Published in final edited form as:
Chin J Sociol. 2015 December ; 1(4): 471–484. doi:10.1177/2057150X15614535.

The Sampling Design of the China Family Panel Studies (CFPS)

Yu Xie and
University of Michigan and Peking University

Ping Lu
Peking University

Abstract

The China Family Panel Studies (CFPS) is an on-going, nearly nationwide, comprehensive, longitudinal social survey that is intended to serve research needs on a large variety of social phenomena in contemporary China. In this paper, we describe the sampling design of the CFPS sample for its 2010 baseline survey and methods for constructing weights to adjust for sampling design and survey nonresponses. Specifically, the CFPS used a multi-stage probability strategy to reduce operation costs and implicit stratification to increase efficiency. Respondents were oversampled in five provinces or administrative equivalents for regional comparisons. We provide operation details for both sampling and weights construction.

Introduction

The China Family Panel Studies (CFPS), launched by Peking University, is a nearly nationwide, comprehensive, longitudinal social survey that is intended to serve research needs on a large variety of social phenomena in contemporary China (Xie and Hu 2014; Xie, Hu, and Zhang 2014). The CFPS is designed to collect individual-, family-, and community-level longitudinal data. The studies focus on the economic, as well as the non-economic, wellbeing of the Chinese population, gathering a wealth of information covering such topics as economic activities, education outcomes, family dynamics and relationships, migration, and health. The baseline CFPS survey in 2010 successfully interviewed 14,960 households and 42,590 individuals living in these households. Four waves of the CFPS (2010, 2011, 2012, and 2014) have been carried out thus far. The CFPS seeks to provide the academic community with the most comprehensive and highest-quality survey data to date on contemporary China.

This paper describes the sampling design of the baseline CFPS survey in 2010 and the methods for weighting the data to make the CFPS sample representative of the Chinese population. In doing so, the paper draws heavily on various documents of the CFPS project in Chinese (Ding 2012; Lu and Xie 2012; Xie 2012; Xie, Qiu, and Lu 2012; Xie, Hu, and Zhang 2014). Interested readers who read Chinese may refer to the original documents referenced above.

Direct all correspondence to Yu Xie, P. O. Box 1248, 426 Thompson Street, Population Studies Center, University of Michigan, Ann Arbor (yuxie@umich.edu), Ann Arbor, MI 48106-1248.

Survey Objectives and Target Sample Sizes

The China Family Panel Studies (CFPS) is a longitudinal survey project that attempts to collect information on a nearly nationally representative sample of families and all family members in those sampled families in the 2010 baseline. See Xie and Hu (2014) for an introduction to the project and its follow-up design. The CFPS sample contains a few features that were designed to meet certain research objectives while overcoming some obstacles.

First of all, we should explain what we mean by the phrase "*nearly* nationally representative." While true national representativeness would be ideal, the CFPS project did not have the resources to conduct longitudinal surveys in remote, minority regions, especially where travelling would be very difficult, and non-Han languages would likely be required. To contain costs, a decision was made not to include Xinjiang, Tibet, Qinghai, Inner Mongolia, Ningxia, and Hainan. Needless to say, Hong Kong, Macao, and Taiwan were also excluded. However, the remaining 25 provinces or administrative equivalents represent 94.5 percent of the total population in China (excluding Hong Kong, Macao and Taiwan).Thus, we state that the CFPS sample is "*nearly* nationally representative" or, for convenience, simply "nationally representative."

China is known to be regionally diverse (Xie 2010; Xie and Hannum 1996; Xie and Zhou 2014). The CFPS project anticipated research needs for studying regional variations. While any national sample would necessarily consist of units that are geographically diverse, such units in different geographic locations are only part of a national sample but do not represent the subnational geographic units (such as provinces) to which they belong, when the larger geographic units (such as provinces) were not used for defining sampling frames. To overcome this problem, the CFPS project chose to oversample populations in five selected provinces (or administrative equivalents): Liaoning, Shanghai, Henan, Guangdong, and Gansu. For convenience, these five are called "large" provinces. With sampling frames within the large provinces, we drew subsamples within them. The resulting subsamples are self-representative of the populations in the five provinces. We collapse the other 20 remaining provinces (or administrative equivalents) together as a large residual group, from which a large subsample is drawn. For convenience, these 20 are called "small" provinces. Note that sampled units within each of the 20 provinces in this subsample are not representative at the province level. Through appropriate weighting, the whole CFPS sample can achieve representativeness of the 25 provinces in China, thereby representing China as a whole. The names of the large and small provinces and target sample size (in terms of households) are provided in Table 1.

In summary, the CFPS sample covered six subpopulations in 2010, i.e., five large provinces and the remaining small provinces. The sample sizes were chosen based on the experiences of other large surveys conducted in China and other countries, as well as the CFPS pilot survey conducted in Beijing, Shanghai, and Guangdong in 2008 and 2009, the cost of data collection, and the precision needs of statistical estimation. The target number of households in the study for each of the large provinces was 1,600. The target number of households in the small provinces was 8,000. Thus, the total target number of households in the study was 16,000.

General Principles for Sampling Design

For two practical reasons, the CFPS used a multi-stage probability strategy. First, China is a vast country. Had a truly random sample been drawn, the cost of conducting face-to-face interviews would have been prohibitive. Second, the CFPS study is interested not only in families and individuals but also in their communities. A hierarchical sampling structure in which families are nested within sampled communities would allow the researchers to study how communities might affect families and individuals (Raudenbush and Bryk 2002; Xie and Hannum 1996). In short, the CFPS sample was designed to be multi-stage both to reduce the operation costs of the survey and to represent the heterogeneity of social contexts. Each of the six CFPS subsamples discussed above was drawn over three stages. At each stage, implicit stratification was employed to improve efficiency.

At the first two stages in the sampling process, official administrative entities were used. The administrative structure in China has two important features: first, it is strictly hierarchical; second, at least in theory, it covers the entire population of China exhaustively, without exception. Because Shanghai Municipality is different from other large provinces, a special provision was made for drawing the Shanghai subsample. For all other five subsamples, we drew counties or their administrative equivalents, districts, (except in the case of Shanghai, to be noted below) at the first stage. We then drew communities in selected counties/districts at the second stage. Communities were defined as either administrative villages (*cunweihui*) for rural areas or resident committees (*juweihui*) for urban areas.

We placed major emphasis in the sampling design on regional representation, as economic development in China has been regionally diverse. Our implicit stratification gave primacy to region so that a good regional representation was ensured. In each province, the capital city was singled out for implicit stratification. When there was a meaningful urban/rural division, this information was always used in stratification. In general, a *district*, a *subdistrict*, or a *resident committee* refers to an urban area; correspondingly, a *county*, a *township*, or a *village* refers to a rural area. Besides the urban/rural divide, we used a continuously measured socioeconomic indicator (SEI) whenever possible for implicit stratification. Depending on data availability, local per-capita Gross Domestic Product (GDP), percent non-agricultural population, or population density was used as an SEI for stratification, in order of preference.

At the third stage, the onsite sampling, households were drawn from a sampled community. The onsite sampling frame, a list of all households, was obtained by mapping out all residential dwelling units in a sampled community. Within this sampling frame, households were selected using systematic sampling. In anticipation of invalid addresses and nonresponses, sampling augmentation was used so that about 25 households would be successfully interviewed from each community.

Sampling Operations at Different Stages

Sampling at Stage One

For the four large provinces other than Shanghai, i.e., Liaoning, Henan, Gansu, and Guangdong, districts (if urban) and counties (if rural) constituted a sampling frame. For implicit stratification, districts and counties were sorted according to auxiliary information as follows:

1. All cities or administrative equivalents within a province were rank ordered, with the capital city at the top, and all the other cities (prefecture-level cities) or equivalents were listed in descending order by a socioeconomic indicator (SEI).
2. Each city or administrative equivalent in this list was divided into three segments: (1) districts, (2) county-level cities, and (3) counties. Within each segment, counties (or county-level cities/districts) were listed in descending order by an SEI.
3. The counties (or county-level cities/districts) thus ordered in a single list constituted primary sampling units (PSUs) within a sampling frame.

Sources for constructing the sampling frames are given in Appendix A. From each of the above four sampling frames, 16 PSUs were selected using systematic probability proportional to size (PPS) sampling, as described in Appendix B.

There are only 19 districts and their administrative equivalents—counties—in Shanghai. To improve efficiency, sampling at stage one in Shanghai took place at a lower administrative level than that of county/district: subdistricts (*jiedao*) for urban areas and townships for rural areas. For implicit stratification, all subdistricts and townships were sorted on a single long list according to auxiliary information, as follows:

1. The 19 districts and counties were listed in descending order by an SEI.
2. Each district/county was further separated into three distinct segments: subdistricts, towns, and townships.
3. Within each segment, subdistricts, towns, and townships – primary sampling units (PSUs) in Shanghai -- were listed in descending order by an SEI.
4. The subdistricts, towns, and townships thus ordered were joined across all districts and counties in a single list, constituting primary sampling units (PSUs) within a sampling frame.

Within this sampling frame, 32 PSUs were selected using systematic probability proportional to size (PPS) sampling (See Appendix B). Sources for constructing the sampling frames are given in Appendix A.

For the 20 small provinces, districts (if urban) or counties (if rural) in all these provinces jointly constituted a large sampling frame. For implicit stratification, districts or counties were sorted according to auxiliary information, as follows:

1. The 20 provinces or administrative equivalents were ranked in descending order by an SEI.
2. Within a province, cities (prefecture-level cities) or their administrative equivalents were ranked with the capital city at the top and all other cities or equivalents in descending order by an SEI.
3. Each prefecture-level city or its equivalent was divided into three segments: (1) districts, (2) county-level cities, and (3) counties. Within each segment, all counties (or county-level cities/districts) – primary sampling units (PSUs) in all 20 small provinces –were listed in descending order by an SEI.
4. The counties (or prefecture-level county-level cities/districts) thus ordered were joined across all provinces in a single list. They constituted primary sampling units (PSUs) within a sample frame.

From this large sampling frame, 80 PSUs were selected using systematic PPS sampling, as described in Appendix B. Sources for constructing the sampling frames are given in Appendix A. We present the distribution of actually sampled districts/counties across the 20 small provinces in Table 2. The distribution is roughly proportional to population distribution across the provinces.

Sampling at Stage Two

The sampling procedure at stage one as described above resulted in a sample of PSUs: 144 districts/counties in provinces other than Shanghai, and 32 subdistricts/towns/townships in Shanghai. We constructed a sampling frame for the second-stage sampling in each of these 144 districts/counties and 32 subdistricts/towns/townships selected at the first-stage sampling. To construct an efficient, high-quality sampling frame, we collected a wealth of data on communities in these PSUs, including name, administrative division code, the latest resident population, the number of out-migrants, and the population density. We use administrative designations to define types of communities. We use “administrative villages” for rural areas and “resident communities” for urban areas.

There are 43,805 communities eligible for stage-two sampling in the 162 districts/counties selected in stage-one sampling.[1]

The average number of communities for each district/county is 272. The population size also varies greatly across communities in these districts/counties. In some communities, the population size is less than 100 persons, whereas in other communities, the population size is greater than 2,000 persons. When communities were fewer than 300 persons, neighboring communities were combined, whereas a very large community was split only if it was selected at stage-two sampling, a subject to be discussed later. We stopped the combination procedure when a combined population reached 300 for the newly combined community. In principle, administrative villages and resident committees could not be combined together. Also, communities in one subdistrict/town/township could not be combined with those in a

[1] The 32 selected subdistricts/towns/townships in Shanghai were located in 18 districts/counties. Thus we collected community data only in these 18 districts/counties.

different subdistrict/town/township. Specific guidelines for combination are given in Xie, Qiu, and Lu (2012). The newly combined units were regarded as virtual communities in the sampling frames of communities. The virtual community's population was the sum of the populations of the merged component communities. Likewise, its area was the sum of the areas of the merged component communities. All communities, original or combined, in each selected country/district, constituted the sampling frame for second-stage sampling.

Ideally, for sampling efficiency, communities should also be sorted according to auxiliary information. Unfortunately, however, auxiliary SEI information at the community level was limited. In the sampling process, we sorted the communities by a natural coding system used by the government agency previously known as the National Population and Family Planning Commission (NPFPC), now part of the National Health and Family Planning Commission. In many areas, the NPFPC simply adopted the coding system used by the National Bureau of Statistics. Because the NPFPC coding system is administratively and/or geographically based, it contained useful information for stratification, approximately reflecting the level of economic development or physical distance from the county government.

In each of the PSUs that were counties/districts, i.e., PSUs outside Shanghai, four communities were randomly drawn through systematic sampling. In Shanghai, two communities were randomly drawn in each of the 32 selected PSUs. The following auxiliary information was used for implicit stratification:

1. When appropriate, a PSU was divided into three segments: subdistricts, towns, and townships.
2. Within each segment, communities were further divided into resident committees and administrative villages, which were sorted within each type of community by an administrative NPFPC coding system.
3. The communities thus ordered in a single list within a county/district constituted primary sampling units (PSUs) within a sampling frame.

From each sampling frame in a selected PSU, either two (in the case of Shanghai) or four (for the other 24 provinces) communities were selected according to a systematic PPS method (Appendix B). In total, 640 communities/villages were drawn in the second stage, 8 of which were virtual communities.

It should be noted that a few communities had been demolished or relocated by the time the sample of communities was drawn. In such cases, we replaced the original communities with neighboring communities that had similar characteristics. In practice, 11 communities were replaced for this reason.

Sampling at Stage Three, Onsite Sampling

To minimize sampling frame errors, we based the third and final sampling stage, the onsite sampling, on sampling frames that were constructed according to maps of dwelling units drawn by CFPS field staff, rather than official records of household registration. Specifically, for each sampled community, CFPS field staff drew a map of all dwelling units. After sorting out unoccupied units, nonresidential units, commercial units, commercial-and-residential units, multi-household units, and multi-residence households, we formed the onsite sampling frame as the list of all dwelling units for the third stage of the sampling process. However, the following few situations still required special attention:

1. Multi-household unit and multi-residence household. The CFPS defined a household as a collection of multiple related persons who shared both a dwelling and economic resources, such as food (Xie and Hu 2014; Xie, Hu, and Zhang 2014). A multi-household unit was defined as a dwelling unit in which two or more households lived. For example, two married brothers still lived together after a household division; a housing unit was rented to several lodgers; elderly persons co-resided with their married children but were economically independent. If we had drawn the CFPS sample strictly according to dwelling units, some targeted households might have been missed. To guard against such errors, a multi-household unit was split, each household being numbered separately. Note that only the dwelling units with different households were considered multi-households, while households with two household registration booklets were not. During the fieldwork of mapping dwelling units, whether or not a dwelling unit was occupied by a single household or multiple households was judged by the field staff observing doorbells, mail boxes, separate entrances, or the number of electricity meters, or by asking informants.

2. A multi-dwelling household. A multi-dwelling household was a household with two or more residential housing units. If the sample had been drawn strictly according to dwelling units, this would have led to the error of over-coverage. We filtered dwelling units by the household head’s name. When we found households that occupied multiple dwelling units, we combined multiple dwelling units belonging to the same household. Note that households with multiple dwelling units currently *in use* were counted as multi-dwelling households. Suppose that a household owned two housing units. If one unit was self-occupied, and the other was unoccupied, this household was actually not a multi-residence household. We considered the unused housing unit to be an “unoccupied unit.” If one unit was self-occupied and the other was leased out, this household was not considered a multi-residence household either. The leased unit was a normal dwelling unit subject to sampling.

3. The boundary of the onsite sampling frame. In general, the boundary of the onsite sampling frame was the boundary of the administrative area at the community level. In some situations, two administrative villages, or an administrative village and a resident committee, were mixed. To resolve the difficulty, we applied the spatial boundary principle so as to classify residents in their communities according to the geographic location of their dwelling units (Xie, Qiu, and Lu 2012). Note that official household registration was also consulted to determine to which community a household belonged.

4. Indetermination of the type of housing unit. In the fieldwork, it was sometimes difficult to know the exact type of a housing unit: a residential unit, a commercial

unit, commercial-and-residential unit, non-occupied unit, non-residential unit, multi-household unit, or multi-dwelling household. In such cases, the housing unit was temporarily given a conservative designation. Specifically, if we were not sure about the type of a housing unit, we treated it as residential. If we were not sure whether several dwelling units belonged to one household, we treated these dwelling units as separate dwelling units. If we were not sure whether a housing unit was a commercial or a commercial-and-residential unit, we treated it as a commercial-and-residential unit.

5. The splitting of large communities. Given a great variation in population size across communities, communities with a very large population (over 10,000) were split in the sampling frames of communities. We did this for two reasons. First, the probability of being sampled would be disproportionately high for very large communities, thereby reducing the efficiency of sampling at the community level. Second, drawing a map of housing units of a large community would be very demanding in fieldwork. Of course, splitting a large community would consume a great deal of human, material, and financial resources. In practice, a very large community was split only if it was selected at stage-two sampling (Xie, Qiu, and Lu 2012). Although this approach reduced cost, it required one additional sampling stage that would increase sampling error and reduce estimation precision.

To enhance the efficiency of the onsite sampling frame, all dwelling housing units were sorted according to the route traveled while drawing the map of the housing units within each selected community, or in clockwise order, starting from the northwest. That is, we used traveling route for implicit stratification. At the end of the procedure, we constructed 640 onsite sampling frames. In Table 3, we present the number of housing units by type and province.

As shown in Table 3, the proportion of effective housing units (normal residential units and commercial-and-residential units) in the onsite sampling frames was 92%. Multi-household units constituted 1.3%, and multi-residence households constituted 3.1%. The remaining 4% of the units were non-residential, non-occupied, or commercial.

Sample augmentation, sampling more households than the target number in each sampled community, was used in anticipation of non-responses. Relative to sample replacement, the sample augmentation approach yields standard response rates that are comparable to those of other surveys. Since eligible households were the target population, we took into account situations such as unoccupied units, commercial units, inaccessible units, and refusal units in sample augmentation. The number of housing units drawn from each selected community is given as follows:

$$n = \frac{\text{target number of housholds}}{1 - (\text{refusal rate} + \text{unoccupation rate} + \text{inaccessibility rate} + \text{ineligibility rate})}$$

The number of households drawn varied by region and urban/rural status, according to the expected response rate. We projected response rates by region and urban/rural status based on the experiences of other surveys conducted in China and other countries, as well as a CFPS pilot survey in Beijing, Shanghai and Guangdong in 2008. The variation in the expected response rate and number of sampled units within a community is given in Table 4. As shown in Table 4, between 28 and 42 households were randomly drawn in each onsite sampling frame, using systematic sampling. The total number of households drawn was 19,986. This number was slightly adjusted in actual fieldwork (Xie, Qiu, and Lu 2012).

Weights

When using data from the CFPS, the researcher should employ appropriate weights to achieve sample representativeness for units of analysis (families or individuals), adjusting for unequal probabilities that arose from sampling design, non-responses, and other factors. We discussed in detail the construction and proper uses of CFPS weights in a technical report (Lu and Xie 2012).

The sampling design weight is the inverse sampling probability of a unit being included through all the sampling stages, i.e., the reciprocal of the product of the sampling rates of the first, second, and third stages. There are two primary reasons why the sampling design weight varies. First, the CFPS sample contains oversamples in five large provinces. Second, although we used the latest administrative data to draw samples according to the PPS principle, as explained in Appendix B, actual population sizes may have been different. Unfortunately, we could verify population sizes only for units that were selected for inclusion. Adjustment thus needs to be made for the discrepancy between the estimated population size used in an earlier stage sampling and the updated population size used in a later stage sampling.

The first design factor for unequal sampling probability, i.e., oversampling in the five large provinces, is also handled in the CFPS data files with an indicator (subsample=1) for a subsample that is nationally representative by design, i.e., without oversamples in the five large provinces (Xie and Hu 2014). We call this subsample the “resampled sample,” in contrast to the complete sample. For this reason, there are two sets of parallel weights--one for the resampled sample and the other for the complete sample.

We constructed weights to correct for non-responses at both the family and the individual levels. By the family level, we mean non-responses to the family questionnaire, more precisely the family member questionnaire--the section listing all family members and their socio-demographic characteristics. By the individual level, we mean non-responses to the adult and child questionnaires. Again, a weight correcting for non-response is the inverse probability of response.

The weight correcting for family-level nonresponses was based on the probability that a sampled household nested within a community would respond. An implicit assumption is ignorability within a neighborhood: families within a selected community are relatively homogenous and thus good substitutes for one another. The weight correcting for individual-level nonresponses was subsequently based on the probability that a family person listed on a completed family member questionnaire, i.e., a person nested within a family, would respond. Further, we model the propensity of individual-level responses based on individuals' socio-demographic characteristics. That is, we further assume ignorability at the personal level: controlling for differences across communities and in observed socio-demographic characteristics, individuals who failed to respond can be approximated by those who did respond. We estimated logistic models to calculate the probability of response as a function of the following socio-demographic variables: urban/rural status, county fixed effect, age, gender, marriage status, education, migration status, family size, presence/non-presence of an elderly person, presence/non-presence of a child, and housing ownership. We used a spline function for age.

The third step for calibrating the weights constructed from the previous stages was post-stratification, in which we aligned the weighted CFPS data in accordance with known key population-level demographic statistics. We did this only at the individual level, i.e., for adult and child questionnaires. We focused on sex, age, and urban/rural status within a sampling frame. After we broke down age into age intervals (16–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, above 80), we formed a full grid for the three categorical variables so as to compare the distribution in the CFPS data to the known distribution obtained from the 2010 China population census within a sampling frame. We adjusted the final individual-level weights for both the resampled and the complete samples so that the resulting distribution was the same as the population distribution.

To minimize the large influences of cases with extremely large weights and preserve the efficiency of the data, we checked on the variation in our weights. We trimmed the extreme weights below the 5th percentile and above the 95th percentile to the 5th and 95th percentiles respectively. Trimming resulted in substantially reducing the variation of the weights (Lu and Xie 2012).

Conclusion

The China Family Panel Studies (CFPS) aims to provide comprehensive, nearly nationally representative, longitudinal data on a variety of social and economic domains in contemporary China. In this paper, we reviewed the sampling design of the CFPS sample for its 2010 baseline survey. There are three main features of the design. First, five provinces or administrative equivalents were oversampled for regional comparisons. Second, the CFPS used a multi-stage probability strategy to reduce operation costs. Third, implicit stratification was used throughout to improve statistical efficiency. Due to the complicated way the CFPS sample was constructed, future research is still needed to evaluate the effectiveness of the CFPS sample in the 2010 baseline survey in representing the 2010 Chinese population.

We also discussed methods for constructing weights to adjust for both sampling design and survey nonresponses. The methods are essentially sequential, following the steps of probability sampling. The weight correcting for a family's non-responses was based on the community within which the family was located. The weight correcting for a person's non-responses further incorporated information about a person's socio-demographic characteristics. In addition, the CFPS also used post-stratification to align key demographic statistics resulting from the sample with population statistics within each sampling frame.

While data from the CFPS may be of use to many researchers, we recommend that all users of the CFPS data know the sampling design of the CFPS survey and apply in their research the appropriate weights that are provided in the data files.

Acknowledgments

The research is supported by financial support from the Natural Science Foundation of China (grant no. 71461137001), the Center for Social Research and ISSS at Peking University, and the Population Studies Center (with support from the National Institute of Child Health and Human Development, R24HD041028), the Survey Research Center, and the Center for Chinese Studies of the University of Michigan. The authors are grateful to Cindy Glovinsky, Jingwei Hu, and Hongwei Xu for their assistance on earlier drafts of the paper.

Biographies

Yu Xie is Otis Dudley Duncan University Distinguished Professor of Sociology, Statistics, and Public Policy, and Research Professor at ISR, University of Michigan and Visiting Chair Professor of Peking University. His main areas of interest are social stratification, demography, statistical methods, Chinese studies, and sociology of science. His recently published works include: *Marriage and Cohabitation* (University of Chicago Press 2007) with Arland Thornton and William Axinn, *Statistical Methods for Categorical Data Analysis* with Daniel Powers (Emerald 2008, second edition), and *Is American Science in Decline?* (Harvard University Press, 2012) with Alexandra Killewald.

Ping Lu is Associate Researcher of Institute of Social Science Survey, Peking University. Her main areas of interest are sampling design, data analysis, and statistical methods.

References

Ding, Hua. Technical Report CFPS-2. Institute of Social Science Survey, Peking University; 2012. “Onsite Sampling Frame Building in CFPS 2010 Baseline Survey” (in Chinese). (http://www.isss.edu.cn/cfps/wd/jsbg/2010jsbg/) [Retrieved December 19, 2014]

Lu, Ping, Xie, Yu. Technical Report CFPS-17. Institute of Social Science Survey, Peking University; 2012. “CFPS 2010 Baseline Survey Weights Calculation” (in Chinese). (http://www.isss.edu.cn/cfps/wd/jsbg/2010jsbg/) [Retrieved December 19, 2014]

Raudenbush, Stephen W., Bryk, Anthony S. Hierarchical Linear Models: Applications and Data Analysis Methods. Second. Thousand Oaks, CA: Sage Publications; 2002.

Xie, Yu. “Understanding Inequality in China” (in Chinese). Shehui. 2010; 30(3):1–20.

Xie, Yu. China Family Panel Studies User’s Manual (in Chinese). Institute of Social Science Survey; 2012. (http://www.isss.edu.cn/cfps/wd/jsbg/2010jsbg/) [Retrieved December 19, 2014]

Xie, Yu, Hannum, Emily. Regional Variation in Earnings Inequality in Reform-Era Urban China. American Journal of Sociology. 1996; 101:950–992.

Xie, Yu, Hu, Jingwei. An Introduction to the China Family Panel Studies (CFPS). Chinese Sociological Review. 2014

Xie, Yu, Hu, Jingwei, Zhang, Chunni. “The China Family Panel Studies: Design and Practice” (in Chinese). Shehui. 2014; 34(2):1–32.

Xie, Yu, Qiu, Zeqi, Lu, Ping. Technical Report CFPS-1. Institute of Social Science Survey, Peking University; 2012. Sampling Design of China Family Panel Studies. (http://www.isss.edu.cn/cfps/wd/jsbg/2010jsbg/) [Retrieved December 19, 2014]

Xie, Yu, Zhou, Xiang. Income Inequality in Today’s China. Proceedings of the National Academy of Sciences (PNAS). 2014; 111:6928–6933.

Appendix A

Sources Used for Constructing Sampling Frames

Stage One--Four Large Provinces

Main sources of information to be used to form the district- or county-level sampling frame include:

1. Provincial Statistical Yearbooks (from 2006 to 2008)
2. County/City Population Statistical Data of the People's Republic of China 2007
3. China City Statistical Yearbook 2007 – 2008
4. China County/City Social and Economic Statistical Yearbook 2008
5. Manual of Administrative Division of the People's Republic of China 2009

Stage One--Shanghai

Administrative information about administrative “street clusters” (if urban) or “towns” (if rural) in Shanghai. Provided by National Population and Family Planning Commission and Fudan University.

Stage Two--All Provinces except Shanghai

Main sources of information to be used to form sampling frame in this stage:

Current administrative data in selected counties/districts, provided by the National Population and Family Planning Commission.

Appendix B

Systematic Probability Proportional to Size (PPS) Sampling with Implicit Stratification

Whenever feasible, units in a sampling frame are sorted according to administrative boundaries and socioeconomic standing measured by a socioeconomic indicator (SEI). After units are sorted, systematic PPS sampling is applied. This sampling procedure implicitly incorporates stratification by the variables that are used in sorting the units in the sampling frame. Stratification is used to gain efficiency and to improve estimation precision.

At the first two stages, sampling occurs at the aggregate level (for example, with counties or villages). Systematic sampling of aggregate units requires that intervals representing them in a sampling frame be proportional to their population sizes. Let us use a numerical example, shown in Table 5, for illustration. In the example, there are 10 districts or counties in the province. Let M_i denote the number of permanent residents within the ith district or county. The distribution of the number of permanent residents for all ten districts or counties is given in the second column of Table 5.

Suppose that we draw a sample of three districts or counties, using PPS sampling of permanent residents in each district or county. That is, $n = 3$. We have:

$$T = \sum_{i=1}^{10} M_i = 142400, n = 3, \text{ and } K = \frac{T}{n} = 47466 \text{ (after rounding)}.$$

First, draw a random integer (R) ranging between 1 and K. In our example, $R = 22020$. Then select districts or counties containing the following three numbers (in the T column): $R = 22020$, $R + K = 22020 + 47466 = 69486$, and $R + 2K = 22020 + 2 \times 47466 = 116952$. Through this procedure, counties or districts $i = 3$, 6, and 9 are selected into the sample. Because the list of counties or districts is ordered by an SEI, representation of the selected counties/districts over the whole spectrum of the SEI variable is ensured.

Table 1

Classification of the 25 provinces covered in the study

Type of provinces	Provinces or equivalents (municipalities/autonomous regions)	Target Number of Households
Large provinces (Self-representative)	Shanghai	1,600
	Liaoning	1,600
	Henan	1,600
	Gansu	1,600
	Guangdong	1,600
Small provinces (non-self-representative)	Jiangsu, Zhejiang, Fujian, Jiangxi, Anhui, Shandong, Hebei, Shanxi, Jilin, Heilongjiang, Guangxi, Hubei, Hunan, Sichuan, Guizhou, Yunnan, Tianjin, Beijing, Chongqing, Shaanxi	8,000

Table 2

The number of districts/counties sampled in each small province

Provinces or equivalents	Number of districts/counties	Provinces or equivalents	Number of districts/counties
Beijing	1	Chongqing	2
Fujian	2	Jiangxi	3
Heilongjiang	5	Shanxi	3
Shanxi	7	Guangxi	3
Anhui	3	Hubei	3
Zhejiang	3	Yunnan	4
Tianjin	1	Guizhou	5
Jiangsu	3	Hunan	6
Jilin	3	Shandong	7
Heibei	8	Sichuan	8

Table 3

Number of households by type

Province	# Households in total	# Normal unit and commercial-and-residential units	# Multi-household units	# Multi-residence households
Anhui	13421	12906	1626	2
Guangxi	6008	5982	2	776
Hubei	22008	19807	1513	774
Jiangsu	12195	12001	849	998
Jiangxi	4592	4537	2	219
Yunnan	17284	17038	461	632
Chongqing	4202	4177	55	186
Guizhou	21332	21189	44	2427
Fujian	7995	6257	647	632
Hunan	20261	19015	173	1761
Zhejiang	8426	7983	423	582
Sichuan	18086	16771	80	1356
Beijing	5490	5374	0	0
Hebei	22266	20178	31	160
Heilongjiang	53654	41409	70	89
Jilin	18437	18360	462	61
Shandong	15312	14583	9	146
Shanxi	18957	17691	237	14
Shaanxi	7553	7230	48	0
Tianjin	4363	3850	14	9
Shanghai	87870	83629	1292	346
Henan	62063	58201	256	4334
Guangdong	62450	53874	31	4022
Gansu	40917	37868	141	194
Liaoning	90080	82451	168	403
Total	645222	592361	8634	20123

Table 4

Onsite sample size of CFPS Baseline Survey

Region	**Type**	**Expected response rate**	**Accessible Sample size (household)**
Regions with low response rate	Resident committees (central areas and suburban villages)	60%	42
	Other administrative villages	70%	36
Regions with intermediate response rate	Resident committees (central areas and suburban villages)	70%	36
	Other administrative villages	80%	32
Regions with high response rate	Resident committees	80%	32
	Administrative villages	90%	28

Note: The division of central areas and suburban villages is based on NBS urban-rural code.

Table 5

An illustrative example of PPS sampling with 10 districts or counties

district/county *i*	**population of district/county *i*, M_i**	**cumulative population, T_i**	**sampled code $R + (j - 1)K$**
1	1160	1160	
2	18160	19320	
3	8360	27680	22020
4	8840	36520	
5	12300	48820	
6	39440	88260	69486
7	12260	100520	
8	14680	115200	
9	10280	125480	116952
10	16920	142400	