User Guide for China Family Panel Studies 2020

Qiong Wu

Yan Sun

Lihong Dai

Qi Zhen

Liping Gu

Yirui Wang

Kaiwen Zhang

Ping Lv

October, 2025

This user guide draws heavily from documentation from prior years written by numerous CFPS staff members. This version was edited by Siyu Wang.

This document may be cited as follows: Wu, Q., Sun, Y., Dai, L., Zhen, Q., Gu, L., Wang, Y., Zhang, K., Lv, P. User Guide for China Family Panel Studies 2020. Institute of Social Science Survey, Peking University, October 2025.

User Guide for China Family Panel Studies 2020

The Earlier waves of the China Family Panel Studies have been described in the User's Manual and the *User Guide for China Family Panel Studies 2018.* This report includes new features that CFPS users may find helpful when analyzing data from CFPS2020.

1. General introduction

Field work for CFPS2020 began in early July 2020 and was completed by the end of 2020. Unlike the predominant use of face-to-face interviews in earlier waves, data collection for CFPS2020 relied primarily on telephone interviews due to the impact of the COVID-19 pandemic. Face-to-face interviews were conducted only when requested and permitted. There were 22,000 families eligible for interviews in 2020, including over 3,000 newly-split families. A total of 554 interviewers participated in data collection. Response rates at the household level were lower compared to previous waves. Using the total sample of over 22,000 families as the denominator, the cross-sectional response rate at the household level for this wave was 62%. Using the number of completed family roster interviews from 2018 as the denominator, the adjacent wave retention rate at the household level was 77%. Upon completion of the family roster interviews, about 66,000 family and individual interviews were generated. They included on family questionnaires, individual-level self-report questionnaires, and child proxy questionnaires. The overall response rate for these interviews was 74%, while the rate for the individual self-reports was 65%. Non-completion for the individual self-reports was due to multiple causes, such as non-completion at the family level, individual level refusals, and a lack of telephone contact information.

1.1 New features in questionnaire design

The CFPS2020 survey included a family roster questionnaire focusing on basic sociodemographic information of all family members and their relationships; a family questionnaire focusing on family income, expenditure, and assets; an individual self-report questionnaire for all respondents aged 10 and above; an individual proxy questionnaire, which is an abbreviated version of the individual self-report questionnaire; and a child proxy questionnaire for respondents aged 0-15. All questionnaires are posted on the project website

under “Documentation” in the “Questionnaires” section.

As a longitudinal survey, CFPS is intended to keep most parts of its questionnaire design consistent across waves. Therefore, the overall design of the 2020 questionnaire followed the basic structure of the 2018 version. However, the project team also incorporated the social context of the pandemic, user demand for special add-on modules, and the overall plan for questionnaire design to add new content. The main additions are as follows:

Pandemic-Related Modules: Content covered education, employment, healthcare, income, family interactions, and other modules, which assessed the scope and depth of the pandemic's impact on respondents and documented their subsequent behavioral adjustments. Users can find specific questions in the “Questionnaires” section on the project website. Related data are shared as restricted data; application procedures can be found on the “Restricted Data” page under the “Data” section on the project website.

Internet Module: Added questions related to online gaming, video watching, and online shopping behavior in the individual questionnaire; added questions related to parental intervention and online learning in the child proxy questionnaire.

Main Job Module: The basic design remained the same as previous waves, but the CFPS team added questions related to non-standard working hours, such as night shifts, weekend work, and flexible work arrangements.

Health Module: Added questions related to exercise intensity and nutritional intake.

Childhood Information Module: Added questions related to family, individual, and parental status at age 14. In addition, we made a small number of minor modifications when needed in each questionnaire. Data users are referred to the full questionnaires published in the project website for details.

1.2 Updates in the weighting variables

CFPS2020 weighting variables include both family-level and individual-level weights. Family-level weights apply to all households that contain gene members and have completed the family questionnaire. Only cross-sectional family-level weights are available, not longitudinal weights. Individual-level weights apply to gene members as defined by CFPS (including those defined at baseline and new genes generated in subsequent waves of the survey), and include both cross-sectional and longitudinal weights.

The cross-sectional individual-level weights are applicable to gene members who completed the individual interview in 2020. The basic algorithm includes the following steps: 1) Calculate the 2020 individual base weight. 2) Calculate the individual attrition weight for 2020. The individual cross-sectional non-response adjustment weight is the product of these two weights. 3) Use the demographic data (by urban/rural, age, gender) from the 2020 National Census to perform full post-stratification adjustment for the CFPS2020 sample across six sampling frames. 4) Finally, trim the weights obtained from the above three steps to avoid excessive weight fluctuation. We released the trimmed individual cross-sectional non-response adjustment weight ('rswt_natcs20n') and the individual cross-sectional post-stratification weight ('rswtps_natcs20n') in the individual and child proxy datasets, respectively. The algorithm for the non-response adjustment weight series is identical to that used for the weights released in 2018.

2. Descriptions of the public-released datasets

We integrated data from the individual self-report questionnaires and individual proxy questionnaires to form a single dataset for all individuals aged 10 and above. Data from the remaining questionnaires (family roster questionnaire, family questionnaire, and child proxy questionnaire) were each placed in a separate dataset. The corresponding codebooks for each dataset are posted under the "Data Description" section within the "Documentation "on the project website. Users can consult the codebooks to understand the meaning of values in the variables. Basic information for the CFPS2020 databases is listed in Table 1.

Table 1: Basic Information of CFPS2020 Public-Use Datasets

Questionnaires	Dataset Name	N	Number of Variables
Family Roster	Famconf	51291	297
Family	Famecon	11620	327
Individual self-report + Individual proxy report	Person	28529	1272
Child proxy report	Childproxy	6985	293

2.1 Family roster database (famconf)

The released 2020 family roster dataset includes 51,291 individual observations from 13,017 families completing the family roster questionnaires in 2020. Gene members account for 84.2% of the sample, core members for 11.1%, and non-core members for 4.7%. The family roster dataset is structured with one line per family member as defined by CFPS, identified by 'pid'. This includes gene members from the 2010 baseline and new family members added in subsequent survey years, along with information on their relatives. Relatives include spouses ('_s' series variables), fathers ('_f' series variables), mothers ('_m' series variables), and up to 10 children ('_c1-_c10'series variables), containing their basic information.

Members of the same household share the same family ID, 'fid20'. Individuals with 'co_a20_p = 1' are defined as members of that family in the current wave. Those with 'co_a20_p = 0' indicate individuals who previously belonged to the 'fid20' household but, for various reasons (financial independence, divorce, deceased, etc.), no longer belong to the current 'fid20' household in this wave, and the family unit where these members currently reside was not successfully interviewed in 2020. If their respective family unit was successfully interviewed, these members would be assigned a new 'fid20'. Users can trace their origin family using family IDs from previous waves ('fid18', 'fid16', etc.).

Financial connection is the basis for defining family membership. 'co_a20_p' indicates the financial connection between the member and 'fid20'. Additionally, the variable 'tb6_a20_p' indicates whether the member physically resides at the household address of 'fid20'. In the CFPS design, members who belong to the same family economically can live at different addresses.

CFPS2020 followed the same rules as the 'TB601_A18*' series variables (reason for leaving home) in the 2018 family roster dataset to generate the corresponding 'TB601_A20*' series. The TB601 variable is based on question A3 in the family roster questionnaire. Open-ended text information collected for category '77' (77. Other reasons [please record respondent's words]) was coded, adding the following categories to the original ones: leaving the country, going to work, going to school, getting divorced, getting married, visiting relatives, separating from family, moving, and going for medical treatment. Apart from this, the variable structure of the CFPS2020 family roster dataset remains consistent with CFPS2018.

For ease of use, we added a few more composite variables in the CFPS2020 family roster

dataset. Appendix 1 lists the variable names and relevant information.

2.2 Family dataset (famecon)

The family dataset is at the household level, with 'fid20', 'as the unique identifier for each family. As mentioned earlier, the family IDs from previous years are identified by 'fid10' - 'fid18'. For families that did not experience splitting between waves, their family ID remains unchanged. However, if a split occurred between waves, the family ID for members still considered part of the "original family" remains the same, while only the family ID for members considered a "new family" changes. Therefore, we cannot rely solely on the family ID to determine if a household's structure has changed or not. Accurately assessing whether the structure of a family is identical to its status in the previous wave relies on comparing the composition of family members in the family roster dataset.

The sample in the family dataset includes original families defined in previous surveys and new families derived in the 2020 survey due to reasons such as marital dissolution or the financial independence of children. The interview mode was face-to-face or telephone (identified by 'iwmode'). The questionnaire content was largely the same for both modes. Due to the pandemic, the proportion of telephone interviews for the CFPS2020 family dataset was 89%, much higher than previous waves.

In the CFPS2020 family dataset, two variables are related to family size: 'familysize20', and 'fml_count'. 'familysize20' is generated by the CFPS data management team after the survey, based on information in the family roster dataset after data cleaning, confirmed by the number of members belonging to the same family (same fid20, and co_a20_p=1). 'fml_count' is a real-time loading variable during the survey, directly loaded from the family roster questionnaire, representing the number of family members as defined by the survey system at the time of the interview. This number served as the reference for the respondent answering the family questionnaire. The two variables are identical in 95.7% of family.

Many questions in the family dataset are related to household finances, e.g., incomes, expenditures, and assets. Two issues are prominent with these variables. First, compared with other items, item-level non-responses for finance questions are relatively high. For some questions, the CFPS arranged unfolding brackets questions following a non-response. For example, if the value for total expenditure is missing, the family respondent would be asked the following question:

*FEXPUB: In the past 12 months, was your family's total expenditure higher/lower than (10,000 / 25,000 / **50,000** / 100,000 / 250,000) yuan?*

The series of questions would start with the middle value in bold. If the respondent answered yes (meaning the total expenditure is higher than 50,000), the series would proceed to the next higher value (100,000) combined with "higher". If the respondent answered no, the series would proceed to the next lower value (25,000) combined with "lower". This process forms a bracket with upper and lower bounds, or reaches the boundary values (i.e., higher than 250,000 or lower than 10,000). The released datasets contain information on these bounds.

The second common issue is the misinterpretation of monetary units in the financial questions, most frequently confusing Yuan with 10,000 Yuan. Such misinterpretation would lead to either reporting error (when respondents mistook one unit for the other) or recording error (when interviewers misheard the units). To minimize such errors, we took a series of steps for variables involving monetary values. We first identified outliers based on their distributions within the same wave and comparisons with previous waves. Then, we tried to verify their credibility using information from other auxiliary variables (e.g., location information to verify house values, income and expenditure to verify assets, etc.). Additionally, we drew upon para-data, such as audio recordings of the interview process, to minimize possible recording errors. Variables that underwent this process included the following: FM401, FQ5, FQ6, FR2, FT301, FT302, FT401N, etc. We updated 468 entries based on interviewer feedback.

2.3 Individual dataset (person)

Consistent with the CFPS2018 design, the CFPS2020 individual dataset includes questionnaire data for all individuals aged 10 and above. Individuals are uniquely identified by 'pid', which remains constant across datasets and years. In addition to the personal identifier 'pid', the person dataset also includes the family ID 'fid20'. Users may link individual level data with family level data using 'pid' as the linking variable.

The CFPS2020 individual dataset contains samples from self-reports, individual proxy reports (for those unable to complete self-reports due to physical reasons), and family proxy reports (for individuals in split-off families, usually answered uniformly by the respondent of the family roster questionnaire). Interview modes were face-to-face and telephone (identified by 'self_iwmode' and 'proxy_iwmode'). Due to the pandemic, the telephone interview proportion for the CFPS2020 person dataset was 87.8%. The main difference between face-to-

face and telephone self-reports lies in the cognitive module: face-to-face interviews included memory and numeracy tests, which were absent from telephone interviews; the interviewer observation module also applied only to face-to-face respondents. Other questions were unified across modes. Self-report samples are identified by the variable 'selfrpt'; when 'selfrpt = 1', the sample is a self-report. In the individual dataset, we use the variable 'PROXYTYPE' to identify whether the sample comes from a family proxy report or an individual proxy report. Note that family proxy samples may come from members of the origin family, while individual proxy reports are usually provided by members of the current family. When integrating self-report and proxy data, we compared questions from both questionnaires. If the question wording was identical across modes, we unified the variable names (using the self-report variable name) and treated them as the same variable. If the question wording differed, we kept the respective variables for self-report and proxy.

For ease of use, we integrated the gender and age variables from the original dataset. The original dataset contained 'GENDER_PRE' and 'GENDER_UPDATE' variables, representing the loaded variable based on prior data or the family roster, and the updated variable via the interview, respectively. Based on this, we generated a composite 'GENDER' variable, incorporating cleaned content, as well as information supplemented from cross-year data and the family roster. Age-related variables include 'AGE', 'IBIRTHY', and 'IBIRTHY_UPDATE', representing age calculated from birth year, loaded birth year information, and updated birth year information, respectively. Missing age values were also imputed based on cross-year datasets and the family roster, all supplemented into the 'AGE' variable, making 'AGE' the most complete source for age information.

For ease of use, the following composite variables were added to the person dataset.

Wage Income from Employed Work ('emp_income'): The 'emp_income' variable in the person dataset is based on system-generated data. The basic algorithm is 'incomeA' (total wage income from general work) + 'incomeB' (wage income from the main job), meaning 'emp_income' primarily reflects the individual's wage income in the past 12 months (covering work up to the most recent year of the interview).

Cognitive Function: As cognitive tests were only administered to face-to-face interviewees, the number of samples completing cognitive tests in CFPS2020 was significantly reduced compared to previous years. The CFPS2020 cognitive test design was the same as CFPS2016, adopting the word recall and numeracy tests from the U.S. Health and Retirement Study (HRS).

The word recall test consists of 4 parallel sets of similar difficulty; respondents were randomly assigned one set for memory assessment. Related test scores include: IWR1 (Immediate Word Recall: Round 1), IWR2 (Immediate Word Recall: Round 2), IWR (Immediate Word Recall: Rounds 1 & 2 combined), and DWR (Delayed Word Recall). The numeracy test uses a two-stage adaptive testing method. Specific test details can be found in Technical Report CFPS-31. Numeracy test scores are provided in two forms: NS_G (Guttman scale) and NS_W (W-score). The variable 'NS_WSE' is also generated, representing the standard error corresponding to 'NS_W'. Specific variable descriptions and scoring methods can be found in Technical Report CFPS-35.

Depressive symptoms (CES-D): CFPS2020 used the Center for Epidemiologic Studies Depression Scale (CES-D) to measure individuals' depression levels. Consistent with the CFPS2018 design, a simplified 8-item version, CES-D8, was generated. The CES-D8 score is calculated by reverse-coding two negatively phrased items and then summing all eight items. CESD8 represents the score based on the 8-item version. Users can handle missing data for CESD as needed for their research. The 2020 dataset retains the raw item scores, allowing users to generate their own scores if preferred. The released dataset also provides the 'CESD20sc' variable, representing a score equivalent to the 20-item version, constructed based on these eight items.

Education variables: The skip logic in the education module is complex. We integrated loaded data and information from different modules to generate four education-related composite variables: the highest completed education level ('CFPS2020EDU'), school departure/current attendance stage ('CFPS2020SCH'), total completed years of education ('CFPS2020EDUY'), and total completed years of education after imputation ('CFPS2020EDUY_IM'). In generating the education composite variables, we first classify respondents into different groups based on the questionnaire skip logic to distinguish between those currently enrolled in school and those who have left. Following the initial compilation of the highest education level and school departure/attendance status, we conduct a logic check and correction of the relationship between these two variables. As a general principle, the school departure/attendance status should be at or above the highest education level obtained. Additionally, we supplement and adjust the 2020 education data by referencing respondents' historical education composite variables from previous survey waves, aiming to achieve logical consistency across years. Building on the cleaned versions of these two variables, we then calculate respondents' total completed years of education. When based on the highest education

level, the years of education equal the number of years corresponding to that degree. When based on the school departure/attendance status, it equals the years associated with the preceding educational stage plus the completed years in any subsequent, unfinished educational stage. If the number of years completed in the unfinished stage is missing, but both the highest education level and school departure/attendance status are available, the resulting years of education variable may contain missing values. We impute these missing values using a hot deck method, creating an imputed version of the education years variable (CFPS2020EDUY_IM). Specifically, we identify the immediately preceding respondent in the sorted list of personal IDs (pid) who shares the same highest education level and also did not complete that stage, but for whom the years of schooling in that stage are not missing, and use their value for years of schooling.

Confirming Employment Status (‘employ’): For details on the design of employment/ unemployment status, users can refer to the "Employment Status Definition" chapter in the "User's Manual" on the project website. Statistically, the definition of unemployment in CFPS differs systematically from the surveyed unemployment rate published by the National Bureau of Statistics. The employ variable for self-respondents is automatically generated by the system based on responses from the [GB Confirming Employment Status] module. The algorithm is as follows: (1) If the respondent met at least one of the following criteria: (i) worked for at least one hour in the past week, (ii) could return to their original job within a definite period or within six months, (iii) was self-employed but currently in an off-season with intention to resume operations, or (iv) was engaged in agricultural work but currently in the slack season, they were classified as employed (employ=1). (2) If the respondent had looked for work in the past month and could start work within two weeks if a job were available, they were classified as unemployed (employ=0). (3) If the respondent had not looked for work in the past month, or could not start work within two weeks if a job were available, they were classified as out of the labor force (employ=3). (4) For all other cases, employ was set to -8.

The individual proxy and family proxy questionnaires did not implement the full [GB Confirming Employment Status] module. However, we supplemented the current employment status for some proxy respondents based on information from item GB1. If the proxy respondent reported working for at least one hour in the past week, the individual was classified as employed (employ=1). In all other cases, due to insufficient information to determine whether the individual was employed, unemployed, or out of the labor force, employ was assigned a value of -10 (Cannot be determined). If the original value of the relevant question

was -8, employ was set to -8.

2.4 Child proxy dataset (childproxy)

The child proxy dataset contains proxy reports from guardians of children aged 0 to 15. Each line represents a child, uniquely identified by 'pid'. Proxy respondent is identified by respc1pid. The proxy respondent is the guardian of the child, often the child's mother or father, and sometimes grandparents or other family members. Consistent with the CFPS2018 setup, the CFPS2020 child proxy dataset contains only guardian proxy data; self-report data for children aged 10-15 are part of the individual dataset.

The child proxy dataset includes child proxy samples, supplemented by a minor proportion of children aged 0-15 from household proxy reports. The dataset uses the identifier variable PROXYTYPE to distinguish between child proxy reports (3) and household proxy reports (2). Since household proxy reports utilize the individual proxy questionnaire, the child proxy questionnaire contains more comprehensive content than the household proxy version. The interview modes were face-to-face (IWmode=1) and telephone (IWmode=2).

2.5 Cross-wave individual core variable dataset (crossyearid)

For all individual samples that ever entered CFPS through the family roster questionnaire, the cross-year ID dataset provides their basic information across all waves since the baseline. A total of 76,982 individuals are in the cross-year ID dataset, including 65,612 gene members, 8，017 members who were core members in at least one wave, and 3,353 non-core members in at least one wave.

Variables in the cross-wave individual dataset are mainly in three categories. The first category is time-constant demographic variables, including 'pid', birth year, gender, ethnicity, and baseline sampling information. They are considered constant across waves in the study design. The second category is time-varying sociodemographic variables, including marital status, education, and household registration (hukou) status. These may change across waves, and thus have a different variable for each wave. The third category refers to interview status, including the entry year of the individual sample, the family ID ('fid') for each wave, whether the individual was financially connected with the corresponding 'fid' of a particular wave, whether the individual lived at the family residence address, whether the individual completed an individual survey, and whether the individual survey was a self-report.

Appendix 1. Variables in the CFPS2020 Family Roster Dataset and Their Corresponding Items in the Questionnaire

Variable name	Variable label	Questionnaire Item	Notes
FID_PROVCD20	Province ID 2020		Provincial GB code cleaned from the address module
FID_COUNTYID20	County ID 2020		Recoded after cleaning the county GB code from the address module
FID_CID20	Community ID 2020		Recoded after cleaning the village/neighborhood community GB code from the address module
FID_URBAN20	Urban area (Census Bureau's definition)		Classification of village/neighborhood community attribute based on NBS urban-rural type for the household's location
SUBSAMPLE	Is it in the national resampling sample?		Defined by subsample status linked with fid_base (defined later)
SUBPOPULATI ON	Sampling subpopulation		Defined by subpopulation status linked with fid_base (defined later)

Variable name	Variable label	Questionnaire Item	Notes
GENETYPE20	Gene Type 2020		Recoded based on the 2020 family member type table and gene member status
fid1*	Family ID for 2010-2018		Household affiliation for each respective year
FAMILYSIZE20	Number of Family Members		The total number of individuals within the same 'fid20''where 'co_a20_p=1'
TB2_A_P	Gender	BC2, E1, D105	Synthesis of newly collected info from the member questionnaire, info collected in the individual questionnaire, and pre-existing gender information
TB1Y_A_P	Birth Year	BC3, E2, D104	Synthesis of newly collected info from the member questionnaire, info collected in the individual questionnaire, and pre-existing birth year information
TB1M_A_P	Birth Month	BC3, E2, D104	Synthesis of newly collected info from the

Variable name	Variable label	Questionnaire Item	Notes
			member questionnaire, info collected in the individual questionnaire, and pre-existing birth month information
TB3_A20_P	Marital Status	BC4, E3	Synthesis of newly collected info from the member questionnaire, info collected in the individual questionnaire, and pre-existing marital status information
TB4_A20_P	Highest Education	BC5, E4	Synthesis of newly collected info from the member questionnaire, info collected in the individual questionnaire, and pre-existing highest education information
HUKOU_A20_P	Hukou Status	BC6, E5, D106	Synthesis of newly collected info from the member questionnaire, info collected in the individual questionnaire, and pre-existing hukou information
TB6_A20_P	Currently Lives in	A2, A201	Prioritizes the leaver's own

Variable name	Variable label	Questionnaire Item	Notes
	this Household		subjective judgment of economic independence; supplemented by the origin family's subjective judgment
CO_A20_P	Economically Connected with this Household	F102, B1	Integrates information on single-person leavers and multi-person leavers from the member questionnaire
OUTPERS_R_WHERE20_P	Region of Residence (Leaver)	G1, H1	Integrates information on single-person leavers and multi-person leavers from the member questionnaire
TB601_A20_P	Reason for Leaving Home (Leaver)	A103	
TB602ACODE_A20_P	Province Code (Leaver)	G101, H101	Integrates province information for single-person and multi-person leavers from the member questionnaire. Values (1-6) retain original options; text from option 77 was categorized and recoded (10-18)
OUTUNIT20	Serial Number of	F1	Assigns leavers from the

Variable name	Variable label	Questionnaire Item	Notes
	Leaving Unit		origin household to different units sequentially based on the number of addresses reported
COREMEMBER20	Is Core Member	"CFPS Family Member Type" table	Value 4 in the table, and core members from previous years
CFPS2020_INTERV_P	Individual Survey Completed in this Wave		Aggregates completion status of various individual questionnaires
ALIVE_A20_P	Is Alive	A3	Synthesis of newly reported death information and pre-existing death information
TA4Y_A20_P	Year of Death	A4	Newly collected year of death information from the member questionnaire
TA4M_A20_P	Month of Death	A4	Newly collected month of death information from the member questionnaire.
TA401_A20_P	Cause of Death	A401	Newly collected cause of death information from the member questionnaire.

Variable name	Variable label	Questionnaire Item	Notes
pid_a_*	'pid' of Father, Mother, Spouse, 10 Children	C2, C3, C4, C5	Integration of newly collected 2020 information and historical family relationship information
C105_A20_P	Reason for Joining the Family	C105	
RTYPE_END20	Meaning of 'rtype' in the Questionnaire		See explanation in the "CFPS Family Member Type" section of the family roster questionnaire
FID_BASE	Baseline Family ID		The origin household at the 2010 baseline survey to which the current 'fid20''traces back
PSU	Primary Sampling Unit (Baseline)		The PSU corresponding to the baseline household during the 2010 baseline sampling
C105_A20_P	Reason for the individual joining the family		C105
ADS1_20	Moved or Not	ADS1	
KZ103_20	Main Language	Z103	

Variable name	Variable label	Questionnaire Item	Notes
	Used in Interview		
INTERVIEWERID20	Interviewer ID		Executive-level information
interrupt	Is Interrupted Sample		Whether the questionnaire was completed