

编者按：近年来，有多项社会调查成为了解社会发展，掌握国情民意的重要工具，特别是成为社会科学研究的重要数据来源。我们发现，部分研究对相关调查的方法了解不充分、指标使用不规范、数据处理不恰当，存在对调查数据的误读和误解，进而影响研究的可靠性，伤害研究结果的有效性。为此，《调研世界》编辑部从 2023 年开始，与各主要调查机构进行联系，希望较系统地对相关调查进行解读。从 2024 年 1 月起，《调研世界》开设专栏，陆续刊发调查解读，以帮助感兴趣的研究者深入了解各调查项目的设计理念、调查方法与特点，以及使用调查数据的注意事项，意在助力相关学术研究高质量开展。

中国家庭追踪调查： 设计理念及数据运用问题

孙 妍 吴 琼 张春泥*

内容摘要：中国家庭追踪调查（CFPS）作为我国首创的全国性、大规模、多学科的长期固定样本追踪调查为研究中国变迁提供了宝贵资料。该项目 2010 年启动全国首轮基线调查，调查样本覆盖全国 25 个省（区、市）的 14960 户家庭。基线调查后，项目保持隔年回访、双年采集数据的追踪模式，截至 2022 年，CFPS 已完成七轮全国调查，前六轮（2010—2020）的数据已面向公众开放。作为学术和政策研究的重要资料，CFPS 注册用户已超过 12 万，每天都有基于 CFPS 的科研成果产出。但 CFPS 在使用上具有一定的门槛，不了解项目理念可能导致数据误用，进而影响研究结果的有效性。为了让研究者能够科学、规范地使用 CFPS 数据，本文将结合 CFPS 在调查设计和管理上的四个特点——综合性、嵌套性、追踪性和规范性，分别解读 CFPS 的设计和管理理念，并对用户在数据使用过程中的常见问题及误区进行回应。

关键词：CFPS；追踪调查；设计理念；数据使用误区

中图分类号：C31 **文献标识码：**A **文章编号：**1004-7794(2024)01-0004-11

DOI: 10.13778/j.cnki.11-3705/c.2024.01.001

* 孙妍，2009 年毕业于北京林业大学，获管理学博士学位，现为北京大学中国社会科学调查中心副研究员，研究方向为社会调查方法。吴琼（通讯作者），2009 年毕业于美国宾州州立大学，获教育与心理测量学博士学位，现为北京大学中国社会科学调查中心副研究员，研究方向为社会学量化研究方法，邮箱：qw@pku.edu.cn。张春泥，2013 年毕业于香港中文大学，获社会学专业博士学位，现为北京大学社会学系长聘副教授，研究方向为社会分层与不平等研究、社会人口学。本研究得到教育部重大课题攻关项目“新中国成立以来社会调查发展进程、经验与展望研究”（20JZD032）的资助。

一、引言

社会调查是向现实社会中的研究对象有目的地、系统性地采集信息和数据的活动。放眼当今世界，社会调查已成为现代化国家了解社会趋势、掌握国情民意、进行学术研究和政策决策的重要工具。我国社会调查的发展也具有应时代之需、经世济民的特点。自 20 世纪 80 年代以来，学术机构、政府机构开始采用现代社会调查的手段了解改革时期的社会特征与社会问题^[1]。21 世纪以来，伴随着社会变迁渗透到各个领域，我国的社会调查也进入了快速发展的阶段。中国家庭追踪调查（China Family Panel Studies，以下简称 CFPS）正是在这一背景下应运而生。

CFPS 是由北京大学中国社会科学调查中心设计和实施的全国性、综合性的长期固定样本追踪调查，旨在通过追踪收集个体、家庭、社区三个层次的数据，反映中国人口、经济、社会等多方面的变迁，意在为现在及将来的研究者提供了解和研究当代中国较为翔实、客观的一手数据^[2]。CFPS 在 2010 年启动全国首轮基线调查，调查样本覆盖全国（不含香港、澳门、台湾、新疆、西藏、青海、内蒙古、宁夏、海南）的 25 个省（区、市）的 14960 户家庭。基线调查后，项目保持隔年回访、双年采集数据的追踪模式，持续回访基线调查中界定出的所有家庭成员及其所在家庭。截至 2022 年，CFPS 已完成七轮全国调查，前六轮（2010—2020）的数据已在官网公开发布。截至 2023 年，CFPS 的注册用户已超过 12 万人次，该调查成为国内学术界最具影响力的数据来源之一。

随着用户数增长、用户来源多样化，引导用户正确使用和深度开发数据成为一项重要议题。为了让研究者能够科学、规范地使用 CFPS 数据，本文将结合 CFPS 在调查设计和管理上的四个特点——综合性、嵌套性、追踪性和规范性，分别解读 CFPS 的设计和管理理念，并对用户在数据使用过程中的常见问题及误区进行回应。

二、综合性理念及其运用

综合性是 CFPS 的一个重要特点。一方面，综合性意味着调查内容包罗万象，CFPS 访题覆盖的主要领域包括就业与经济、婚姻与家庭、儿童发展、教育、养老与代际关系、健康、心理与主观福祉，旨在兼顾众多学科及领域的研究需要；另一方面，综合性意味着调查内容适用于不同社会人群，CFPS 的受访者覆盖全年龄段、城乡不同阶层的人口，对各类群体所关注的调查内容各有侧重。在综合性的理念下，问卷设计如何容纳各类主题的访题，让访题适用于各类群体的实际情况，并能够全面、综合地反映社会事实？为此，CFPS 在问卷设计上采用以分类分时点采集的方法来决定历次调查内容，以定制化的方式来对不同人群形成不同的采集内容，用标准化采集的方式来提高数据在不同研究场景下的可比性。

（一）以综合性为理念的问卷设计特征

1. 分类分时点采集。

随着时间推移及与各学科研究需求的持续对接，CFPS 的调查内容不断丰富，积累的访题越来越多，产生了拓展调查内容的需要。这一背景下，如何在有限的调查访问时间下合理安排问卷采集的内容？结合追踪调查固定样本重复观测的特点，CFPS 将调查信息分为基线信息、核心信息及扩展信息三类，针对不同类型的信息采取不同的数据采集策略。其中，基线信息属于 CFPS 数据库中每个样本不可或缺的基础内容，这些内容具有不会重复发生且不可改变的特性，例如，出生年月、性别、初婚情况、第一份工作经历都是典型的基线信息。对此，项目采用一次性提问的方式，当样本首次回答 CFPS 问卷时或首次发生某种特定行为时进行采集，之后如无特殊原因不做重复采集。核心信息是 CFPS 测量社会现象变迁的核心指标，是历年调查问卷的主体内容。

核心信息具有可变的数据特性，如受教育经历、工作经历、婚姻史、主观态度等会随着受访者不同人生阶段而变化，对此，CFPS 在不同轮次尽可能运用相同的访题进行重复采集，以供用户构建可比的研究指标。扩展信息通常具有较强的时点属性，可根据特定时期的研究需要而增加，在整个项目实施中仅在某个轮次中出现，不作为追踪调查的内容。

2. 定制化采集。

鉴于 CFPS 受访者构成的高度异质性，CFPS 借助计算机辅助调查模式，让计算机系统根据每名受访者的不同情况定制访题，形成不同的问卷路径和模块组合。决定受访者在访问中的答题路径或问卷内部逻辑跳转的因素有以下四个方面。一是样本的属性。CFPS 针对不同类型人员设定不同层级的信息采集需求，基因成员^①的信息采集最全面，覆盖问卷中所有适用的调查指标；对家庭成员中非基因的直系亲属，重点采集其在调查时点的信息，不采集教育史、迁移史等需要较长回溯时段的信息。二是当轮调查前已有数据的完整程度。对于首次访问的初访样本，需要先补齐基线信息，在此基础上再补齐特定时段内的回溯信息；对于追踪样本，通常以最近一轮调查的状态为出发点，采集两轮调查间的变化。三是样本的前序回答。根据本轮访问中受访者对前序访题的回答，动态决定后续访题是否需要展开提问。例如，针对在学人群提问上学模块访题；在学人群中，仅对高中及以下阶段的学生提问辅导班的参与情况等。四是轮替及随机化设计。为最大限度降低受访者在单轮访问中的负担，CFPS 采用轮替及随机化的技术指定某类访题的适用对象。轮替策略主要运用于主观评价题组，根据调查轮次或受访者年龄确定某一题组是否需要被提问。由于 CFPS 主观评价题较多，在不同年份会启用不同的主观评价题组，部分题组会间隔 1~2 轮（或更长周期）提问一次。与年龄相关的主观题组则采用固定年龄组采集的模式，仅向访问时符合特定年龄的受访对象提问。

3. 标准化采集。

CFPS 的数据收集依靠受访者自报的信息，访题设计和受访者对访题的理解差异均会影响数据采集的准确性。为了最大限度地保证跨轮调查间的数据可比性，CFPS 尽量使用标准题项采集客观数据，也会根据不同类型信息的可及性分别向不同受访对象采集。例如，就业状态是许多研究领域普遍关注的变量，相较于直接提问受访者“当前是否有工作”，CFPS 从 2012 年之后使用了国际劳工组织（ILO）建议的一系列客观访题，通过 7 道是否题的回答综合判断受访者当前就业状态。这一方式有效地减少了由于受访者不了解就业状态界定标准而导致的错报。家庭总收入是研究者常用的衡量家庭经济状况的变量，但很少有受访者能够在调查时直接给出现成的年收入总额，且直接提问总额的方式会因受访者对“收入”概念内涵的理解不同或记忆偏差导致与实际情况存在较大偏误。对此，CFPS 在家庭经济问卷中将家庭收入分为经营性收入、工资性收入、转移性收入、财产性收入、其他收入五类，先分项采集、后汇总确认，帮助受访者明确概念具体内涵的同时辅助回忆，避免遗漏或重复。例如，调查工资性收入时，家庭经济问卷中会首先识别出从事受雇工作的家庭成员名单，然后提问家庭中这些成员的该项收入之和。考虑到工资性收入具有较强的个人属性，除在家庭经济问卷中汇总提问外，在个人自答问卷的工作模块还会详细询问个人的工资水平及各项福利收入，以此适度纠正家庭经济回答人因未全面掌握信息而导致的数

（二）相关数据库使用的注意事项

上述以综合性为特征的问卷所采集的数据会包括大量的逻辑跳转，这意味着用户需要熟悉不同轮次调查的分类分时点采集安排，认真阅读问卷标注，了解模块进入条件及访题前后的逻辑跳转，以确定某个访题的回答对象属性。由于调查内容综合了不同来源的信息，也意味着受访者在使用数

^① 基因成员是指 2010 年基线调查界定出的所有家庭成员及其今后新生的血缘/领养子女，下文将有详述。

据时要对信息进行重组,根据研究需要构造变量。以下两类用户关注的问题就体现了上述两个要求。

1. 缺失值处理。

刚开始接触 CFPS 的数据用户可能感到最困惑的一个问题就是不少变量存在大量的缺失值,而且这些缺失值以不同的数值标注。据前文介绍,CFPS 问卷中呈现的访题并非适用于所有受访对象,而是系统根据不同受访对象的不同状态动态定制访问内容、生成访题,由于这类因逻辑跳转导致的数据缺失属于系统性缺失,在变量取值时统一用“-8”表示。以家庭经济库 FM4 变量(该经营的净利润(元))为例,在 2020 年数据中,有 10541 名受访者(占样本 90.7%)在该变量上的取值是“-8”,这是由于这部分受访者的所在家庭未从事经营活动而直接跳过该题。此外,在访问中,受访者也可能在某些问题上由于缺乏答题所需知识或缺乏回答意愿而回答“不知道”或“拒绝回答”,这两类属于受访者原因造成的缺失,在变量取值上分别用“-1”和“-2”表示。对于这些缺失值编码,如用户使用时不做处理,直接将-1、-2 和-8 与其他有效数值放在一起分析,不仅影响有效样本数的识别,还会导致分析结果出现严重偏误。

因此,当一个变量存在缺失值时,用户首先需要结合问卷了解缺失值产生的原因,避免盲盒式操作。如果是受访者原因造成的缺失值,可以根据缺失值比例及具体研究需求选择合适的缺失值处理方法。CFPS 用户常用的缺失值处理方法包括整行观测删除、均值或回归方式插补、hotdeck 插补等,也有用户使用更复杂的链式方程方法插补^[4-5]。如果是逻辑跳转形成的系统性缺失,用户需要厘清问卷设计逻辑,从缺失原因入手处理缺失值。典型的处理系统性缺失数据的方法有以下几类:第一,由于没有参与某类活动而跳过具体访题的,相关变量可以直接取零值。例如,受访家庭因为没有参与经营活动而跳过 FM4(该经营的净利润(元))这道题时,这些家庭的经营利润实质上是零值,应结合前序筛选题将“-8”赋值为 0。第二,由于样本不符合筛选条件而跳过后续访题。例如,2020 年 CFPS 个人自答问卷中的 QM801(婚姻满意度),其跳转进入条件是当前婚姻状态为“有配偶”或“同居”(即 EA0=2 或 3)。对于非在婚或同居的人群,QM801 取值为-8。由于婚姻满意度对这类人群不适用,因此存在此类系统性缺失的样本可以整条删除。第三,对于追访样本,之前轮次已经采集过的基线信息不再重复采集,例如,QA701(民族成分)的跳转进入条件是 minzu 变量为空,即往期调查尚未采集到该受访者的有效民族信息。如若往期调查已采集过此信息,本轮调查会直接跳过该题并赋值-8。这种情况需要用户结合更早轮次的 CFPS 数据,找到相关数值进行填充。

2. 综合变量的应用。

上文介绍了 CFPS 尽量使用标准题项采集客观数据的原因。标准化采集使 CFPS 中很多变量的信息以间接、分项的方式获得。例如,家庭年收入分为五大类收入分别提问,这种方式便于采集到准确度更高的数据,但对用户却意味着问卷中没有单独一个在研究中能够直接使用的“家庭总收入”的变量。为了便于用户使用,CFPS 项目组根据问卷设计逻辑构建了一些常用的综合变量。例如,2020 年教育相关的综合变量(见表 1)用到了教育史模块原始采集的最高学历(w01),还用到了问卷中“上学确认”等其他模块、关系库学历信息以及往期调查中所采集的受访者教育信息。综合变量 cfps2020edu 是将当前轮次 w01 所获得的数据与往期调查受访者最高学历数据相融合所生成的、截至本轮调查受访者已完成的最高学历综合变量,其缺失值只有不到 1%。综合变量 cfps2020sch 反映的是受访者截至当前轮次最后的离校或在校阶段。对于当前上学人群来说,cfps2020sch 数值一般要高于 cfps2020edu,例如,在读大学本科的受访者 cfps2020edu 取值是“高中”,而其 cfps2020sch 取值则是“本科”。cfps2020eduy 反映了受访者受教育总年限,例如,大部分本科毕业生教育年限为 16。若用户不了解访题在问卷中的位置及前序条件,仅通过关键词搜索标签为“最高学历”的变量,只会找到 w01,而这道访题只对往期调查未采集最高

学历或 45 岁以下的受访者提问，因此约有一半的样本该变量取值显示为“-8 不适用”。

项目组已在官网发布综合变量的生成算法（参见综合变量查询表）。绝大部分情况下，项目组推荐用户直接使用项目组在公开数据集中提供的综合变量。用户若希望自己生成更适用于研究的变量，可以查看公开数据集中的原始问卷变量，也可以结合各轮次的清理技术报告了解更多综合变量信息。

表 1 CFPS2020 教育相关变量

变量名	变量标签	解释	类型
w01	最高学历	补全或更新最高学历	问卷变量
cfps2020edu	已完成的最高学历	最高学历综合变量	综合变量
cfps2020sch	离校/上学阶段	在学阶段综合变量	综合变量
cfps2020eduy	受教育年限	受教育年限综合变量	综合变量

三、嵌套性理念及其运用

中国社会具有多层次结构的特点，体现为个体嵌套在层层的社会组织（如工作单位、村庄、家庭等）中。在各类嵌套结构中，家庭是最基础、最直接影响个人角色、地位、行为和态度的社会组织^[3]。因此，CFPS 以个人和家庭为研究主体，关注研究主体的嵌套性，充分考虑了社会现象的异质性和复杂性。这一理念体现在嵌套式的数据采集策略中。

（一）以嵌套性为理念的数据结构设计

1. 个人—家庭—社区三级嵌套关系。

社区、家庭和个人是 CFPS 调查的三个重要层次。社区层次调查关注受访者家庭所在社区的政治环境、村/居面貌、基础设施、人口、资源、交通、医疗卫生和财政收支等，CFPS 在 2010 年和 2014 年通过社区问卷形式向家庭所在村委会或居委会干部采集信息；在后续年份为适应样本家庭日益分散的特点将社区关键指标纳入家庭及个人问卷中。家庭层次调查采集家庭结构与成员关系、生活条件、社会交往、收入支出和资产状况等，该层包括两套问卷，一是针对受访家庭人口结构的家庭成员问卷，二是采集家庭经济活动、财富状况、生活条件和以家庭为单位社会交往的家庭经济问卷，这两套问卷通常会请家中熟悉相关情况的成员来完成。个人层次调查则详细了解每位受访者的教育、职业、收入、婚姻、心理与身体状况、观念与态度等，通常由受访者本人回答，对 10 岁以下的未成年人，由其主要抚养人代答。CFPS 通过三级嵌套关系的设计，使得社区、家庭和个人三者不再是孤立的，能够在数据上关联起来。

2. 个体成员与家庭关系相嵌套的家户全员数据采集。

不同于大多数其他入户调查只在抽中家户中随机抽取一人访问，CFPS 对受访家庭中所有基因成员及其直系亲属均进行个人问卷的访问，实现个体成员与家庭关系相嵌套的网络结构。由于全员访问，CFPS 根据受访者年龄将需要访问的个人分为三组：0~9 岁、10~15 岁、16 岁及以上，针对不同年龄组采取不同的个人信息采集策略。0~9 岁低龄组的个人信息由其主要抚养人提供，10~15 岁青少年组的个人信息通过主要抚养人代答及受访者自答相结合的方式获得，16 岁及以上的成人组则主要采取受访者自答的方式采集信息。考虑到存在极少量成人受访者因客观条件限制（如聋哑、健康原因等）无法完成个人自答问卷，项目组设计出个人代答问卷，可由其他家庭成员来提供受访者最基本的背景信息。由此，CFPS 在个人层次形成了少儿家长代答问卷、个人自答问卷、个人代答问卷这三类问卷形式。其中，10~15 岁青少年组的问卷数据采集形式进行过调整。2016 年及之前轮次的调查，根据受访对象的类别，将少儿家长代答问卷及个人自答问卷的内容均并入少儿问卷库中统一管理；2018 年起，根据回答人的属性，分别生成少儿家长代答问卷及个人自答问卷分库管理。

（二）相关数据库使用的注意事项

由于上述嵌套关系，CFPS 在每轮都至少存在四个数据集，包括家庭关系库、家庭经济库、个人库（CFPS2018 年之前的成人库）、少儿库。家庭经济库以家庭为单位，每行代表一个家庭；其余三个库均以个人为单位，每行代表一个个体。在 2010 年、2014 年的调查中，CFPS 还采集了村居问卷，形成村居库。当用户分析使用到不同层次的变量时，需要进行多个数据集的合并与关联。对于各数据集的数据结构不够了解的用户很容易发生错误。下面分不同类型介绍 CFPS 不同数据集合并时的常见问题。

1. 个人数据一对一横向合并。

CFPS 数据采集流程是从家庭成员问卷开始，记录家庭所有成员的信息后识别出该家庭需要访问的个体，继而生成个人问卷并开展个人数据采集。CFPS 为每位受访者构建了一个唯一的识别码 pid，pid 在不同库和不同轮次间保持恒定不变，是能够作为个人层面数据集匹配的链接变量，实现数据库的一对一横向合并（merge）。

不少用户困惑于家庭关系库中的部分观测对象并不存在于个人库中，对此通常有两种情况：一是家庭成员中只有基因成员及其直系亲属（即核心成员）需要完成个人问卷，与基因成员没有直系亲属关系的其他家庭成员不需要完成个人问卷，因此这类人员不会出现在后续的个人库中。第二类情况是家庭中需要回答个人问卷的受访者可能由于联系不上、拒访等因素无法完访个人问卷，也就没有个人库数据。

2. 一对多横向合并。

一对多横向合并 CFPS 数据使用中也常见。第一个典型例子是个人数据集与家庭数据集的合并。例如，研究者要分析家庭收入对个体心理健康的影响，前者数据在家庭经济库，后者在个人库。CFPS 的嵌套设计意味着一个家庭中存在多名受访成员，这些成员共享一套家庭层面的数据，因此家庭经济数据集和个人数据集合并是一对多横向合并，合并应以唯一识别的家户号 fidXX（XX 表示相应轮次的年份，例如 fid20 是 2020 年的家户号）作为链接变量。一些用户发现在合并家庭经济库和个人库时，也会出现少量不能完全匹配的现象，如有些存在于家庭经济库中的家户号在个人库中找不到任何个人问卷，或存在于个人问卷的家户号在家庭经济库中无法找到相应的家户记录。上述这些无法完全匹配的情况仍与访问流程相关：当每个家庭的家庭成员问卷完访后，会同时生成家庭经济问卷和个人问卷，这两类问卷独立存在，需要受访者分别作答，因此会存在两类未同时完访的情况。

第二个典型例子出现在同一家庭的亲子数据匹配中。例如，如果一位母亲在家庭内有多孩子，将母亲数据匹配到少儿库中作为家长信息时就会面临一对多匹配的问题。对此，首先可以通过家庭关系库查找到每位孩子（pid）母亲的样本编码（pid_m），将母亲信息用孩子的 pid 匹配到孩子的个人数据库中。然后，在成人库中提取所需纳入孩子个人库中的母亲的特征变量，将成人库的 pid 重命名为 pid_m，以 pid_m 为链接变量与孩子个人库合并。子女和父母数据的横向合并也可能存在不完全匹配的问题：一是在家庭关系库中，有些个体的 pid_m 变量值缺失，这是由于孩子的母亲没有被认定为家庭成员且家庭成员问卷的回答人没有提供关于孩子母亲的任何信息。二是家庭关系库中虽然存在 pid_m，但是在当年没有产生其对应的个人问卷，如母亲并非 CFPS 界定的家庭成员或不满足访问条件。三是虽然系统生成了母亲的有效个人问卷，但母亲没有完成个人问卷访问。

3. 纵向合并。

横向合并的目标在于扩充变量的丰富程度，而纵向合并（append）的目标在于扩充观测数。CFPS 中纵向合并的典型例子是 2016 年及之前轮次调查中的少儿库和成人库的纵向合并，少儿库包含 0~15 岁的个体样本，成人库包含 16 岁及以上的个体样本。需要留意的是，CFPS 每一轮次

中少儿库和成人库相同变量在这两个库中的变量名并不一定相同。用户在纵向合并之前，需要对变量名进行统一化操作，避免纵向合并时出错。例如，2010 年成人库中身高的变量名是 qp1，少儿库中身高变量名是 wa104，在纵向合并之前需要让两个变量名保持一致。

从 2018 年开始，CFPS 对个人层面数据集进行了结构调整。少儿库依然覆盖 0~15 岁样本，但只包含少儿家长代答问卷的内容，命名为“少儿家长代答库”；另一个库涵盖 10 岁及以上的个人自答问卷数据以及个人无法回答而由其他家人代答的个人代答问卷数据，命名为“个人库”。如用户将少儿家长代答库和个人库直接进行纵向合并，会造成一部分 10~15 岁样本重复。对此，可以在纵向合并完成之后删除 pid 重复的样本，或在纵向合并之前只保留少儿家长代答库中 10 岁以下样本或个人库中 15 岁以上样本。无论使用哪种办法，均需在合并前完成变量名统一，且需重点关注不同库中 10~15 岁样本相同变量的数值不一致问题。

四、追踪性理念及其运用

时间性是研究社会变迁的一个重要概念。不同于大多数国内调查采用的一次性或重复截面设计，CFPS 充分考虑了社会现象的动态趋势与个体随时间变化的特征，采用了对固定样本群体在不同时点上进行重复观察的追踪设计，其在数据丰富程度上有着截面数据无法比拟的优势。但追踪调查在界定访问对象的设计上会比截面调查更为复杂，也要持续应对样本流失的挑战。在数据使用上，构造追踪数据是数据使用的难点。

（一）以追踪性为理念的固定样本的追访设计

1. 追踪策略。

CFPS 借鉴国际追踪调查项目实施的成功经验，并结合中国家庭变迁的特点，开创性地设计出一套具备样本总体自我更新的追踪策略。2010 年基线调查结束后，所有接受访问家庭中的所有家庭成员及其在基线调查后新出生的血缘子女和新领养的 10 岁以下子女被视为“CFPS 基因成员”，成为永久追踪的受访对象。后续追踪调查实施时，CFPS 基因成员所在家庭作为项目追踪的目标家庭。随着基线家庭的裂变，CFPS 追踪家庭规模整体呈现出增长的趋势。与此同时，受访家庭内部成员结构的变化也在家庭成员问卷中被记录下来，项目以此为基础，动态界定每个家庭每期个人问卷的受访对象，针对不同类型的人员采取不同的追踪策略。基因成员是 CFPS 永久追踪的对象，而基因成员当前所在家庭内的非基因直系亲属（父、母、配偶、子女）仅会成为当轮调查的访问对象，当这些非基因直系亲属迁出基因成员所在的家庭单元时则停止追踪。自我更新的追踪策略通过新家庭及新成员的纳入，一定程度上可以记录中国家庭结构的变迁，并部分抵消人口死亡、样本流失等原因引发的样本规模递减问题。

自我更新的追踪策略需要数据用户从三个方面加以理解：第一，在每个访问家庭中，并非所有的家庭成员都需要完成问卷采集详细个人信息，只有基因成员及其非基因的直系亲属（即核心成员）需要完成个人问卷。第二，并非每轮接受访问的个人样本都会成为 CFPS 永久追踪的对象。CFPS 仅对基因成员采取永久追踪的策略；对于其他非基因的直系亲属，只有当其与非基因成员处在同一家庭时才需要完成个人问卷，而当家庭关系断裂时，这部分人员将不再成为 CFPS 的追踪样本。第三，对于永久追踪的基因成员，其个人是恒定的，但其所属的家庭可能是动态变化的。

2. 追踪效果。

遵循 CFPS 追踪策略，除基线调查完访家庭会被持续追踪外，基因成员后续新组建的家庭也会被纳入到后续追踪中。表 2 中展示了以家庭层面完访为判断标准下，前五轮全国追踪家庭层面应答率。整体而言，随着时间的推移，追踪调查实施的难度逐渐加大，不可避免地存在样本流失，

家庭层面应答率均呈现出下降的趋势。2018 年之前，每轮保持 3~5% 的下降，2020 年因新冠肺炎疫情原因，访问模式由依靠实地面访为主转为电话访问为主，访问模式的转化拉低了样本应答率。

表 2CFPS 家庭层面应答及追踪率

	2012	2014	2016	2018	2020
截面应答率	85.7%	83.4%	80.1%	76.2%	63.4%
跨轮追踪率	85.1%	89.5%	87.8%	86.4%	75.7%

除了截面应答率外，跨轮追踪率及重复数据采集的轮次数也是度量追踪调查实施情况的重要指标。跨轮追踪率是以前一轮完访的样本为分母，计算此类样本在后续轮次完访的比例。表 2 展示了 CFPS 家庭层面跨轮追踪率。在受新冠肺炎疫情冲击前，CFPS 家庭层面跨轮追踪率始终保持在 85% 以上水平；2020 年调查模式从面访至电访的转变使少部分原本可以接受访问的家庭成为了无应答样本，导致家庭层面的跨轮追踪率下滑。

值得注意的是，CFPS 在项目实施过程中仅会移除死亡样本，往期调查中因各种原因未完访的样本将持续保留在样本库中，后期继续尝试追踪。因此，每轮调查完访样本中都包含一部分上轮未完访而本轮重新接受访问的样本，而这部分样本的完访情况无法在跨轮追踪率中体现。如计算家庭层面累计完访轮次数，约 85% 的样本至少拥有家庭层面上两轮调查数据，约四成样本拥有六轮完整的追踪数据。

CFPS 在每个家庭中访问基因成员及其非基因的直系亲属，对基因成员永久追踪，直系亲属则在关系存续时追踪。表 3 展示了 CFPS 符合追踪条件的基因成员在各轮次的应答率。根据 CFPS 项目访问规则，调查时居住在军队、监狱、寺庙、养老院这四类机构的家庭成员，以及去世、离家出走的家庭成员无须访问，这些人员不包括在应答率的分母中。在 CFPS 项目执行的前三个轮次中，基因成员截面应答率维持在 74% 以上；在第四、五轮次，截面应答率每轮下降约 5 个百分点，2020 轮次则因访问模式转换出现大幅下滑。个人跨轮应答率在疫情之前相对平稳，每轮调查时约八成左右的前期访问的受访者（包括基因和非基因）会继续接受访问。从累计完访轮次数来看，有 84.5% 的个人样本至少存在两轮或以上数据，其中六轮调查均接受访问的个人在数据库中占比为 22.3%。CFPS 在每个家庭需要通过家庭成员问卷数据识别出需要访问的成员，进而生成个人问卷。换言之，个人问卷完访是家庭层面完访后叠加个人问卷完访的组合行为。由此，可以解释为何个人层面的应答率及累计完访轮次数低于家庭。

表 3CFPS 个人应答及追踪率

	2010	2012	2014	2016	2018	2020
基因成员应答率	74.8%	74.3%	74.0%	69.9%	65.4%	49.2%
个人跨轮追踪率		79.7%	83.2%	80.9%	79.0%	63.7%

（二）跨年数据库使用的注意事项

构造追踪数据的核心在于跨年数据的匹配。CFPS 个体识别码 pid 是恒定的，直接通过 pid 可以进行个体层面不同数据集的合并。在跨轮个人数据库合并时，用户会遇到跨年间样本不能完全匹配的问题。导致跨轮无法完全匹配的原因主要来自以下三方面。第一，随着年龄增长，个体在不同轮次调查的数据会出现在不同个人数据库中。例如，2010 年 14 岁的受访者，其 2010 年数据保存在少儿库中；到 2012 年追访时，其年龄为 16 岁，应回答个人自答问卷，其 2012 年数据保存在成人库中。如果用户直接将 2010 和 2012 年的少儿库合并，会导致该个体的 2012 年数据缺失。第二，曾经接受过访问的样本并不能保证其在后续轮次都被一直成功追访，无法联系上、拒访、去世、不符合访问条件（如离境）都可能造成后续轮次这些样本的流失。样本流失在大型追

踪调查中是常见现象。CFPS 的相对优势是每一轮次均会尝试联系曾经进入过 CFPS 的所有符合条件的样本，让这些流失样本在后续轮次依然有机会重新被访问到。第三，非基因直系亲属只有居住在基因成员所在的家庭时才需要完成个人问卷，当关系断裂时，这部分人员就无须继续访问。对此，用户可根据数据库中 gene 变量判断所关注的人员是否属于 CFPS 基因成员。

不同于个人识别码的恒定，个人所属的家户号 fidXX 在跨年之间有可能发生变化，通过家户号 fidXX 在家庭层面进行跨年合并的做法并不妥当，因为跨年之间相同的家户号并不一定代表完全相同的家庭，也许只是关联家庭。家庭在跨年之间的变化既有可能是家庭规模变动（如新生儿加入或家庭成员去世），也有可能是家庭结构变化（如夫妻离异搬离或成年儿女经济独立，组建自己家庭）。如图 1 所示，对家庭规模变动，如果仅是人数变化，跨年间家户号不会改变。例如，一个 2018 年的受访家庭在 2020 年追访时添了一名新生儿，其家户号不会变动。因此，对这个家庭来说，fid20 和 fid18 一致。而对于家庭结构发生变动，跨年之间家户号可能会改变。当原先的一个家庭分裂成 N 个家庭时，会产生 N-1 个新家户号。这 N 个家庭中的其中一个会沿用原有家户号，其他家庭则被分配新家户号。一般来说，原家庭中最先成功接受家庭成员问卷访问的受访者，其所在家庭会沿用原有家户号。假设 2018 年有一个三口之家，家庭成员包括老两口及其子女，该家庭在 2018 年的家户号 fid18 等于 A。2020 年追访时，由于该子女经济独立，原来的三口之家分裂成两个家庭单元，2020 年的家户号除 A 之外还会新生成 A1。若 2020 年先成功访问到老两口中的一位，完成了家庭成员问卷，则老两口所在家庭的家户号会沿用 A，子女所在家庭采用新的家户号 A1；相反，若 2020 年先成功访问到子女，完成了其家庭成员问卷，则子女所在家庭的家户号会沿用 A，而老两口所在家庭会使用新的家户号 A1^①。

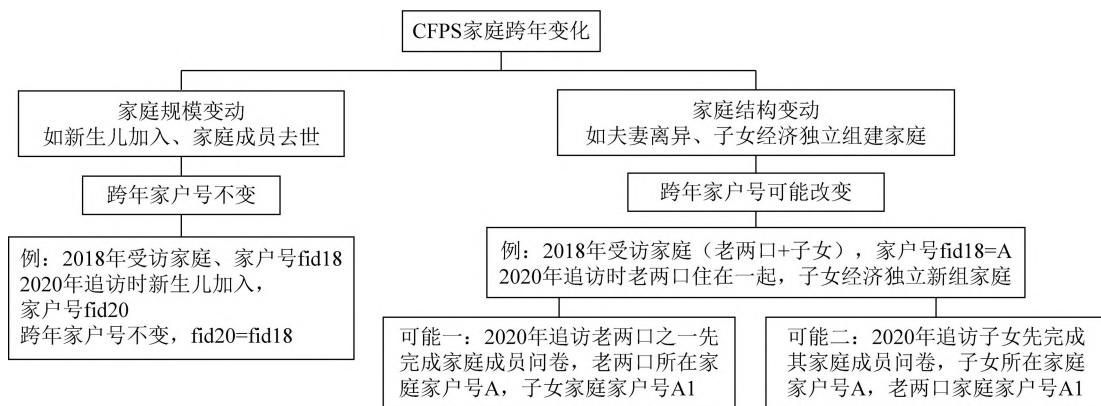


图 1 CFPS 家庭跨年变化引发的家户号改变情况

由上可知，无论家户号在跨年之间是否相同，家庭的规模和结构都有可能变化。因此，用户如要进行跨年家庭的分析，不能直接通过 fidXX 是否相同来确定家庭是否能完全匹配。更稳妥的做法是根据每轮家庭关系库中家庭所有成员的列表来判断跨年间的家庭结构是否完全一致。

五、规范性的理念及其运用

（一）以规范性为理念的数据管理模式

为了最大程度保护受访者隐私信息，CFPS 参照国内外标准，对数据实施分级共享机制。对于

^① CFPS2012~CFPS2020 数据中的家户号是如上述方法分配的。但是从 CFPS2022 数据清理时项目组在分配家户号时会尽量避免原有家户号的消失。关心此类问题的用户可以关注相应年份的清理技术报告，掌握最新信息。

可能识别到个体的隐私变量，CFPS 将其直接从共享数据集中删除。在用户可用的共享数据集中，项目组将变量根据其安全等级和机构及合作方要求等，分成三个等级。大部分问卷变量不涉及用户隐私，划为公开数据集，用户可在官方数据平台注册申请、下载使用。除公开数据集之外，还有两级属于限制数据。其中，一般限制性数据的安全等级稍高于公开数据集，项目组会要求用户提交额外的申请，说明研究目的，申请通过审核后方可在平台上下下载使用。机房使用类限制数据是目前可供用户使用的最高安全级别的数据，该类数据只能在北京大学中国社会科学调查中心的机房中使用。

以地址信息为例，公开数据集中只包含具体省级名称，省级以下的区县及村居均以项目编制的代码形式放置于公开数据集中。项目编制的代码与国标码一一对应，只要国标码不变，项目编制的区县码和村居码原则上保持不变，用户可以进行跨年比较。通过项目编制的代码，用户可以知道哪些样本居住在同一个区县或村居，能够进行不同轮次数据间的比较，也能判断哪些样本进行了跨区县或跨村居的迁移，但是用户不能从公开数据集中知晓具体区县名称或村居名称。任何通过非官方渠道获得的 CFPS 数据都会违反 CFPS 数据使用协议，影响由此产出的科研成果。

除了公开数据集中的省级地址名称、区县及村居编码外，CFPS 项目组提供了多种渠道支持用户通过规范途径进行相关研究。项目组依据 2010 年人口普查和统计年鉴等资料，构建了 2010 年区县宏观经济变量数据库，包含项目编制的区县码（用户可以基于此与 CFPS 公开数据集进行合并）、GDP、人均 GDP、人口数、就业率、平均受教育年限、劳动年龄人口比例、老年人口比例、10~19 岁人口性别比例、非农业户口人口比率。为了降低数据被反向识别的可能性，所有宏观变量的取值均进行过模糊化处理，数据库中不包含真实的区县名或代码。这部分数据属于一般限制性数据，用户可在官网限制数据页面申请使用区县宏观经济变量数据库，并在申请书中承诺不在此基础上探索可以识别出省级以下具体地址的信息。

机房使用类限制数据中包含具体的市级、区县名称，可以满足部分用户希望将 CFPS 数据与区县或市级层面的外部数据（如空气质量、营商环境等）进行合并的需求。这类分析需要用到真实的区县名（或市级名称）与外部数据进行链接，该类信息只能在北京大学中国社会科学调查中心的限制类数据机房中使用。用户通过申请后可以在机房中进行数据合并及数据分析，合并后的数据集不可带出机房，分析结果及使用代码在提交项目组安全审核后，方可用于科研发表。

有关限制类数据使用规范的最新信息，用户可在官网“数据中心”栏下方的“限制数据”页面查看，务必通过官方渠道申请数据并依照数据使用协议合规使用。

（二）数据共享的注意事项

近年来，越来越多国内外期刊要求或鼓励论文发表的同时附上相关的程序代码或者数据，以供读者复现论文中的相关发现。该举措的意义一方面在于提高相关研究质量，避免学术造假等不诚信行为；另一方面也能提高研究效率，避免重复性的数据采集^[6]。国外绝大部分高质量经济学期刊都明确要求作者提供能供读者复制论文结果的数据或数据申请渠道及分析代码。近些年，国内经济学期刊也陆续提出此要求。

对于 CFPS 用户来说，应如何回应期刊这一要求？能否将 CFPS 微观数据直接放在期刊网站上？首先，在 CFPS 用户注册申请数据的用户协议中已明确指出，用户不能将 CFPS 微观数据直接放置于包括期刊网站在内的第三方平台上，无论是原始完整数据，还是整理提取后的数据（如提取部分变量、进行清理后的子样本数据集等）。CFPS 微观数据的版权归北京大学中国社会科学调查中心所有，不可在未经授权的情况下被分享于其他平台。其次，作者拥有对自己编写的分析代码的版权，可以根据发表需求将分析代码（如 Stata 代码）置于合适地方，只要代码中不包含违反 CFPS 用户协议规定的信息即可。最后，CFPS 数据是公开申请的数据，作者只需在论文中提供 CFPS 数据规范申请方法，感兴趣的读者可以通过官方渠道申请，再根据作者提供的代码复制论文结果。此举有助于

保留数据处理细节，帮助研究成果复现。总之，从建立诚信的学术氛围角度来说，项目组不允许用户直接分享数据，建议仅向期刊提供基于 CFPS 发布数据集的分析代码及官方数据申请方法。

六、总 结

本文从 CFPS 项目综合性、嵌套性、追踪性和规范性四个方面介绍了项目设计和管理特点，并针对每个方面用户关切的数据使用问题进行了回应。CFPS 自 2008 年开始实施预调查，2010 年开展全国基线调查，至今已完成七轮全国调查。作为一个数据资源丰富的大型调查项目，在数据使用上有一定的门槛，对于其设计理念不够了解的用户可能会对数据产生误解，进而影响研究结果的有效性。希望本文的详细介绍能够帮助用户充分利用 CFPS 的已有资源，正确规范地使用 CFPS 数据，产出更多有价值、有潜力、有创新的研究成果。

参考文献

- [1] 李炜. 与时俱进: 社会学恢复重建以来调查研究的发展[J]. 社会学研究, 2016, 31(6): 73-94+243.
- [2] 谢宇, 胡婧炜, 张春泥. 中国家庭追踪调查: 理念与实践[J]. 社会, 2014, 34(2): 1-32.
- [3] 金勇进. 缺失数据的插补调整[J]. 数理统计与管理, 2001(6): 47-53.
- [4] Si Yajuan, Steve Heeringa, David Johnson, Roderick J A Little, Wenshuo Liu, Fabian Pfeffer and Trivellore Raghunathan. Multiple Imputation with Massive Data: An Application to the Panel Study of Income Dynamics[J]. Journal of Survey Statistics and Methodology. 2021, 11(1): 260-283.
- [5] 张学文, 陈凯华. 数字时代的开放科学: 理论探索与未来展望[J]. 科学学研究, 2022, 40(2): 203-208.

China Family Panel Studies: Design Concepts and Common Data Misuses

Sun Yan¹ Wu Qiong¹ Zhang Chunni²

(1. Institute of Social Science Survey, Peking University;
2. Department of Sociology, Peking University)

Abstract: China Family Panel Studies (CFPS) is the first national, large-scale, and multidisciplinary long-term panel study in China. It provides valuable information for studying China's changes. CFPS launched its first national baseline survey in 2010, covering 14960 households in 25 provinces. After the baseline survey, the project follows up the households and individuals every other year. Up to 2022, CFPS has completed seven rounds of national surveys, and the data from the first six rounds (2010-2020) have been publicly released. As an important data source of academic and policy research, CFPS has over 120,000 registered users and CFPS-related research findings are published on a daily basis. However, using CFPS data can be challenging for new users. Those unfamiliar with its design concepts may have misunderstandings about the data, and common misuses lead to biased and even erroneous research results. In order to enable researchers to use CFPS data in a scientific and standard manner, this article will present four features of CFPS in survey design and data management: comprehensiveness, nesting, traceability, and standardization. Frequently asked questions from the users are also addressed to minimize common misuses by users.

Key words: CFPS; Panel Study; Design Concept; Data Misuse

(责任编辑: 黄 煌)