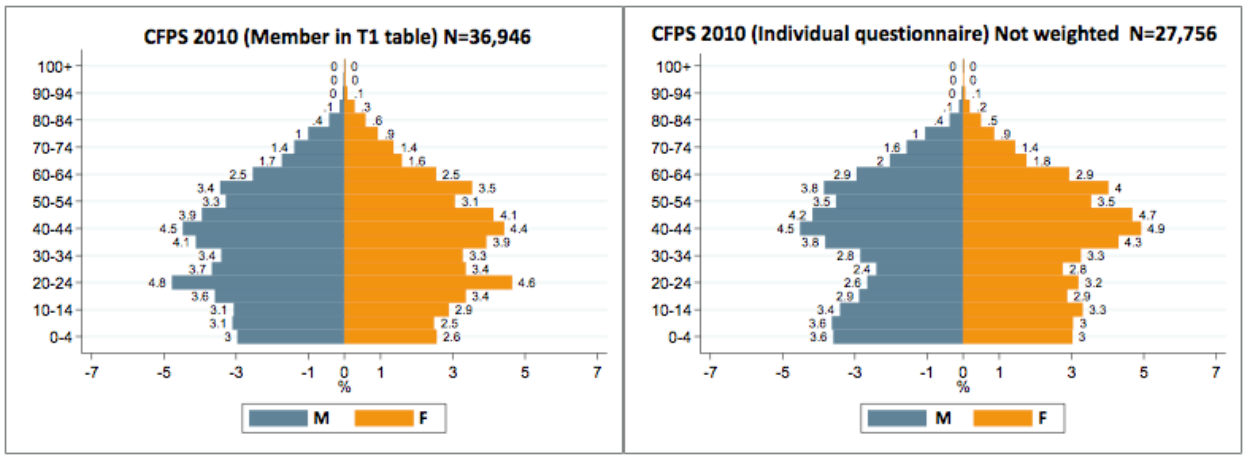
Version	All dataset	releaseversion	Each dataset contains a version variable. Users can view the latest version number on the webpage "CFPS data and document" on the CFPS website to confirm if the dataset used is the latest version.
Family size	Family relationship	familysize	Number of family members (including those living at home and non-coresident family members with economic relationships)
Number of generation	Family relationship	Generation	Number of generation calculated based on the family structure and relationship
If family member	Family relationsihp	co_aXX_p	Dummy variable, indicating whether the individual is a family member in round X (or sharing the kitchen), which is the criteria for identifying a family member.
Background variable about parents	Family relationship Adult Child	fbirthy、feduc、foccupcode、foccupisoc 、fparty、mbirthy、meduc、moccupcode、moccupisoc 、mparty、fbirthy12、feduc12、mbirthy12、meduc12、	Based on information collected in CFPS 2010 and 2012, composite variables are generated that include parents' birth year, highest education attained, two types of main occupation code, political identity in the 2010 dataset, and parents' birth year, highest education attained in 2012 dataset.
Related household ID in family economic questionnaire	Family	overlapfidX，overlapfidXtype	Variable in CFPS 2014 and later rounds. X=1,2,3,4, indicating that the related household has overlapping family members in the family questionnaire, and the type of the overlap. For more

and type of relation			information, see CFPS-34: *Data processing of CFPS 2014 and data introduction*
Urban-rural category	Adult, child	urban urban12 urban14	urban，urban12，urban14 indicates the urban-rural status (according to the definition of National Bureau of Statistics) of the village in CFPS2010, CFPS2012, CFPS2014 respectively. Each variable is included only in the datset of the respective survey year.
Whether there is self-report questionniare	Adult, child	selfrpt	Whether this observation has information from self-report questionniare
Type of self-report questionnaire	Adult, child	self_IWmode	Whether the self-report questionnaire was completed through face-to-face interview or telephone interview
Interrupt status of self report	Adult, child	Interrupt_SF	Whether the self-report questionnaire was interrupted
Whether there is proxy report	Adult, child	proxyrpt	Whether this observation has information from a proxy-report questionniare
Type of proxy report	Adult, child	proxy_IWmode	Whether the proxy-report questionnaire was completed through face-to-face interview or telephone interview
Interrupt status of proxy report	Adult, child	Interrupt_PR	Whether the proxy report questionnaire was interrupted

8. CFPS 2010 Baseline Survey Preliminary Findings and Evaluations[80]

8.1 Age-Sex Distributions

Figure 18 presents the age-sex population pyramids based on the resampled national sample of the CFPS 2010 baseline survey, tabular data of the 2010 national census,[81] and the 2010 Chinese General Social Survey. In the population pyramids, we break ages down to 5-year groups across the range from age 0 to age 100 and above, and present sex-specific relative sizes of each age group. For CFPS, we present the age-sex structures of both the T1 table members (all co-residing household members) and the respondents who have completed the individual questionnaires (adults and children). According to the design, all co-residing family members were target subjects of the survey. However, in the actual surveys, only some of the family members answered the questionnaires due to physical absences or refusals. Therefore, the age-sex pyramid based on all T1 table members reflects the characteristics of the sampled national population, and the one based on all completed individual questionnaires reflects the characteristics of the respondents who completed those questionnaires. Similarly, for CGSS 2010 we also give the age-sex structures of all the family members in the interviewed households and of the actual respondents—the former represents the population, while the latter represents the age-sex characteristics of all survey respondents.



[80] All results in this chapter come from the resampled national sample.
[81] The data comes from Tables of the 2010 Sixth National Population Census "T3-01 National Population by Age and Sex."

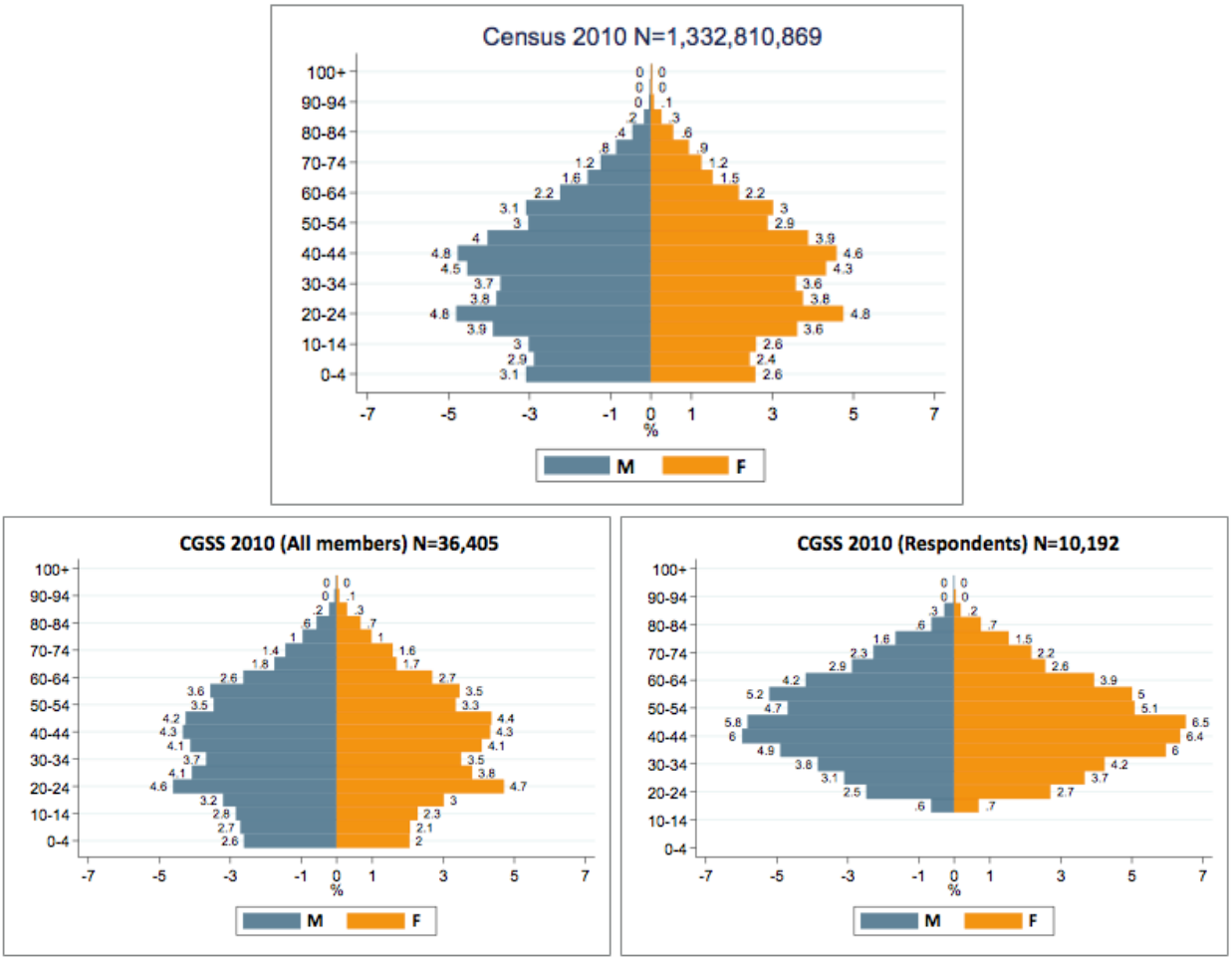
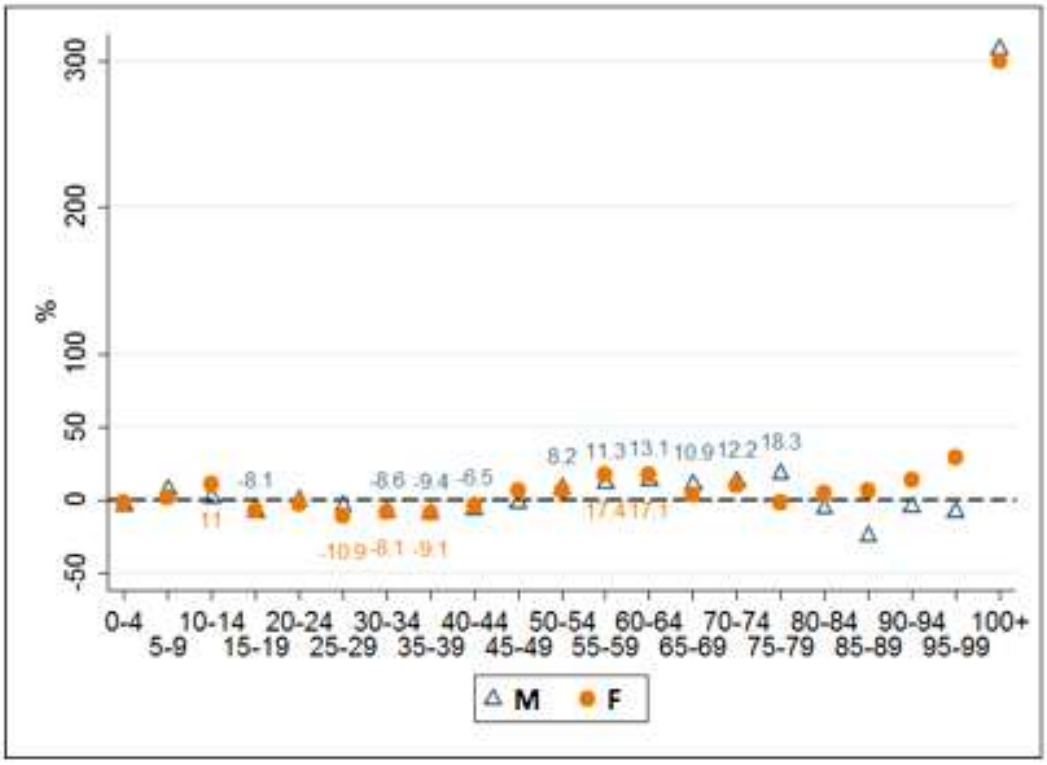


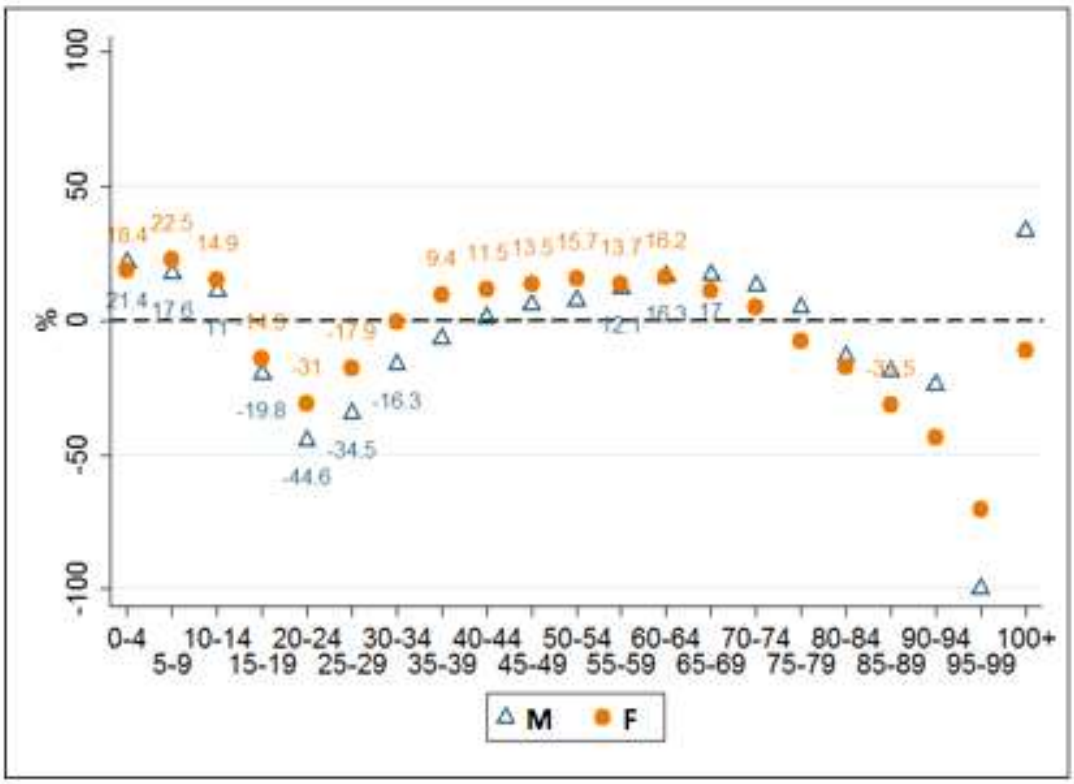
Figure 18. Population Pyramids of CFPS 2010, Census 2010, and CGSS 2010

We compared the age-sex structures of both CFPS T1 members and all members of interviewed households in CGSS with the population structure, as in the 2010 national census. We found that the shapes of the pyramids of these two surveys are rather similar to census data: the 20-24 and 40-44 age groups are the largest, and the young and old groups are the smallest. Specific to each age group, the CFPS data is closer to the census data. We applied log-rate models to test the differences between the two survey data structures and the census population structure, with the population age-sex frequency distribution as the exposure. This method assumed that the census data was the population itself and the data from CFPS and CGSS were probability samples drawn from the population. If the samples' age-sex structures were consistent with the population structure, the sampling probability of each individual should be the same. We included gender, age and their interactions into the full model and compared it to the null model, which produced a Chi-square value for model comparison. The Chi-square value of comparison between the T1 table members of CFPS and the census is 253.70 (*df*=41, p=0.0000), and the Chi-square value for CGSS is 529.85 (*df*=41, p=0.0000). The data compositions of both surveys are significantly different from the population age-sex structure, but CFPS is closer to the census, which has a smaller Chi-square value.



Note: Points with value labels are significantly different from the census data at 0.01 level. Those without value labels are not significantly different.

Figure 19. Differences between CFPS 2010 T1 Members and the 2010 National Census



Note: Points with value labels are significantly different from the census data at the 0.01 level. Those without value labels are not significantly different.
Figure 20. Differences between CFPS 2010 Individual Respondents and T1 Members

Figure 19 shows the differences between the T1 members of CFPS and the census population in the frequency distribution of age-sex groups. The triangles and dots in the figure correspond to male and female groups, respectively. If there are no differences, the value should be 0. Values greater than 0 mean that the relative sizes of the age-sex groups in CFPS are larger than the corresponding groups in the census, while values smaller than 0 mean that the proportions are smaller. We marked out the significant differences at the 0.01 level in the figure. We can see that the sampling probability of females aged 10-14 is relatively higher, that of males aged 15-19 is lower, those of males and females aged 25-44 are lower, and those of males aged 50-79 and females aged 55-64 are higher. In general, the age-sex structure of the T1 members of CFPS is consistent with the 2010 national census.

In terms of the age-sex structure of the CGSS respondents (Figure 18), there is no representation of the age group 0-14 and the proportion of the age group 40-60 was higher, as only one family member over age 18 in each selected household was interviewed in the CGSS. Compared with CGSS, the respondents of CFPS individual questionnaires had a wider age-sex distribution, with respondents from all age groups. The age groups over age 30 in CFPS data are closer to the census, but the proportions of younger groups (aged 15-19, 20-24, and 24-29) are much lower, since these individuals are more likely to go out of town for work or study. We examined the differences in the age-sex structure between all T1 members and all respondents of individual questionnaires. We took the frequency of each age-sex group of individual respondents as the dependent variable and the frequency of each age-sex group of T1 members as the exposure. Comparing the full log-rate and the null model, we reached a Chi-square value of 1,059.8 ($df = 41$, $p = 0.0000$), which means that the age-sex structure of the actual respondents is significantly different from the structure of the ideal sample according to the survey design. Figure 20 shows the differences between all individual respondents and all T1 members of CFPS: the actually interviewed proportions of both males and females aged 0-14 are higher, those of the females aged 15-19 and of males aged 15-34 are lower, those of females aged 35-64 and of males aged 55-69 are higher, and those of females aged 85-89 are lower. Such differences in the age-sex structure of the actual respondents would affect the accuracy of statistical inference, but can be solved by weighting on unit nonresponse. We briefly introduce the calculation methods of the weighting on unit nonresponse. For further information about the weight, see *Weight Calculation* (CFPS-17).

8.2 Family Size and Household Types

In the CFPS 2010 resampled national sample, the average family size calculated from the T1 table is 3.8. Conditional on co-residence, the number is 3.3. The CFPS family size is larger than that based on the 2010 national census. A t-test shows that the difference is significant at the 0.001 level. Broken down to rural and urban areas, the CFPS family sizes are larger than those calculated from the census, which is true for all T1 members as well as for the present T1

members. Figure 21 also lists the average family sizes in CGSS, which are very close to CFPS both in national averages and by rural and urban areas.

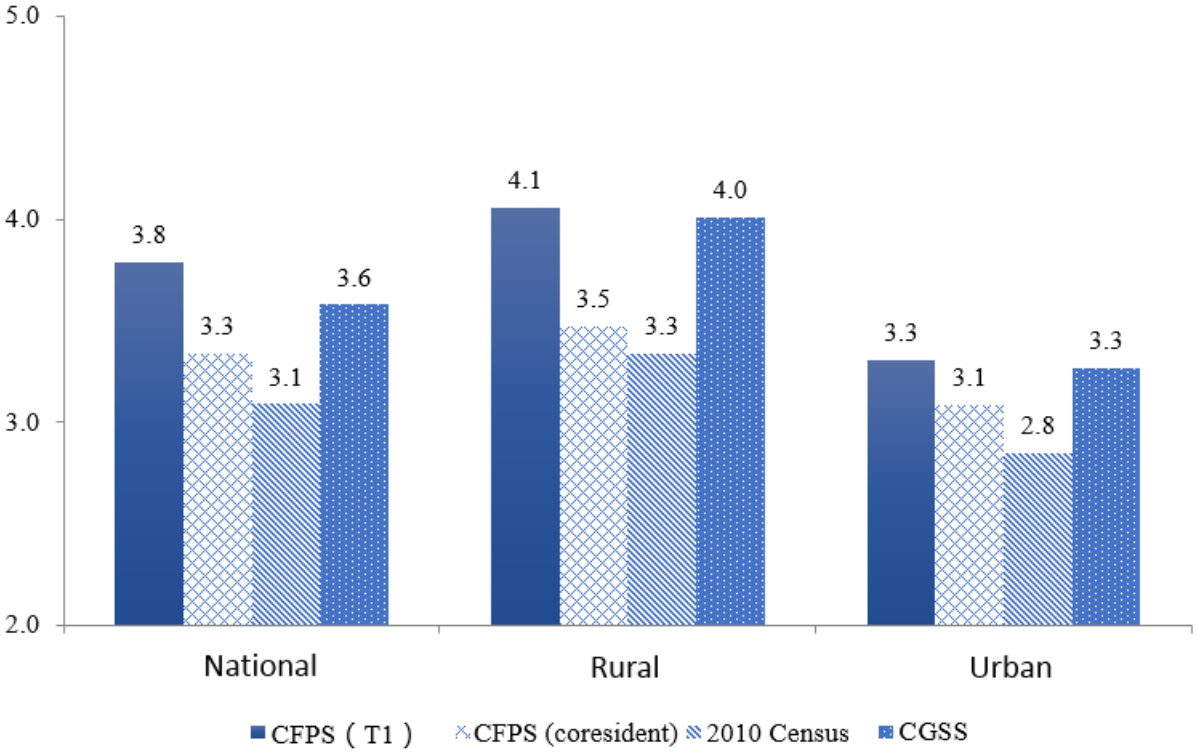


Figure 21. Average Household Size, by Rural and Urban

As shown in Figure 22, in the CFPS 2010 resampled national sample, based on the family members of the T1 table, one-generation households account for 20.2% of all households, two-generations households account for 48.7%, and three-or-more-generation households account for 31.2%. If the calculation is restricted to the members present in the home, the proportions are 29.3%, 42.8%, and 28.0%, respectively. The proportion of three-or-more-generation households is larger than that in the census while the proportion of one-generation households is lower. The Chi-square test is significant at the 0.001 level. In comparison, the family types in CGSS are closer to the CFPS results.

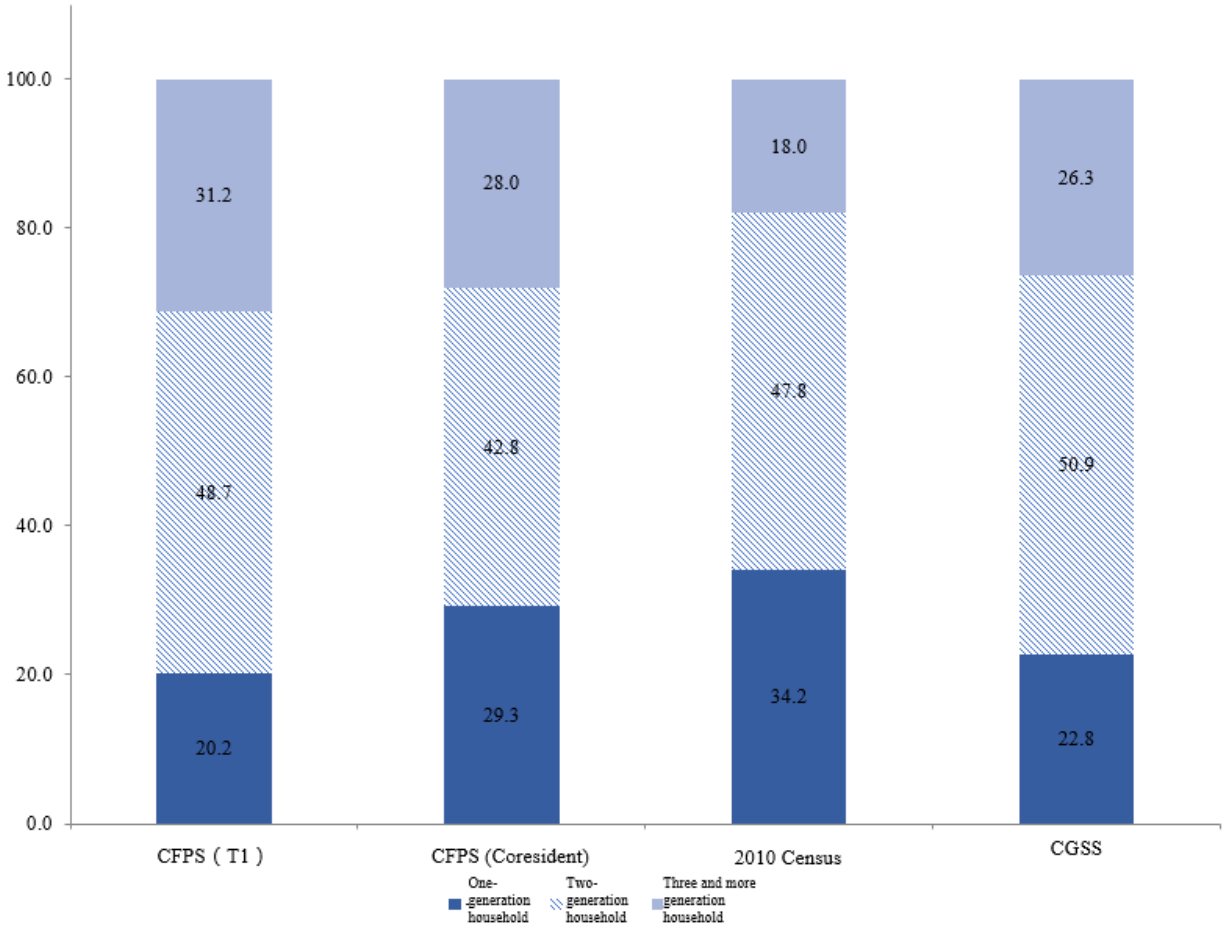


Figure 22. Household Types at the National Level

8.3 Family Income

Table 42 shows mean family incomes and Gini coefficients based on three different national samples: CGSS, CFPS, and CHFS (China Household Financial Survey). The results from CGSS and CFPS are closer. The Gini coefficient for urban families is higher in CGSS. However, in terms of mean family income and Gini coefficient, CHFS gives substantially higher estimates than CGSS and CFPS. The mean urban family income of CHFS is 87,071 *yuan*, which is nearly twice the amount of the CFPS estimate. The Gini coefficients for both rural families and urban families are greater than 0.65 in CHFS.

Table 42. Sample Size, Mean Family Income & Gini Coefficient: CGSS, CFPS, and CHFS

		CGSS2010	**CFPS2010 (adjusted)**	**CHFS2011 (24 provinces)**
Rural	Sample size	5,313	5,883	4,162
	Mean income	22,125.8	28,826.4	32,285.8
	Gini	0.495	0.498	0.675
Urban	Sample size	3,849	3,248	3,619

Mean income	53,494.0	44,917.6	87,071.4
Gini	0.535	0.470	0.655

Note: The CHFS 2011 data includes 25 provinces, which include Qinghai but not Fujian compared to CFPS 2010. The Qinghai data were excluded to improve comparability. In addition, zero or negative family incomes were excluded.

Tables 43 and 44 show the distributions of rural and urban family incomes based on these three samples. As stated above, there are two versions of CFPS 2010 income data: the unadjusted and the adjusted. The unadjusted distribution is calculated based on the original data from the survey and the adjusted distribution, which uses the total rural family income that includes the value of self-consumed products.[82] As shown in these two tables, the income distributions of CFPS and CGSS are very close for both rural and urban families. However, the CHFS data presents a rather polarized distribution. For the lower income (below 25th percentile), the proportion given by CHFS is much lower than that in the other two samples, especially for rural families. On the other hand, the proportion of higher income families (above 75th percentile) is much higher, especially for urban families.

Table 43. Distribution of Total Rural Family Income: CFPS, CGSS, and CHFS (*yuan*)

Percentile	CGSS2010	CFPS2010 (unadjusted)	CFPS2010 (adjusted)	CHFS2011 (24 provinces)
5th	2000	1940	2300	720
10th	3240	3300	4210	1750
25th	8000	9000	10000	5220
50th	15000	18000	20000	13250
75th	28880	32000	34205	32249
90th	45000	53600	56121	61025
95th	60000	74000	77628	93000

Note: The CHFS 2011 data include 25 provinces nationwide, with Qinghai added to replace Fujian. Therefore, the Qinghai Province data were excluded from the samples in order to make the data more comparable. In addition, family samples with income lower than 0 are excluded here.

Table 44. Distribution of Total Urban Family Income: CFPS, CGSS, and CHFS (*yuan*)

Percentile	CGSS2010	CFPS2010 (unadjusted)	CFPS2010 (adjusted)	CHFS2011 (24 provinces)
5th	6000	5000	5000	1700
10th	10000	9640	10000	5170
25th	20000	17800	18000	19001
50th	30000	30200	30300	38450
75th	55000	55000	55000	78000

[82] See detailed adjustment methods in Technical report CFPS-14.

90th	100000	86000	86100	169000
95th	150000	120000	120000	262500

Note: The CHFS 2011 data include 25 provinces, which include Qinghai but not Fujian compared to CFPS 2010. The Qinghai data were excluded from CFPS 2010 to improve comparability. In addition, zero or negative family incomes were excluded.

Tables 45 and 46 describe the percentile distributions of rural/urban family incomes based on the data of the three surveys. The results of CFPS and CGSS are very slightly below the 50th percentile while CHFS proportions are relatively lower. Particularly, in the CHFS rural sample, the total income of the families at the lower half accounts for only 10% of the total incomes of all the families. For families in the 5th and 10th percentiles, the results of CHFS are also significantly lower than in the other two surveys—as shown in both Tables 42 and 43. For families in higher percentiles (75th, 90th, 95th), the rural sample results of CGSS and CFPS are very close, and the CGSS urban sample shows a higher level of clustering along the distribution. However, CHFS has the highest level of clustering along the distribution in both rural and urban areas. Particularly, for the rural sample, CHFS shows that the top 5% income group has 43.1% of the total income.

Table 45. Percentile Distributions of Rural Family Income: CGSS, CFPS, and CHFS (%)

Percentile	CGSS 2010	CFPS 2010 (unadjusted)	CFPS 2010 (adjusted)	CHFS 2011 (24 provinces)
Below 5th	0.3	0.2	0.2	0.0
Below 10th	0.9	0.7	0.8	0.2
Below 25th	5.5	4.1	4.6	2.0
Below 50th	18.0	16.3	17.5	10.0
Above 75th	59.6	61.2	59.7	72.5
Above 90th	36.0	38.5	37.1	53.4
Above 95th	24.1	27.1	25.9	43.1

Table 46. Percentile Distributions of Urban Family Income: CGSS, CFPS, and CHFS (%)

Percentile	CGSS 2010	CFPS 2010 (unadjusted)	CFPS 2010 (adjusted)	CHFS 2011 (24 provinces)
Below 5th	0.3	0.3	0.3	0.1
Below 10th	1.4	1.1	1.4	0.4
Below 25th	7.7	5.7	6.2	3.6
Below 50th	16.9	19.0	19.1	14.0
Above 75th	63.0	57.6	57.6	65.6
Above 90th	39.2	35.0	35.0	43.9
Above 95th	30.1	22.8	22.9	33.2

8.4 Urban-Rural Distribution

We made a comparison of the rural-urban distribution between the CFPS 2010, 2010 census and CHFS 2011. As CGSS used stratified sampling for rural and urban areas and oversampled urban areas, the data from CGSS is not suitable for comparison with CFPS in this case.

Table 47 shows that the distributions of the T1 members and individual respondents of CFPS are very close based on the rural-urban division standard defined by the National Bureau of Statistics, though the Chi-square test of the two-way contingency table shows statistically significant differences ($\chi^2(1) = 5.0157$, $p = 0.025$). In contrast to the nearly even rural-urban distribution (the proportion of respondents in urban areas is higher than those in rural areas by 0.6%) of the 2010 national census, the CFPS data shows that the proportion of rural residents is nearly 10% higher than the urban residents. Both the log-rate analysis ("rural" as the reference group and census as the exposure) ($\chi^2(1) = 462.83$, $p = 0.000$) and the Chi-square test of the two-way contingency table ($\chi^2(1) = 462.03$, $p = 0.000$) shows a significant difference in the rural-urban distribution between CFPS and the census. This may be due to sampling differences or response rates. Also, we classified "towns" into the urban areas without considering that they may also have rural areas within their administrative or geographical boundaries. Due to lack of lower levels of census data, we are currently unable to discover the true reasons. Compared with CFPS, the rural-urban distribution of CHFS is closer to the census data, although significant difference still exists ($\chi^2(1) \approx 150000$，$p = 0.000$).

In terms of the division by the type of village/neighborhood community, the frequency distributions of T1 members and individual respondents of CFPS are very close, although the Chi-square test of the two-way contingency table is significant ($\chi^2(1) = 10.8773$，$p = 0.001$). Approximately 70% of the respondents in CFPS were living in communities under the administration of village committees.

The rural/urban division by *hukou* in the CFPS data is fairly close to the census data, which is about 70% rural and 30% urban. Although the Chi-square test of the two-way contingency table is significant, the Chi-square value is only 96.6. The frequency distribution of CHFS data is also very similar to that of the census data, although the distributions are statistically significantly different ($\chi2(1) \approx 150000$，$p = 0.000$).

Table 47. Rural-Urban Distribution (%)

	CFPS 2010			
	T1 members	Individual respondents	2010 census	CHFS 2010[b]
By National Bureau of Statistics				
Rural	55.4	54.5	49.7	48.7
Urban[a]	44.7	45.5	50.3	51.3
N	36,571	27,444	1,332,810,869	528,808,705
By Village/Neighborhood Community				
Village	69.6	68.4	—	—

Neighborhood community	30.4	31.6	—	—
N	36,571	27,444	—	—
By *hukou*				
Agricultural	—	73.6	70.9	63.3
Non-agricultural	—	26.4	29.1	36.7
N	—	27,204	131,9046,434	440,104,614

[a] "Towns" and "cities" of the census data were merged as "urban."
[b] The data of CHFS are weighted.

8.5 Educational Level

As shown in Table 48, the distribution of the highest educational level of T1 members of the CFPS 2010 baseline survey is very close to the distribution in the 2010 national census. The difference between these two distributions is mainly the slightly lower proportion of junior middle school graduates among the T1 members (a difference of about 5%). Apart from this, the differences in other educational levels are basically less than 2%. The Chi-square test of the two-way contingency table shows a significant difference ($\chi^2(6) = 493.1084$， $p = 0.000$). We took the group "illiterate/semi-illiterate" as the reference group and the census data as the exposure. The log-rate analysis shows a significant difference in the proportions of "junior middle school," "senior middle school," "junior college," and "graduate."

Table 48. Highest Level of Education, Age 6 and above (%)

	CFPS 2010		2010 census	CGSS 2010
	T1 members	Individual respondents		
Illiterate/semi-illiterate	24.31	29.7	22.9	22.7
Primary school	21.54	23.6	19.9	19.0
Junior middle school	32.05	26.2	37.6	26.3
Senior middle school	14.35	13.2	11.8	18.3
Junior college	4.65	4.3	4.7	7.7
College	2.93	2.9	2.8	5.5
Graduate	0.2	0.2	0.3	0.6
Sample size (N)	29,974	23,219	111,601,269	10,173

The comparison between CFPS and CGSS shows that the respondents whose highest educational level is junior middle school have a higher proportion among the T1 members of CFPS than that of the CGSS, but the proportions are lower for the categories of "senior middle school,""junior college," "college," and "graduate." We tested the differences to evaluate the quality of the CFPS data: for the comparison between CGSS and the census, the Chi-square test result of the two-way contingency table is significant, with a value of about 1200, much higher than the Chi-square value from comparing CFPS and the census; log-rate analysis shows that all the groups of CGSS are significantly different from the reference groups except for "primary

school." Therefore, we suspect that the source of such differences may be CGSS' over-sampling of senior middle school graduates and insufficient representation of lower educational levels.

Compared with CFPS T1 data, the respondents of the individual questionnaires in CFPS have a lower educational level, which might be caused by underestimations. Family members with lower levels of education are highly likely to stay at home; thus, this group of people are more likely to complete individual questionnaires, while family members with higher levels of education are more likely not to be present at the time of interview. In addition, it is possible that T1 respondents who answered questions on behalf of their family members intended to report higher levels of education than the actual levels.[83]

8.6 Marital Status

In addition, we compared the distributions of marital status of the population aged 15 or above between CFPS, CHFS, CGSS, and the 2010 national census to test the quality of the marriage data of CFPS. Table 49 lists the results of the comparisons. The CFPS data again has two versions—the T1 family members and the respondents to the individual questionnaires. The CHFS asks about marital status for all family members. The CGSS only includes marital status for respondents aged 18 or above. The statistical results show that the marital status distribution of T1 members of CFPS is very close to the distribution for census. The two distributions are essentially identical, although the huge sample size of the census leads to a significant difference in the Chi-square test of the two-way contingency table ($\chi^2(3) = 23.5398$， $p = 0.000$), as well as in the log-rate analysis with the census data as the exposure (χ^2 (3) =24.15, $p = 0.000$) .

Table 49. Distribution of Marital Status, Ages 15 or above (%)

		CFPS 2010				
		T1 members	Individual respondents	2010 Census	CHFS 2011[a]	CGSS 2010[b]
Total	Unmarried	20.8	14.6	21.6	18.2	8.1
	Married	72.2	78.5	71.3	76.5	82.8
	Divorced	1.2	1.2	1.4	1.3	2.1
	Widowed	5.8	5.8	5.7	4.0	7.0
	N	30,642	22,197	105,542,243	24,693	10,154
Male	Unmarried	24.1	17.0	21.6	21.1	10.1
	Married	71.2	78.2	71.3	75.3	83.6
	Divorced	1.4	1.4	1.4	1.2	2.1
	Widowed	3.3	3.4	5.7	2.3	4.1
	N	15,454	10,732	52,943,450	12,352	4,932
Female	Unmarried	17.3	12.3	18.5	15.2	6.3
	Married	73.2	78.8	72.3	77.7	82.0

[83] Technical report: CFPS-21.

Divorced	1.0	0.9	1.2	1.4	2.0
Widowed	8.5	8.0	8.0	5.7	9.7
N	15,188	11,465	52,598,793	12,341	5,222

[a] The data of CFPS 2010 are weighted.
[b] The data of CGSS 2010 represent the population at ages 18 or above.

As the proportion of youth among CFPS individual respondents is much lower than the census, there is a difference in the marital status distribution between CFPS individual respondents and the census data ($\chi^2(3) = 669.7521$， $p = 0.000$). The proportion of unmarried individuals is lower, while proportions of divorced and widowed are no different from those in the census. This indicates that the sample of CFPS individual respondents have only one source of bias—under-representation of youth.

The distribution of CHFS data is also quite similar to the distribution in the census but the difference is larger than that between CFPS T1 members and the census ($\chi^2(3) = 157.7124$, p=0.000). Because CGSS focused on family members aged 18 or above and over-sampled the urban population, the marriage distribution is significantly different from other data.

The sex-specific marital status distributions of CFPS T1 family members are also consistent with the census data, with a higher proportion of unmarried males—but the female results are essentially the same.

9. Weights Calculations

9.1 Baseline Weights

CFPS 2010 provided weights for family, adult, and child data sets for national full sample and national resampled sample. The national full sample weight is the combined weights of five "large provinces" (Shanghai, Henan, Gansu, Liaoning, and Guangdong) and "small provinces" (other provinces among 25 provinces/cities/autonomous regions). The national resampled sample weight is the combined resample weights of the five "large provinces" and "small provinces." Weight calculations take into account sampling design weights, non-response adjustment weights, post-hoc stratification adjustment weights, and trimming of the weights.

The sampling design weights are the inverses of the multiplying sampling probabilities from the first, second, and third stage sampling.[84] As for the resampled sample, the calculation also took into account the probabilities of resample from the sample districts/counties in the first stage.

Non-response adjustment weights adjust the non-responses at the family member and family/individual questionnaire levels. At the family member level, the non-response weight is based on weighted adjustment method, which used the number of completed family roster questionnaires over the total number of families of the community as the weighted adjustment coefficient. At the family/individual questionnaire level, the non-response weight is also based on weighted adjustment method. The adjustment coefficient is the number of family households that completed the questionnaires over the total number of family households expected to answer the survey. The non-response weight for the individual questionnaire uses a logistic modeling, estimating the likelihood of answering the questionnaire based on two stages as the adjustment coefficient. In the first stage, the individual sample was divided into contacted sample and non-contacted sample using supplemental information from the data to build the logistic regression model to obtain the probability of answering surveys at the individual connecting level. In the second stage, the connected samples are divided into refusal samples and non-refusal samples to build the logistic regression model and obtain the probability of a respondent refusing the interview.

The post-stratification aimed to reduce the sampling errors and increase the accuracy of the estimation by adjusting the structural biases of the sample due to the complexity of the sample design, the diversity in the field investigation, and the non-response situation. CFPS used gender, age, and rural-urban division to construct the post-stratification.

The trimming of weights restricted the adjustment coefficients of non-response and post-stratification in a certain range to control the standard errors after the non-response and post-stratification adjustments. In addition, the trimming restricted the final weights after the sampling design weight, non-response adjustment, and post-stratification adjustment in a certain range to secure the efficiency of the estimation.

Finally, in order to make the final weights and the population total equal, more adjustments were needed. After the adjustments, we finally obtained the full sample weights data sets of 25

[84] The third-stage sampling seeks to adjust for multiple eligible households at the same sampled address. One household was randomly selected in such case.

provinces and the national representative sample weights of the resamples from the 25 provinces.[85] Each data set contained weights for family, adult, child full samples and the resampled samples. The full sample weights dataset of 25 provinces is a combination of the weights of "large provinces" and "small provinces."

The usage of weights is associated with data users' target population and types of data sets. Detailed information regarding the different types of data sets and the represented total population can be found in Table 27. Table 50 lists the weighting variables in CFPS.

Table 50. Names and Labels of the Weighting Variables in CFPS

Data Set	Variable Name	Variable Label
CFPS 2010		
Family	fswt_nat	Family weight-national full sample
	fswt_res	Family weight-national resample
Adult/Child	rswt_nat	Individual weight-national full sample
	rswt_res	Individual weight-national resample
CFPS 2012		
Family	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_restpn1012	Panel weight(family level):nationally representative subsample
Adult/Child	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_natpn1012	Panel weight(individual level):nationally representative subsample
CFPS 2014		
Family	fswt_natcs14	Cross-sectional weight(family level):total sample
	fswt_rescs14	Cross-sectional weight(family level):nationally representative subsample
	fswt_natpn1014	Panel weight(family level):total sample
	fswt_restpn1014	Panel weight(family level):nationally representative subsample
Adult/Child	rswt_natcs14	Cross-sectional weight(individual level):total sample
	rswt_rescs14	Cross-sectional weight(individual level):nationally representative subsample
	rswt_natpn1014	Panel weight(individual level):total sample

[85] The weight is calculated separately for adults and children in 2010. Starting from 2012, adults and children are taken as a whole when calculating the sample weight.

rswt_natpn1014	Panel weight(individual level):nationally representative subsample

For detailed calculation methods of the weights and the weighted analysis results of each CFPS data set, see *CFPS Baseline Survey Weights Calculation* (CFPS-17).9. Technical Reports

9.2 Weights in the follow-up survey

As stated before, CFPS only tracks gene members and their families in follow-up surveys.With the birth of new gene members, the death of existing gene members, and family splitting for reasons such as divorce, our sampling frames are changing. Also, because of the inevitable non-responses and attritions, both the sample and the population are changing. Given these changes, users should employ appropriate weight adjustment to achieve sample representativeness and more efficient statistical inference.

CFPS weights are applicable to the gene members, including cross-sectional weights and panel weights. Cross-sectional weights are available for both original gene members from 2010 and new gene members, and panel weights are available for baseline gene members from 2010. We have cross-sectional weights and panel weights at both the individual and family levels. There are six sampling frames in CFPS and the weights are corrected for each sampling frame separately. The weight construction for individuals and families of the national full sample is as follows. The processes of the other sampling frames are similar to this.

9.2.1 Individual level weight in follow-up survey

The adjustment of respondent weights in the CFPS follow-up survey includes original individual weights for 2010 gene members, weights for new gene members, non-response adjustment weights, post-stratification adjustment weights, and trimming.

The CFPS follow-up survey aims to track all previous gene members and new gene members (newly-born and adopted children of 2010 gene members). The original weights for 2010 gene members are their baseline weights adjusted for non-response. It is worth noting that a county in Sichuan Province was not surveyed in 2012 for administrative reasons, so this county was regarded as non-responsive at the county level. The county was eliminated from the 2010 data in the weight correction of CFPS 2012 survey before we performed the county-level non-response adjustment within Sichuan Province.

For the new gene members, we used the average of the individual weights of their parents as their follow-up weights. If only one parent was a CFPS gene member, then that parent's individual weight became the child's weight.

In order to enhance the accuracy of the weights, the non-response adjustment of an individual weight used the family roster data and adult/child data from both the baseline survey and subsequent survey. We used the survey data to obtain the non-response adjustment weight at the individual level based on the calculation of propensity weights in the Logistic Model. To be more specific, the individual samples in the follow-up years were categorized as complete cases and incomplete cases. Logistic regression models used age, the square of age, gender, family size, whether there was an old man or child at home, urban status, number of generations within the family, home ownership, and interview status in different waves. All individual samples including both the adults and children were treated as a whole during the weight calculation. Logistic models were estimated within each of the six sampling frames.

The sampling design and complications of the CFPS field survey, together with non-response and loss of samples, resulted in some structural differences between the sample and the population, thus affecting the accuracy of the estimation. To adjust the sample structure, minimize sampling error, and enhance estimation accuracy, the individual sample data needed to be post-stratified. Gender, age, and urban status are very important indicators at the individual level, so we used variables of urban (urban or rural), gender (male or female), age (16-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, >80) to do the post-stratification. The data used in the post-stratification were the latest available official population data. For example, we used the sixth national census data for CFPS 2010 and 2012, and the 2014 population sample survey for CFPS 2014. For a very small number of missing in age and gender in the CFPS individual questionnaire, we used mean and median imputation.

To avoid the problem of large variance and loss of efficiency of estimation in the individual-level data due to extreme weights, we trimmed the individual level weights at the 5th and 95th percentiles.

In conclusion, the product of all the individual adjustment weights mentioned above was the final individual weight. Individual panel weights are not available for new gene members, who have only the cross-sectional weights.

9.2.2 Family-level weight in the follow-up survey

The adjustment of family weights includes original family weights from the previous wave, non-response adjustment weights, post-stratification adjustment weights and trimming.

The original weight of every family (split family included as well) is the average of non-response adjustment weights among all the gene members within that family.

Not all families complete all the questionnaires. Families without individual questionnaires also need adjustment of non-response weight. We use the interview status of family samples based on the response rate of AAPOR RR1 to get the non-response adjustment rate of families through the method of calculating the weight at county level.

We also use trimming to adjust the weights at the family level mentioned above by regarding the 5th and 95th percentiles as the boundaries of outliers.

Trimming causes a discrepancy between the total weights of the sample and the populaton, so further adjustment is needed. We simply regard every sampling frame as even-distributed and correct the weights to achieve the consistency between the sample and the population.

In conclusion, the product of the adjustment weights at the family level mentioned above is the final family-based weight.

10. Technical Reports

In addition to this handbook, a series of technical reports are being released. The reports address different aspects of the CFPS project, and help users to gain a better understanding of the survey and the data. A list of the technical reports currently available is provided below. More reports will be added in future work on the survey project.

CFPS-1: *Sampling Design*. Yu Xie, Zeqi Qiu, Ping Lv, 2012. (in Chinese)

CFPS-2: *Third-stage Sampling Frame Construction*. Hua Ding, 2012. (in Chinese)

CFPS-3: *The Implementation Report*. Hua Ding, 2012. (in Chinese)

CFPS-4: *Quality Supervision Report*. Jie Yan, Yi Sun, Xueliang Teng, Liying Ren, Yan Sun, 2012. (in Chinese)

CFPS-5: *Sample Contact Results*. Yan Sun, 2012. (in Chinese)

CFPS-6: *Compilation and Data Cleaning of Family Relations Data Set*. Yuhuan Sun, Yu Xie, Jingwei Hu, Chunni Zhang, Qi Xu, Guoying Huang, 2012. (in Chinese)

CFPS-7: *Data Cleaning of the Family Relations Data Set*. Qi Xu, Chunni Zhang, Yuhuan Sun, Jingwei Hu, Ping Lv, 2012. (in Chinese)

CFPS-8: *Occupational and Industry Codes*. Liying Ren, Li Li, Chao Ma, 2012. (in Chinese)

CFPS-9: *Career Expectation Codes*. Yu Xie, Wangyang Li, Chao Ma, Guoying Huang, Airan Liu, 2012. (in Chinese)

CFPS-10: *Conversion of Occupational Codes and Construction of Socioeconomic Indices (CFPS-10)*. Guoying Huang, Yu Xie, 2012. (in Chinese)

CFPS-11: *Composite Variables (I): Verbal and Mathematical Tests*. Hongwei Xu, Weixiang Luo, 2012. (in Chinese)

CFPS-12: *Composite Variables (II): Educational Level and Depression Scale*. Yu Xie, Qi Xu, Chunni Zhang, Hongwei Xu, 2012. (in Chinese)

CFPS-13: *Composite Variables (III): "Best Variables"for Age and Marriage*. Chunni Zhang, Qi Xu, Yan Sun, 2012. (in Chinese)

CFPS-14: *Adjustments of Rural Family Income*. Yu Xie, Chunni Zhang, Guoying Huang, Qi Xu, Hongwei Xu, 2012. (in Chinese)

CFPS-15: *Income and Expenditure Data*. Yan Shen, Xiaoyan Lei, 2012. (in Chinese)

CFPS-16: *Income Comparisons between CGSS, CHIP, CHFS, and CFPS.* Qi Xu, Chunni Zhang, Xiang Zhou, Yu Xie, 2012. (in Chinese)

CFPS-17: *Weight Calculation*. Ping Lv, Yu Xie, 2012. (in Chinese)

CFPS-18: *Sample Maintenance*. Ping Lv, 2012. (in Chinese)

CFPS-19: *Poverty Rate Comparisons between CFPS, CGSS, CHIP, and CHFS*. Chunni Zhang, Qi Xu, Xiang Zhou, Xiaobo Zhang, Yu Xie, 2012. (in Chinese)

CFPS-21: *Collection, Cleaning and Evaluation on Education Level.* To be released. (in Chinese)

CFPS-22: *Composite Variables (IV): Parents' Social Status.* Chunni Zhang, Hua Ye, Lihong Dai, Jingwei Hu, Yu Xie, 2013 (in Chinese)

CFPS-23: *Masking of the CFPS county level variables*. Yahong Cui, Qiong Wu, Hongwei Xu, Guangzhou Wang, 2014.

CFPS-24: *Number of biological children and their demographic information*. Zheng Mu, Yu Xie, 2014.

CFPS-25: *Data processing of CFPS 2012 and data introduction*. Qiong Wu, Lihong Dai, Yahong Cui, Wenjia Zhang, 2014.

CFPS-26: *Psychological scales in CFPS 2012.* Weixiang Luo, Lingwei Wu, 2014.

CFPS-27: *Adjustment of the income data in CFPS 2012*. Qi Xu, Chunni Zhang, 2014.

CFPS-28: *Dialect coding in CFPS*. Lingwei Wu, Wenjia Zhang, 2014.

CFPS-29: *Data processing report for asset data in CFPS 2010 and 2012*. Yongai Jin, Yu Xie, 2014.

CFPS-30: *Constructing a composite variable of main occupation in CFPS 2012*. Wangyang Li, Jingwei Hu, Yu Xie, Qiong Wu, 2014.

CFPS-31: *Number series test in CFPS 2012*. Hongwei Xu, Yu Xie, 2015.

CFPS-33: *Constructing the family roster data set in CFPS 2012*. Lihong Dai, Yan Sun, Qi Xu, Qiong Wu, 2015.

CFPS-34: *Data processing of CFPS 2014 and data introduction* . Qiong Wu, Lihong Dai, Cong Zhang, Yulei Wang, Wenjia Zhang, 2016.

11. References

McKinleyK, T. and K. Griffin. "The Distribution of Land in Rural China." *Journal of Peasant Studies* 21(1) (1993): 71-84.

Xie, Yu. "Evidence-Based Research on China: A Historical Imperative." *Chinese Sociological Review* 44(1)(2011): 14.

Xie, Yu and Jingwei Hu. "An Introduction to the China Family Panel Studies (CFPS)." *Chinese Sociological Review* 47(1)(2014): 3-29.

Xie, Yu and Ping Lu. "The Sampling Design of the China Family Panel Studies (CFPS)." *Chinese Journal of Sociology* 1(4)(2015): 471-484.

The Institute of Social Science Survey, Peking University, 2009. *China Report 2009*. Beijing: Peking University Press. (in Chinese)

The Institute of Social Science Survey, Peking University, 2010. *China Report 2010*. Beijing: Peking University Press. (in Chinese)

The Institute of Social Science Survey, Peking University, 2011. *China Report 2011*. Beijing: Peking University Press. (in Chinese)

Ren, Qiang and Xie Yu. 2011. “Statistical Analysis of Longitudinal Data.” *Population Research* 35(6): 3-12.

Sun, Yan, Yan Jie, Ding Hua, Gu Jiafeng, Liu Yue, Yao Jiahui and Zou Yanhui. 2011. *China Family Panel Studies 2010 – Interviewer Training Handbook*. Beijing: Peking University Press. (in Chinese)

Xie, Yu. 2010. “Understanding the Inequality in China.” *Chinese Journal of Sociology* 30(3): 1-20.

Xie, Yu and Miranda Brown. 2009. “Between Heaven and Earth: Dual Accountability of Han Officials.” *Chinese Journal of Sociology* 31(4): 1-28.

Xie, Yu. 2011. “Evidence-Based Research on China: A Historical Imperative.” *Chinese Sociological Review* 44(1)：14-25.

Xie, Yu. 2012. Socialogical methodology and quantitative research (2nd edition). Beijing: Social Science Academic Press (China).

Xie, Yu, Jingwei Hu, & Chunni Zhang. 2014. The China Family Panel Studies: Design and Practice. Chinese Journal of Sociology, 34 (2): 1-32.

Xie, Yu, Xiaobo Zhang, Jianxin Li, Xuejun Yu, & Qiang Ren. 2014. China report 2014. Beijing: Peking University Press.

Xie, Yu, Xiaobo Zhang, Jianxin Li, & Qiang Ren. 2016. China report 2016. Beijing: Peking University Press.