CFPS China Family Panel Studies

中国家庭追踪调查
用户手册

（第三版）

谢宇 张晓波 涂平 任强
孙妍 吕萍 丁华 胡婧炜 吴琼

2017.7.30

目录

前言

中国家庭追踪调查（CFPS）经过多年筹备以及2008、2009年两年的预调查，于2010年正式开始基线调查，此后又分别于2012、2014、2016年开展了三轮全样本的追踪调查。此外，在2011年CFPS还对部分样本开展了一轮小规模的样本维护调查。众所周知，社会调查是一项复杂而细致的工作，从最初理念的设计、调查方法的使用，到执行过程中访员素质的培养、质控程序的设置，再到后期数据库的构建、数据的清理等，各个方面都可能影响到数据的学术使用价值。CFPS 基线样本覆盖 25 个省/市/自治区，代表了中国 95%的人口，2010年基线调查共采访14960户家庭、42590位个人，并对个人样本展开长期的追踪调查，是国内第一个如此大规模的、综合性的、以学术为目的的社会追踪调查项目，并期望能够成为最具权威性的研究中国家庭以及中国社会的调查项目——因此其复杂程度可想而知。虽然我们收集到了比较全面的、高质量的、高使用价值的数据——这一点是让人欣慰的，但是，我们也为此付出了代价。我们的设计、执行以及数据库都相对复杂，给用户使用数据带来了一定的难度。本用户手册正是出于方便用户的目的，希望通过简单的语言，尽可能全面、详细、具体地为用户提供在使用数据过程中需要了解的各方面信息。其中第一和第二版关注基线调查及相关数据库，这一版在前两版的基础上加入了关于后续追踪调查的设计、执行和数据的介绍。从总体上看，CFPS用户手册包括以下具体内容：

一，CFPS项目的设计理念与方法。如抽样的方法、加权的方法、测量工具的使用、问卷的设计、追踪策略，等等。

二，项目在实际运作过程中的各方面的操作细节。如绘图、住宅与住户过滤、访员访问程序的控制、数据质量的控制、样本的维护，等等。

三，数据管理与数据库构建。包括数据库的结构与内容、数据的清理、综合变量的构建，职业编码，等等。

四，技术报告索引。针对某些主题或者专业领域，我们为用户准备了一些独立的技术报告，使用户能够更进一步了解我们的项目与数据。此用户手册参考和引用了这些技术报告的部分内容，也为这些技术报告提供了索引。对于技术报告已有的内容，用户手册不再做详细介绍。

五，数据质量评估。通过与普查数据及其它一些数据的比较，对 CFPS 调查数据的质量进行简要评估。

本手册的很多内容来自调查过程中北京大学中国社会科学调查中心（ISSS）员工、学者以及助理积累下来的会议记录、文件、手册和技术报告；手册中关于 CFPS2010 年基线调查数据的初步统计分析与评估的图表由张春泥、许琪、周翔、徐宏伟和黄国英绘制；第一和第二版手册由胡婧炜负责资料整理与编辑，由张春泥负责校对。第三版手册由吴琼负责资料整理与编辑，由胡婧炜负责校对。此外，张欣、李汪洋、喻文姗、靳永爱在手册写作过程中提供了帮助。

我们衷心希望该手册能够为用户使用数据提供最大便利，如果因数据或者相关文件的更新需要对用户手册的内容进行变动，我们将会发布更新版本。如您在使用此手册的过程中发现问题和错误，我们恳请您批评指正。如果您有疑问或好的建议，也欢迎和我们联系。您合理的意见与建议，我们将会采纳至新的版本。

致谢

大量的工作人员为 CFPS 项目付出了巨大的努力与无私的贡献，CFPS 取得的成果凝聚了集体的智慧与心血，我们在此对所有为 CFPS 做出贡献的人员表示感谢。

以北京大学为主、包括国内外很多高校在内的诸多专业领域的专家和学者参与了 CFPS 调查的问卷设计工作，主要有：白建军、陈育德、陈玉宇、丁小浩、顾佳峰、郭志刚、黄桂田、李国平、李建新、李强、刘世定、卢云峰、乔晓春、邱泽奇、沈明明、沈艳、孙妍、涂平、吴琼、萧群、谢宇、徐湘林、严洁、杨伯溆、姚洋、袁瑞军、岳昌君、张春泥、张千帆、张拓红、张晓波、赵耀辉、周晓林、邹艳辉、蔡禾（中山大学）、郭有德（复旦大学）、雷洪（华中科技大学）、李路路（人民大学）、李培林（中国社会科学院）、李实（北京师范大学）、李友梅（上海大学）、刘精明（清华大学）、刘玉照（上海大学）、仇立平（上海大学）、任静娴（清华大学）、宋哲（清华大学）、王广州（中国社会科学院）、王正位（清华大学）、唐文方（美国爱荷华大学）、翁定军（上海大学）、吴晓刚（香港科技大学）、张伟强（清华大学）、周从意（清华大学）、祝建华（香港城市大学）、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）、Jean Yeung（National University of Singapore）。CFPS 问卷几乎涵盖了社会科学的所有研究领域，感谢这些专家学者无私贡献其在专业领域的学识与见解，他们为问卷内容的丰富、完善与优化提供了大量富有建设性的意见与建议。

同时，也要感谢为调查的执行付出了大量努力的工作人员。在繁忙的调查季中，为了保证调查的顺利进行，他们经常加班加点，甚至放弃节假日的休息时间，坚守在自己的工作岗位，耐心细致地处理调查过程中遇到的各类问题，保证了调查的顺利进行。他们是：蔡禾、曹宇龙、陈敏燕、丛维、陈佳波、慈丽娟、丁华、葛新兴、葛彬、顾春杰、顾佳峰、郭振威、韩俊丽、黄长群、洪洋、贾丹丽、贾小婧、李国华、李冉、李力、李胜文、李友梅、梁玉成、刘月、吕萍、吕杰、马耘、马腾宇、马文婷、马超、孟夏、彭德金、钱萍、邱泽奇、仇新晨、任莉颖、宋式斌、司玮、沈玉芳、邵娜、孙婷、孙帅、孙妍、孙翊、孙玉环、孙彩琴、滕学亮、杨旭、申容、万婷、王涛、王艳梅、王琪尧、王京、王堃、王雪音、魏晓雯、许琪、严洁、杨倩、杨思佳、姚佳慧、叶雪、易静、尹文茂、于双、于学军、藏好兵、张海东、张蓝心、张曼、张雅欣、张永建、周芸、周红苹、周滢滢、朱庭威、朱陈玲、邹艳辉。

CFPS 访问量之大、数据类型之多，使得问卷数据的管理与清理任务也相当繁杂。CFPS 数据团队在工作上一直精益求精，力求为用户提供质量可靠、使用便利的数据，在此，对他们的辛勤工作表示感谢，他们有：白玲、崔雅红、陈嘉、戴利红、胡婧炜、黄国英、靳永爱、李力、李汪洋、骆为祥、吕萍、马超、旎莎、任莉颖、任前平、任强、孙玉环、谭之博、王佳、王隆玉、王雪音、王骁、王玉磊、武玲蔚、吴琼、项军、谢宇、徐宏伟、许琪、严洁、阎溆、姚佳慧、於嘉、张春泥、张婧申、张聪、张文佳、张欣、赵端、赵方圆。

CFPS 在执行过程中得到了前国家人口和计划生育委员会、国家统计局、民政部、上海大学、中山大学的大力协助。在调查设计、技术支持等方面，美国密歇根大学社会研究中心给予了很多指导与帮助，双方建立起了良好的合作关系。另外，在资金方面，CFPS 得到了北京大学和国家自然科学基金的资助。我们也对这些单位表示感谢。

最后，也是最为重要的是，我们要感谢历年调查中辛苦奔波于一线的近两千多名访员。实地访问是社会调查中最重要也是最为艰辛的一环，我们的访员克服了调查过程中自然环

境、交通条件、天气因素带来的各种困难，出色地完成了调查工作。他们的辛勤工作换来了今天能为我们所用的高质量数据。我们更要感谢理解和支持我们调查工作的受访者，他们的积极配合是我们的调查能够顺利进行、数据能够真实反映社会状况的保障，没有他们，就不可能有 CFPS 如此珍贵的数据资料。

1. CFPS 概述

1.1 项目背景与调查概况

中国家庭追踪调查（CFPS）是一项全国性、综合性的社会追踪调查项目，旨在通过追踪收集个体、家庭、社区三个层次的数据，反映中国社会、经济、人口、教育和健康的变迁，为学术研究和公共政策分析提供数据基础（谢宇、胡婧炜、张春泥，2014； Xie & Hu，2014）。

CFPS 重点关注中国居民的经济与非经济福利，以及包括经济活动、教育获得、家庭关系与家庭动态、人口迁移、身心健康等在内的诸多研究主题。CFPS 的目标样本规模为 16000 户，调查对象为中国（不含香港、澳门、台湾以及新疆维吾尔自治区、西藏自治区、青海省、内蒙古自治区、宁夏回族自治区、海南省）25 个省/市/自治区中的家庭户和样本家庭户中的所有家庭成员（Xie & Lu，2015）。其中，居住在传统居民住宅内的、家中至少有一人拥有中国国籍的一个独立经济单元，便可视为一个满足项目访问条件的家庭户。[1] CFPS 定义的家庭成员指样本家户中经济上联系在一起的直系亲属[2]，或经济上联系在一起、与该家庭有血缘/婚姻/领养关系且连续居住时间满 3 个月的非直系亲属。

CFPS 于 2007 年开始前期工作，2008、2009 年在北京、上海、广东三地总共 2400 户家庭开展了初访与追访的预调查。2010 年，CFPS 在全国 25 个省/市/自治区正式实施基线调查，共发放样本 19986 户，最终完成了 14960 户家庭、33600 名成人，8990 名少儿的访问。此次调查在家户层面累积应答率为 81.25%，合作率为 96.58%，联系率为 84.13%，拒绝率为 2.67%；在个人层面应答率为 84.14%，合作率为 87.01%，联系率为 96.70%，拒绝率为 8.47%。[3]

CFPS2010 年基线调查为本地调查，重点完成了对样本村/居内的样本家户和家庭成员以及外出到本区/县范围内的个人的访问，同时也在家庭成员问卷中通过他人代答的方式收集了调查当时不在家的家庭成员的基本信息。经 2010 年基线调查界定出来的与家庭有血缘/婚姻/领养关系的所有家庭成员，他们作为 CFPS 的基因成员，将成为调查的永久追踪对象。这些基因成员今后新生的血缘/领养子女同样被视为基因成员，因此也将接受永久性追踪调查。

[1] 最初我们还要求受访家庭户中至少有一名成员在抽样社区居住时间满 6 个月，但在执行过程中，这一条件被取消，实际被这一条件过滤掉的仅有极少数家户。
[2] 关于直系亲属的界定参见孙妍等（2011）。
[3] 此数据按照 AAPOR 标准进行计算，参见技术报告：CFPS-5。

在此后的追踪调查中，基因成员在家中的非基因直系亲属（父母、配偶、子女）在 CFPS 中被定义为调查当年基因成员所在家庭的核心成员；而基因成员所在家庭的既不是基因成员也不是核心成员的家庭成员被称为非核心成员。在 CFPS 中，只有基因成员是被永久性追踪的；核心成员与基因成员关系存续时使用与基因成员同样的问卷进行访问，关系断裂时停止访问；非核心成员则仅通过他人代答的方式采集最基本的个人信息，其本人不作为 CFPS 的访问对象。

CFPS 2010 年基线调查全部采用面访形式，从 2012 年追踪调查起开始实行以面访为主、电话访问为辅的混合调查模式。CFPS 共有社区问卷、家庭成员问卷、家庭问卷、成人问卷和少儿问卷五种主体问卷类型。CFPS 2012 和 CFPS 2014 的电访问卷是在同期面访问卷基础上精简的版本，CFPS 2016 的面访和电访问卷则实施了高度的整合，除了个人问卷的认知测试外，其它内容完全一致。此外，CFPS 从 2012 年追踪调查起添加了代答问卷，通过在家的家庭成员代答的方式收集物理外出个人的基本信息[4]。

除 2010 年基线调查以及之后常规的两年一次的全样本追踪调查（CFPS 2012、2014、2016）外，CFPS 在 2011 年还对部分样本进行了一次小规模的样本维护调查。我们在本报告中将着重介绍基线调查及常规的全样本追踪调查的情况，仅在必要的时候会简要提及 2011 年的样本维护调查。

CFPS 由北京大学研究团队设计，由北京大学以及自然科学基金资助，由北京大学中国社会科学调查中心（ISSS）负责实施，并在执行中得到原国家人口和计划生育委员会、民政部的大力支持。

1.2 设计思路

1.2.1 中国社会的变化[5]

中国正在经历一场巨大的社会变革，其范围之广、速度之快、影响力之大在人类历史上史无前例。可以说，自 20 世纪末以来中国正在发生的这场变革在世界历史长期进程中的

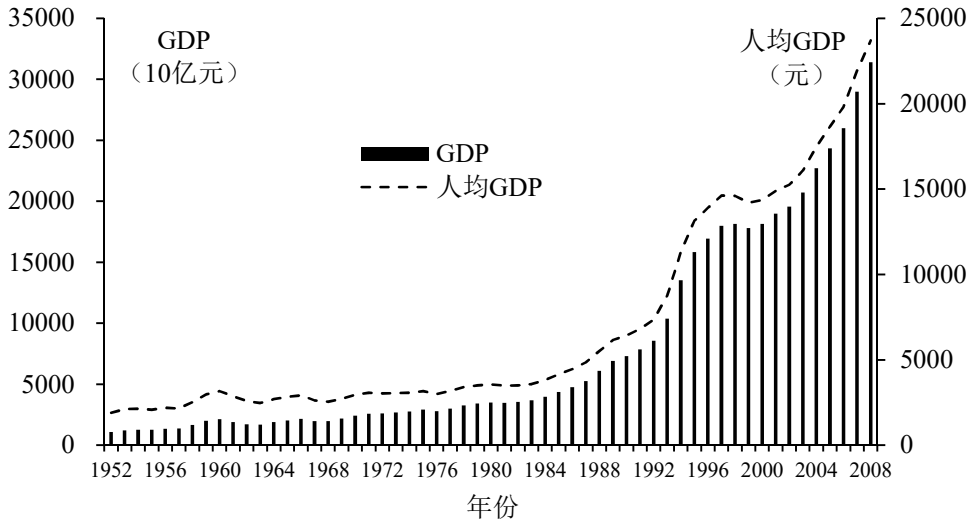
[4] 物理外出是指与受访家庭存在经济上的联系（即 CFPS 定义的同一家户的家庭成员），但物理居住地址不在同处。

[5] 1.2.1、1.2.2、1.2.3 部分的内容引自 Xie（2011）。

重要性并不亚于 14 世纪意大利文艺复兴、16 世纪德国宗教改革、18 世纪英国工业革命这样一些通常被认为是历史转折点的事件。中国的这场急剧、大规模且不可逆转的社会变革是多方面的，我们可以从经济增长、教育普及和人口转型三个方面看出这些变化的程度与速度。

中国的经济从 20 世纪 80 年代开始出现了大规模、持续、快速的发展。从图 1 可以看出，自 1978 年经济改革以来，中国的 GDP 和人均 GDP 显著增长。扣除通货膨胀因素，人均 GDP 在 1978 到 2008 年的年增长率为 6.7%，而美国即使是在黄金工业化时期（1860-1930），其人均 GDP 年增长率也仅为 1.5%，[6] 远远低于中国近些年来的人均 GDP 年增长率。

中国人的受教育水平在近年来也明显提高，这尤其体现在高等教育阶段。从图 2 可以看出，除文化大革命时期（1966-1976）外，中国大学生人数长期以来以平稳增长为主，但从 20 世纪 90 年代末起激增。中国受过高等教育的年轻人数目快速增长既是中国近年来经济增长的结果，也是经济增长的原因。



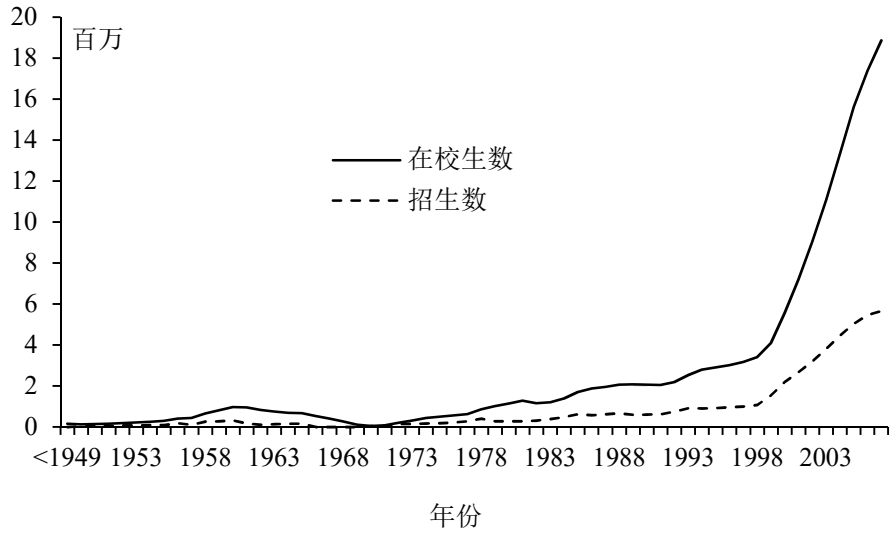
注：全部数据以 2008 年人民币的价格计算，扣除通货膨胀因素的影响。2005-2008 年数据在第二次经济普查的基础上进行过调整。

资料来源：国家统计局，2010，《新中国六十年统计资料汇编 1949-2008》，中国统计出版社。

[6] 根据 Measuring Worth（2011）提供的数据计算。http://www.measuringworth.com/。

国家统计局，2010，《中国统计年鉴 2010》，中国统计出版社。

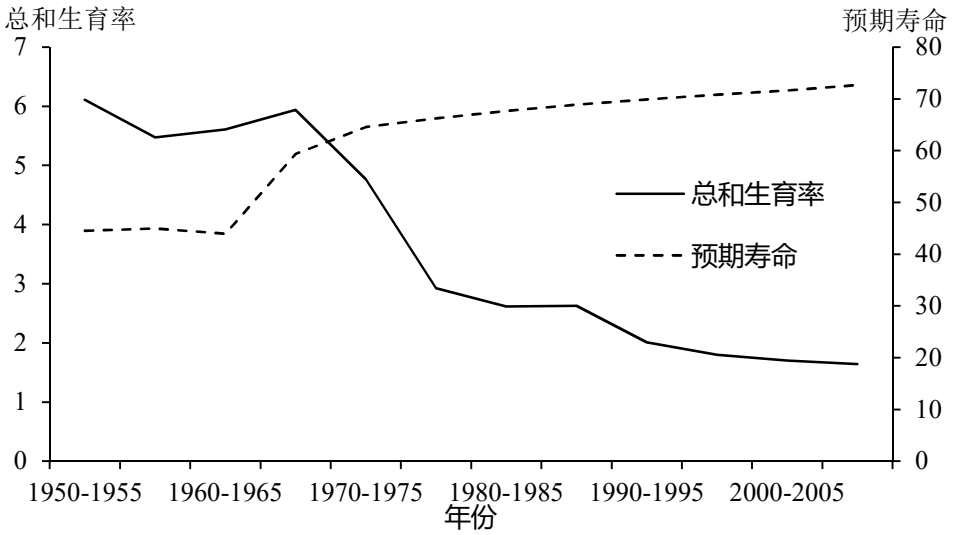
图 1. 中国 GDP 与人均 GDP 发展趋势， 1952-2008 （以 2008 年为基准年进行调整）



资料来源：《中国教育年鉴》（1984，1986-1988，1989-2008），人民教育出版社。

图 2. 中国高等教育的扩张，1949-2007

在这几十年间，中国同时完成了从高生育率、高死亡率到低生育率、低死亡率的人口转型。从图 3 我们可以清楚地看到，中国的总和生育率（TFR）自 20 世纪 70 年末以来急剧下降，从 6 降为 2，正好处在人口更替水平。预期寿命自 20 世纪 50 年代以来稳步提高，到 21 世纪初达到 70 岁，与发达国家 1970 年左右的水平相当，并远远超过其他欠发达国家。



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

图 3. 总和生育率和预期寿命，1950-2010

可以说，上述这三个领域的变化是中国在过去 30 到 40 年间最重要的社会变化。当然，在其他领域，很多重要的社会变化也正在发生并值得社会科学研究者关注，比如，社会不平等的日益加剧、离婚率的上升、婚前同居现象的增多、大规模的劳动力流动，等等。CFPS 项目的出发点，就是让学者有高质量的、全方位的、长期追踪的数据来更深刻地研究这些社会现象及其变迁。

1.2.2 实证研究在中国的重要性

中国所经历的这场社会变迁，不仅其独特性本身具有研究价值，同时它还具有改变世界历史进程的潜力。在过去将近三个世纪的时间里，西方社会主导了世界发展的潮流，其演变路径几乎成为了“现代化”、“发展”和“进步”的同义词。我们知道，西方社会的演变建立在两大支柱的基础之上：一是作为唯一合法政治体制的民主制度，二是作为唯一可行经济体制的自由市场。而如今，中国自近三个世纪以来头一次在经济发展模式上对西方提出了严峻的挑战。在既没有西方式民主政治体制、也没有真正的市场经济体制的情况下，中国经济的高速、平稳增长却持续了 30 多年。相比之下，美国等西方经济体在近年来却面临滞涨与衰退。这是否说明中国模式可能也是一种切实可行，甚至更好的发展路径呢？

这是一个有趣且迫切需要关注的问题。对这个问题的研究要求我们必须搁置对中国的

先入之见——这些先入之见或来自对别国经验的照搬，或出于理论上的推断。要了解中国，我们必须将中国置于它自身的历史、文化、政治和经济情境中；要了解中国，我们还必须客观地对待我们的研究对象，将我们的研究构筑在经验证据的基础上而非纯粹的想象上。

一方面，中国正在发生的这些变化的独特性只存在于今天的中国——它既不是别处任何一种社会变革的翻版，也不会在未来的中国重现。我们可以将其理解为一种仅发生在当代中国背景下的社会现象。这种背景包括中国当前的政治、经济、文化和社会环境，这是别处所不具备的。社会理论应该在社会情境中构建，社会研究应该在社会情境中进行，我们需要设计出在理论和方法论上具有创新性的研究框架，专门用于研究当代中国的社会现象。

另一方面，中国的社会科学如这个社会的其他方面一样正在经历迅速的变化。随时间的推移，以意见为主、意识形态化的、思辨式的讨论已渐渐失去了市场，实证研究开始被越来越多的社会科学家所接受。这是必然的趋势。无论是公众、政府还是学术界都欢迎高质量的实证研究，因为这样的社会研究符合他们的利益。首先，与社会科学家一样，普通的中国人也不完全理解他们的国家正在经历的一切，因此，他们自然有兴趣也切实需要去理解那些最直接影响他们生活的变化。其次，中国的政策制定者在决策之前也需要掌握更准确的信息或依据，以便理性决策。最后，中国是重要的世界大国，它在所有重要的领域，从艺术、体育和金融，到自然科学、技术和世界和平，都发挥着不可忽视的作用。我们希望未来中国在社会科学领域也能做出得到世界范围认可和赞誉的卓越贡献，而这必须依赖于实证科学研究。

1.2.3 CFPS 设计意图

综上所述，中国巨大的社会变革给当今的社会科学家们提出了挑战，也提供了机遇。如果我们不能理解这些变革，我们对社会的理解仍不全面。中国的实证研究是历史对当今社会科学家提出的迫切要求，而好的实证研究必须依赖于高质量的调查数据。从更长远的角度来看，当代社会科学家不可能完全理解中国正经历的这些变化，所以我们更不应该错过这样一个收集经验资料的黄金时期，为未来的社会科学家对当代中国产生超越于我们这个时代的理解创造可能性。所以，无论是出于当代社会科学实证研究的需要，还是出于将来研究者了解中国现阶段历史的需要，我们都应该珍惜这个独一无二的机会来记录中国发生的各项重大

历史变化，保留中国社会的这段历史，这也是 CFPS 收集数据的根本目的。以此目的为出发点，CFPS 的设计基于以下三大社会特征展开：

首先，社会现象是多维度的。社会现象的不同维度之间不是相互割裂的，而是紧密相连的。人们生活的不同方面，比如家庭背景、社会网络、住房、教育、职业、收入、健康等，都互相影响。

其次，社会现象是多层次的。宏观的国家政策、文化传统、历史事件，中观的区域经济、城市设施、社区环境，以及微观的家庭结构、代际关系、个人教育和职业，等等，无一不影响社会中每个个体的生命历程。

最后，社会现象具有时间上的持续性。过去的事件影响现在的行为，现在的经历、经验影响将来的决策。很多社会事件看似简单，但其实都是随时间积累的结果，有着错综复杂的因果联系和动态的发展历程，如人口的流动趋势、家庭支出的动态等等。

基于社会现象的多维度与多层次特征，CFPS 从社区、家庭、个人三个层次展开调查，在社区层次收集政治环境、村居面貌、基础设施、人口、资源、交通、医疗卫生、财政收支等多方面的宏观或者中观数据，在家庭层次收集家庭结构与关系、生活条件、社会交往、收入支出、资产状况等多维度的数据，在个人层次收集个人的教育、职业、收入、心理与生理状况、婚姻等方面的信息。通过这样一种设计，我们的研究对象不再是孤立的，在研究中个体、家庭、社会三者可以进行很好的关联。

社会现象在时间上的持续性加深了社会系统的复杂性，提高了社会研究的难度。时间性是研究社会现象与社会变化的一个重要概念。从方法论角度来讲，时间是一种信息，社会科学研究需要了解什么事情先发生，先被观察到，什么事情后发生的，后被观察到。比如说，人的行为随着个人经验、经历而变化，这与时间紧密相关。[7] 追踪调查关注动态的现象与时间维度上的变异，是研究社会现象时间性的一个非常有效的途径。它通过对同一人群（同一样本）在不同时点上的重复观察，可以掌握不同个体在不同时间的状况，从而帮助研究者更好地判断随时间发展的因果关系以及推断总体的变化趋势，对于总体异质性、因果推论以及状态变化这样一些社会科学领域的重要研究课题有着非常重要的价值。[8] 追踪调查虽然成本高昂、设计复杂、操作困难，但相比横向调查与趋势调查，它能够提供更多的信息，得到更

[7] 谢宇（2012）。
[8] 任强、谢宇（2011）。

多有效的资料，并将给科学研究带来更大的回报。正因为如此，CFPS 在设计之初便决定采用追踪调查的方式采集固定调查对象在不同时点上的数据，即，对 CFPS2010 年基线调查界定出来的基因成员及其今后的血缘/领养子女长期进行追踪调查。

我们希望 CFPS 项目通过这种多维度、多层次、长期追踪的数据收集方式，能够提供给今天以及将来的研究者最为全面、可信的一手数据，帮助他们做出更好、更科学的研究，也希望能够为国家政策的制定提供更为可靠的实证依据。

1.2.4 研究单元

社会科学研究的真正本质是变异性。[9] 在一个总体中，每个个体都不一样。虽然他们从属于同一个总体，但他们在具体特征上却存在很大差异。如我们所知道的，在当前中国社会，不同的个人在受教育程度、收入、生活习惯、健康状况、社会关系等各个方面都不一样。因为差异性的存在，我们不能把不同的个体单元等同看待。由于个人是反映人类社会变异性的最基本单元，许多社会现象，如健康、幸福感、工作等，最终也都会反映到个人层次，所以我们要了解社会，就必须了解在社会中不同的个人在生活质量、经济地位和社会角色等方面的差异。正因为如此，CFPS 最基本、最重要的研究单元是社会中的个人，CFPS 长期追踪的落脚点也是这些个人。

社会中的个人在生活质量、经济地位和社会角色等方面的差异不仅仅来自于其刚出生时的基因和家庭背景，而且还会在生命过程中受社会环境和个人特有经历的影响而逐渐变化。所以，个人的差异是一定的历史时间和一定的社会结构的产物，而影响个人差异的最重要的社会结构便是家庭。

首先，家庭构成了个人社会化最初始也最重要的环境。家庭赋予个人最初始的社会地位；个人从出生起在相当长的一段成长过程中要通过家庭来接触、学习社会规则；家庭环境对个人的态度、行为、期望也有持续性的影响。因此，要研究个人，必须要了解个人所处的家庭。

其次，中国人常以家庭为单位从事经济活动和社会交往。中国社会的一些重要社会现象，如经济生活、居住、抚育子女、赡养老人等，均是在家庭层面进行的。CFPS 关注中国社会，就必须从家庭层面对这些相关的重要主题开展调查。

[9] 谢宇（2012）。

再次，家庭是代际关系的主要机制，是代际传递的重要桥梁。对社会中代际关系与代际流动的研究，如，父母社会地位如何影响小孩，家庭资源在不同子女间如何分配，成年儿女的资源如何向父母转移等，需建立在一个清晰全面的家庭结构基础上，需要对相关家庭成员的信息有详细了解。

再次，家庭是研究婚姻与性别机制的重要平台。绝大多数成年男女都会结婚，夫妻来自不同的家庭，通过婚姻/同居的形式组成新的家庭。社会地位与资源在通过婚姻形式组合家庭的过程中得以重新分配与组合，男女在社会/经济成果分享、劳动分工等方面的性别差异也会在婚姻和家庭中体现。

最后，家庭在中国文化中更是具有非同寻常的意义。中国人崇敬祖先，重视孝道，婚姻上讲究门当户对、传宗接代，事业上追求光宗耀祖，这些传统的价值观念无一不体现出家庭/家族在人们生活中的重要地位。家庭是个人在物质与精神情感方面极其重要的支持来源，而对家庭的回报也是个人的义务与责任。虽然随着中国社会的变化，传统的家庭观念也在受到侵蚀，但从中国家庭中父母对小孩的投资、家庭的社会关系网络对家庭成员的影响、家庭内部资源的转移等诸多方面仍可以看出家庭在中国文化中非同一般的重要性。

综上所述，对中国社会的了解与研究不可能脱离家庭而实现，因而家庭也是 CFPS 一个重要的研究和调查单元。CFPS 对家庭关系和家庭成员信息开展了全方位的深度调查，建立起了可明确定位家庭成员间关系的精确的家庭结构网络，并详细采集了家庭经济社会生活的各方面的详细资料，希望能够给人们了解和研究中国社会提供更广阔的视野。

1.3 国际比较

CFPS 在设计的初期借鉴了世界上一些先进的调查项目的方法、工具与成功经验，主要有 PSID、NLSY、HRS 等。同时，CFPS 也具备自己的一些特点与优势，以满足各领域研究者的不同需求。

PSID （Panel Study of Income Dynamics）[10]始于 1968 年，由美国密歇根大学设计并实施，是美国当前最具权威的对家庭经济的跟踪调查项目。该项目最初设计的目的是为了研究贫穷以及蓝盾·强森（Lyndon Johnson）“对贫穷作战”（War on Poverty）计划对人民经济福

[10] 关于 PSID 的介绍，参考 http://psidonline.isr.umich.edu/default.aspx。

利的效应，之后研究的主题逐渐扩展到雇佣、收入、财富、住房、食品开支、转移支付、婚姻与生育等方面。项目最初的样本量为 5000 户家庭，主要采用电话访问的形式，在每个家户抽取一位成年人作答。从 1997 年开始，PSID 增加了为少儿设计的专项调查（The Child Development Supplement to the PSID，PSID-CDS）。

NLSY （National Longitudinal Surveys of Youth）[11] 是反映美国的年轻人群体进入劳动力市场状况的一个权威性追踪调查项目，由俄亥俄州立大学设计、芝加哥大学全国民意研究中心（National Opinion Research Center，NORC）具体实施。该项目分别在 1979 年和 1997 年开始了对两批年轻人的长期追踪调查，1979 年的调查样本由 12686 名 14-22 岁的年轻人组成，1997 年调查的样本主要由 9000 名 12 -16 岁的年轻人组成。NLSY 的研究主题集中于年轻人的人力资本和劳动力市场活动，调查内容包括学校教育、雇佣状况、职业培训、工作时间、收入与资产、态度与行为、健康、政治参与等多个方面。

HRS （Health and Retirement Study）[12] 项目启动于 1992 年，由美国密歇根大学设计实施，是美国关于老龄问题研究最具影响力的一个长期追踪调查项目。HRS 调查对象为 50 岁以上的老年人，样本量约为 26000 人。HRS 关注老年人劳动力市场参与和健康的变化情况，及其与社会、经济、心理、退休之间的关系，调查内容包括收入、工作、资产、养老规划、健康保险、生理健康、认知能力、健康护理等多诸多方面。

总的来说，PSID、NLSY、HRS 都是美国大型的、具有全国代表性的追踪调查项目，具有大规模的样本量。它们虽然都属于专题研究，但同时有着广泛的调查内容，可供不同学科、不同兴趣的研究者进行各种主题的研究。这三个项目的所有数据向研究者开放，是相关领域众多科学研究成果的重要数据来源。CFPS 借鉴和吸收了这些项目的重要优势，希望能够为中国的社会研究做出同样重要的贡献。与上述调查类似，CFPS 同样具有全国的代表性和大规模的样本量，在基线调查中共访问 14960 户家庭，界定出需要长期追踪的基因成员共 57155 人。CFPS 调查的内容也非常具体全面，它作为一个多层次、多维度、长期性的综合追踪调查项目，涉及到少儿、成人整个生命历程的各类重要事件，同时还有针对家庭关系、家庭经济和社区等各个方面的专门设计。

在吸收上述调查项目成功经验的同时，CFPS 也做了一些改进。如上文所述，我们认为

[11] 关于 NLSY 的介绍，参考 http://www.bls.gov/nls/nlsy79.htm、http://www.bls.gov/nls/nlsy97.htm。
[12] 关于 HRS 的介绍，参考 http://hrsonline.isr.umich.edu。

社会结构的多层次性是中国社会的一个重要特征，在中国的多层次社会结构中，家庭又尤其重要。因此，CFPS 对家庭关系与家庭成员的信息开展了全方位、更具深度的调查。一般的调查项目对家庭关系的调查通常以家庭中的一两位受访者为核心，采集其家庭背景信息以及其他家庭成员的情况，收集到的信息极为有限。相比之下，CFPS 的调查对象不仅仅是一位或少数几位成人，而是覆盖了所有家庭成员。凡是满足条件的家庭成员（包括儿童在内）均需要自答/代答其相应的个人问卷，从而提供更详细、准确、全面的数据。此外，CFPS 独有的设计[13] 使得研究者不仅可以知道家庭成员间最直接的父母、配偶、子女关系，还可以推断出相互间的一些间接亲属关系；不仅可以了解家庭中每一名受访者的父母、子女、配偶的情况，还可以了解其跨代的祖父母、孙子女，以及同辈的兄弟姐妹的情况；不仅可以清楚地得到与受访者同住的家庭成员的各类具体信息，还可以得到与受访者不同住的直系亲属的一些重要的社会人口信息。研究者通过 CFPS 的个人编码体系，可以对家庭成员进行准确定位，从而可以得出一个精确的家庭关系网络。CFPS 的设计为研究者提供了更具研究价值的家庭结构与家庭成员信息，这也是 CFPS 相比其它调查最为突出的一个优势。

1.4 技术手段

CFPS 作为一项全国性、综合性的追踪调查项目，其样本规模之大、覆盖范围之广、设计之复杂，使得传统的纸笔调查方式已不再可能适应其繁重而复杂的调查与管理任务，也不可能满足其多样化的设计与调查需求。CFPS 在 2010 年使用了 CAPI（计算机辅助面访调查）访问模式，从 2012 年开始又增加了 CATI（计算机辅助电访调查）访问模式。这些计算机辅助调查方式的使用保证了调查的效率与质量。

计算机辅助调查依赖于访问管理系统来实现，这是一个专业的调查访问软件，它不仅可以辅助访员对受访者进行问卷调查，而且可以帮助访员管理各种访问信息。另外，通过访问管理系统，访员可以方便地与总部建立联系，及时解决在访问过程中出现的各种问题。具体来说，它的主要功能有：

⑴ 通过问卷的电子化使得复杂的问卷设计成为可能。借助电子化问卷，CFPS 不仅可以使用多选、单选、表格、循环、区间等多种提问方式，同时还能够通过设计各种复杂的逻辑跳转条件，针对人群的不同特征量身定制问题。而且，电子化问卷还可以通过设置硬检查、

[13] 这里的设计指的是 T 表格设计，在下文中将有详细介绍。

软检查的方式，现场对一些不符合逻辑或者常识的答案给出提示，使访员能够即时和受访者进行沟通并修正数据。

⑵ 样本的快捷管理。通过访问管理系统，总部可以实时为访员远程发放样本，并根据实地调查的需要在访员之间进行调查任务的调换。同时，系统还可以记录每个受访者或者受访家庭的详细信息，包括联系地址、访问方式、支付状况等，为督导和访员的工作提供方便。

⑶ 实时数据传输。通过访问管理系统，访员可以即时与总部实现数据的传输与交换，总部可以方便地了解调查的进度并实现远程控制。针对调查中发现的数据问题，总部也可以及时发现原因，并联络访员解决问题。此外，访问管理系统也帮助我们省去了传统纸笔调查中数据录入的环节，通过系统我们可以随时进行数据清理和分析工作。

⑷ 实时的访员行为监督与质量控制。访问管理系统可以通过录音等方式记录并监督访员的电脑操作行为。如果总部发现访员存在不规范访问行为，可以及时告知访员进行改进。

⑸ 并行数据分析。访问管理系统可以完整地收集一套访问过程中的并行数据（paradata），如访员在每道问题上的停留时间、访员对问题答案的修改记录，等等。对这些数据的分析可以为今后进行更为合理与完善的调查设计提供科学依据。

2. 抽样

2.1 抽样设计

CFPS 的样本覆盖中国除香港、澳门、台湾、新疆、西藏、青海、内蒙古、宁夏和海南之外的 25 个省/市/自治区的人口。这 25 个省/市/自治区的人口约占全国总人口（不含港、澳、台）的 95%，因此，CFPS 的样本可以视为一个全国代表性样本。

CFPS 最初目标样本规模为 16000 户，其中，有 8000 户从上海、辽宁、河南、甘肃、广东五个独立子样本框（称为“大省”）过度抽样（oversampling）得到，每个“大省”1600 户。另有 8000 户则从其他 20 个省份共同构成的一个独立子样本框（称为“小省”）抽取（Xie & Lu，2015）（见图 4，表 1）。5 个“大省”的子样本具有地区自代表性，可以进行省级推断以及地区间比较。5 个“大省”样本框在二次抽样后，与“小省”样本框共同构成具有全国代表性的总样本框。[14]

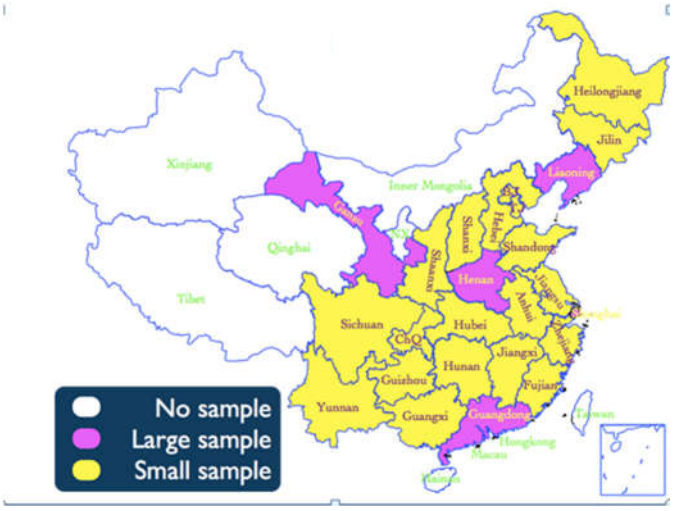


图 4. CFPS 样本来源区

考虑到中国社会有很大的地区差异，同时为了减少调查的运作成本，CFPS 抽样采用了内隐分层的（implicit stratification）、多阶段、多层次、与人口规模成比例的概率抽样方式（PPS）。行政区划和社会经济水平是主要的分层变量。在同级行政层以地方人均 GDP 作为社会经济

[14] 我们将此样本框中样本称为再抽样样本或整合样本。

水平的排序指标；在无法获得 GDP 指标的条件下，则采用非农人口比例或人口密度作为替代指标。

表 1. CFPS 目标样本规模

省/市/自治区类型	省/市/自治区	目标户数	过度抽样比率
“大省”	上海市	1600	10.28
	辽宁省	1600	4.45
	河南省	1600	2.04
	甘肃省	1600	7.30
	广东省	1600	2.02
“小省”	江苏省、浙江省、福建省、江西省、安徽省、山东省、河北省、山西省、吉林省、黑龙江省、广西壮族自治区、湖北省、湖南省、四川省、贵州省、云南省、天津市、北京市、重庆市、陕西省	8000	1.00

CFPS 每个子样本框的样本都通过三个阶段抽取得到。第一阶段样本（PSU）为行政性区/县，第二阶段样本（SSU）为行政性村/居委会，第三阶段（末端）样本（TSU）为家庭户（见表 2）。[15] CFPS 前两个阶段的抽样使用官方的行政区划资料，第三阶段则使用地图地址法构建末端抽样框，并采用随机起点的循环等距抽样方式抽取样本家户。考虑到每个地区的应答率，2010 年的实际操作参考了 2008 年和 2009 年预调查所得的预估应答率，采用按应答率比例扩大样本规模的方法，依据系统抽样原则共抽取了 19986 个居住地址，以保证获得预计的有效样本家户数量（见表 3）。

表 2. CFPS 三阶段抽样

阶段	广东、甘肃、辽宁、河南 4“大省”	上海“大省”	“小省”	总计
第一阶段	4×16 个区县=64 个区县	32 个街道（乡镇）	80 个区县	144 个样本区县+32 个样本街道（乡镇）
第二阶段	64×4 个村居=256 个村居	32×2 个村居=64 个村居	80×4 个村居=320 个村居	640 个村居
第三阶段	640×[28，42]户			19986 户

[15] 上海因其不同于其它“大省”，样本的抽取略有不同，具体可参考技术报告：CFPS-1。

表 3. CFPS2010 年基线调查末端样本量[16]

地区	类型	预计应答率	接触样本数量
低回答率地区	居委会（主城区和城乡结合部的村委会[17]）	60%	42
	其他村委会	70%	36
一般回答率地区	居委会（主城区和城乡结合部的村委会）	70%	36
	其他村委会	80%	32
高回答率地区	居委会	80%	32
	村委会	90%	28

值得一提的是，考虑到官方对于农村与城市的划分已难以反映中国快速城市化的现实，CFPS 抽样没有再采用将农村与城市分开抽样的传统方式，而是将中国社会作为一个整体进行抽样。我们在社区层面收集了样本社区是属于居委会还是村委会的信息，在家庭层面收集了家庭从事农业生产与非农经营的信息；在个人层面收集了个人的户籍信息以及个人从事农业工作与非农工作的信息。用户可以通过这些实际情况来判定样本的农村/城市属性，而不单纯依赖于行政区划。

抽样的具体设计方法与实施方式可参见技术报告《中国家庭追踪调查 2010 年抽样设计（CFPS-1）》。

2.2 末端抽样框制作

考虑到我国目前人口流动性大、人户分离严重的情况，我们认为如果仅利用村/居委会的户籍花名册进行抽样，会导致很大一部分住户的信息遗漏。为了得到一个完整覆盖样本村/居所有住户的末端抽样框，提高末端抽样的精度，在末端抽样之前，我们通过纸笔作图的方式，对样本村/居的地图进行了实地绘制。 我们先于 2009 年初至 2009 年 8 月在北京、河

[16] 转引自技术报告：CFPS-1。
[17] 主城区、城乡结合部的划分由国家统计局设计管理司的城乡代码确定。

北的四个村/居开展了多次试点工作，在详细了解了不同类型村居在建筑物特征、建筑物编号规则、住户列表清单制作方法、村居现有地图和住户清单可用性等方面的信息后，制定了村居抽样框制作的基本方案。为测试该方案的可行性，在 2009 年 11 月至 12 月，我们在甘肃和浙江各抽取了 4 个村居进行了预调查。预调查结束后，我们总结各方面的经验，进一步改进了我们的地图绘制方法和住户列表清单制作方案。

自 2009 年 12 月至 2010 年 4 月，我们开展了 23 批绘图员培训，每批培训时间 3 天，共培训绘图员 243 名，培训内容包括建筑物绘制、建筑物编号、辅助材料收集、建筑物列表清单和住户列表清单制作等。

绘图工作正式开始于 2009 年 12 月，至 2010 年 6 月结束，共获得了 649 个村居的纸笔绘制地图、村居基本情况统计表及村居住户列表清单。为了保证绘图质量，我们采用了多种手段进行了多次核查。具体的绘图方法与核查标准可参见技术报告《中国家庭追踪调查 2010 年基线调查末端抽样框制作（CFPS-2）》。

抽样组在对绘图资料进行整理，对一户多宅、一宅多户、地址类型无法确认等特殊问题进行处理后，[18] 便开始进行第三阶段的末端抽样。

3. 问卷设计[19]

3.1 总述

CFPS 的主体问卷包括村居问卷、家庭成员问卷、家庭问卷、少儿问卷和成人问卷五类。调查在社区、家庭和个人三个层面进行（图 5）：在社区层面，CFPS 通过村居问卷对各样本村/居进行一个整体的访问，主要了解该村/居的基础设施、人口结构、政策实施、经济情况、社会服务等信息；在家庭层面，由一位家庭成员回答一份关于家庭成员信息与成员间关系的家庭成员问卷以及一份反映家庭整体情况的家庭问卷；在个人层面，对于符合资格的个人，16 岁以下者回答少儿问卷，16 岁及以上者回答成人问卷。其中，少儿问卷分为代答和自答两个部分，10 岁以下的少儿，由其监护人回答代答部分问题；10 岁至 15 岁的少儿，除监护人回答代答部分问题外，本人还需完成自答部分问题。

[18] 更详细的内容可参考技术报告：CFPS-1。
[19] 如需更了解更详细的问卷设计相关内容，请参考孙妍等（2011）。

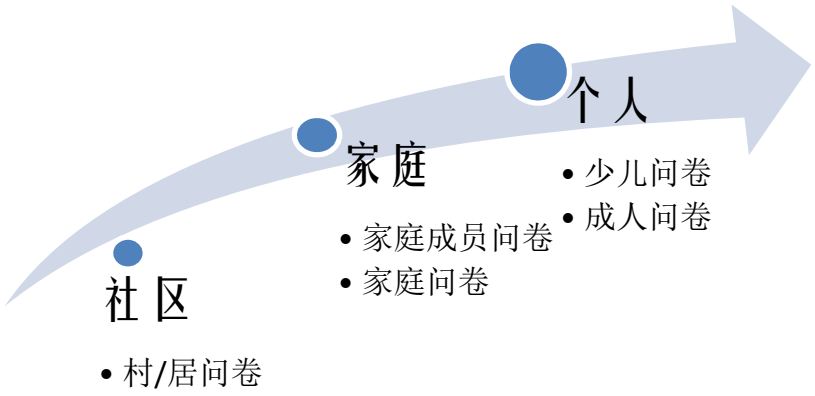


图 5. CFPS 主体问卷设计层次图

除 5 套主体问卷外，CFPS 在 2010 的基线调查中还设计了住宅过滤与住户过滤两套小的执行操作问卷。其中，住宅过滤主要通过访员实地观察与判断完成，由系统自动控制操作流程，没有程式化的问卷。住户过滤则和传统问卷一样，需要通过面访受访者完成。2010 年基线调查全部问卷的结构流程图见图 6。

家庭成员问卷是家庭中成员的个人问卷产生的前提。如前文所述，CFPS 将 2010 年的所有家庭成员及其今后新生/领养的子女定义为基因成员，将基因成员在家的非基因的直系亲属（父母、配偶、子女）定义为核心成员，将基因成员所在家庭的既不是基因成员也不是核心成员的家庭成员定义为非核心成员。在家庭成员问卷完成后，系统会根据每一位在家的基因成员和核心成员的年龄，产生相应的个人问卷；对不在家的基因成员和核心成员，先由在家成员完成一份代答问卷后，将该样本转入调配系统，安排访员在其所在地对其进行面访或电访。[20]对于非核心成员则停止访问，不再产生个人问卷。此外，如果调查当时家庭成员由于出家、服刑、参军/服役、出境四类原因不在家，当年不对其进行访问。

[20] 在 2010 年的实际调查中，我们仅开展了本地调查，即完成了对样本村/居内的样本家户和家庭成员以及外出到本区/县范围内的个人的访问，对于外出到样本区/县外的不在家家庭成员都没有进行个人访问。但是，我们也在家庭成员问卷中通过他人代答的方式收集了他们的基本信息。

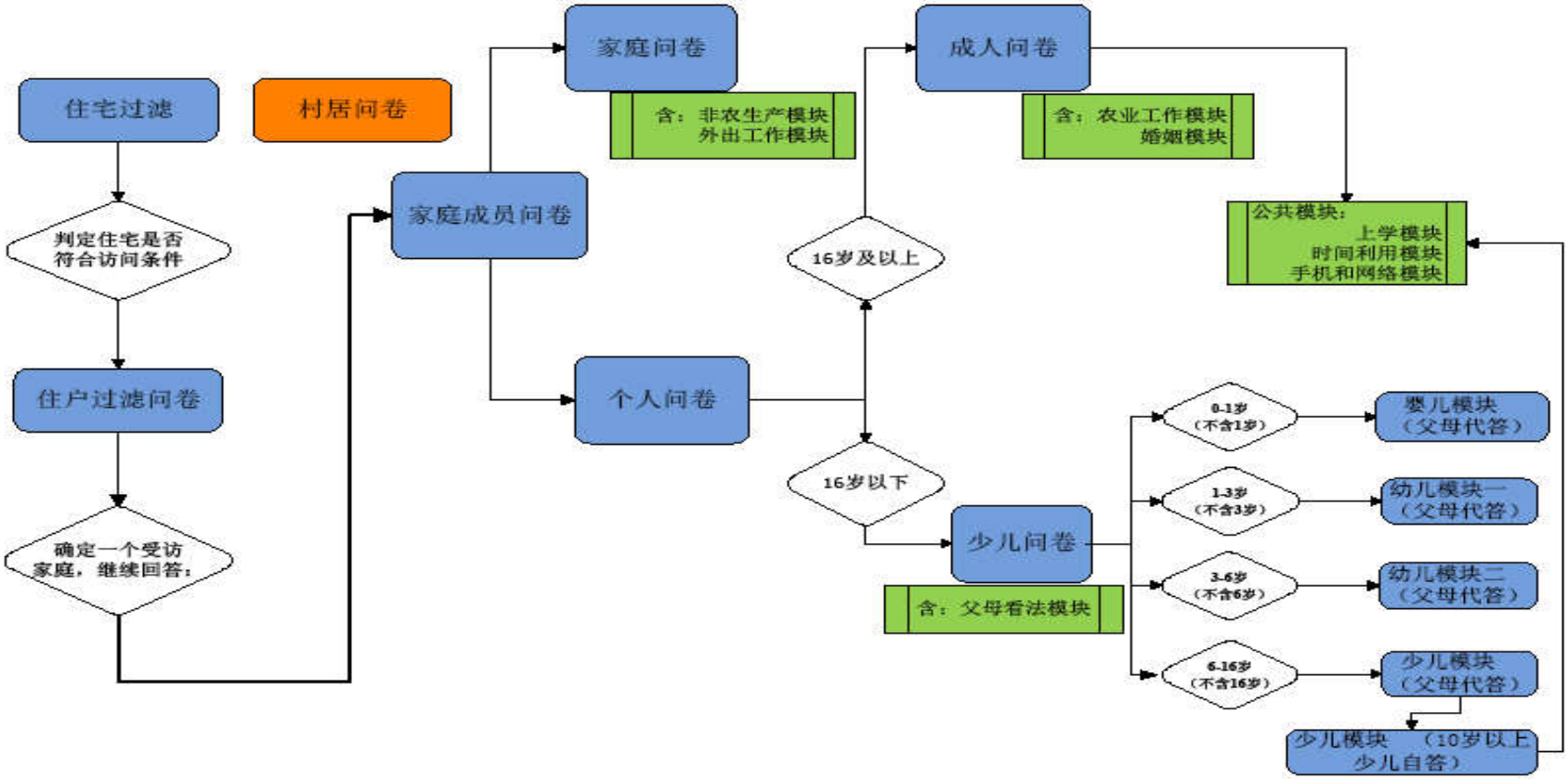


图 6. CFPS 2010 年基线调查问卷结构图

自 2012 年追踪调查开始，CFPS 从上一轮调查界定出的家庭户开始展开家庭层面和个人层面的访问，不再进行住宅或住户过滤。对于基因成员另组建的家庭，系统会自动分配一套新家庭成员问卷，以该家庭中的基因成员为起点，构建另组家庭的成员结构。对于去世人员，由家人回答与去世人员死亡相关的信息后，终止调查，保留个人 ID。追踪调查问卷结构流程图见图 7。

CFPS 采用模块化的设计方式，每个受访家户或个人的问卷内容根据其具体情况，由相关的不同模块组合而成。计算机辅助调查系统为我们在访问当中即时调用相关问题模块、建立个性化问卷提供了方便。比如说，对正在上学的受访者调用上学模块，对有工作的受访者调用工作模块。我们不对农村和城市分开使用不同问卷也是出于同样的道理。

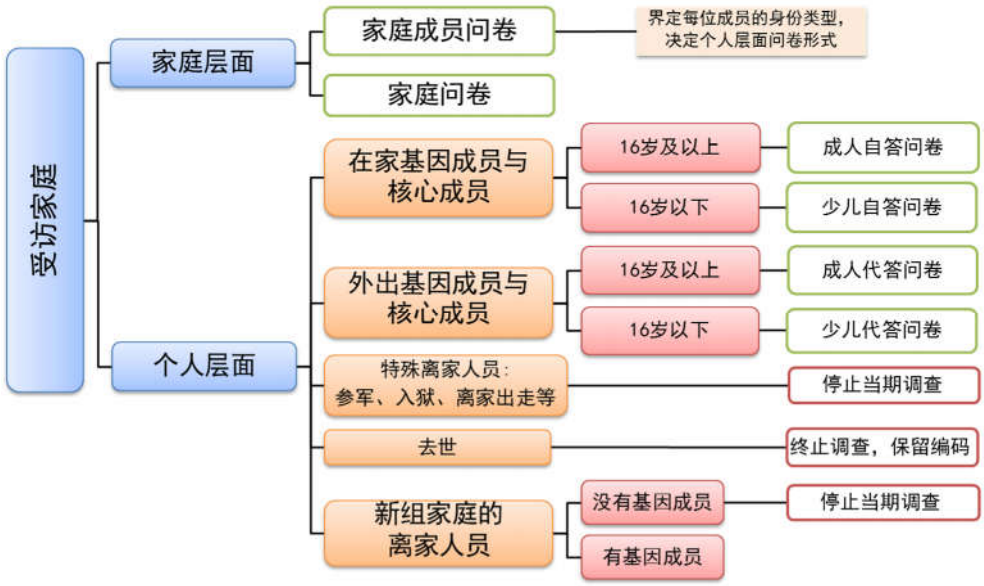


图 7. CFPS 追踪调查问卷结构图

3.2 村居问卷

村居问卷的主要目的是了解村（农村社区）或者居（城市社区）的设施、人口、政治、经济、历史、政策等相关情况。2010 年基线调查的问卷的流程与内容见下图 8 与表 4 第 2 列。

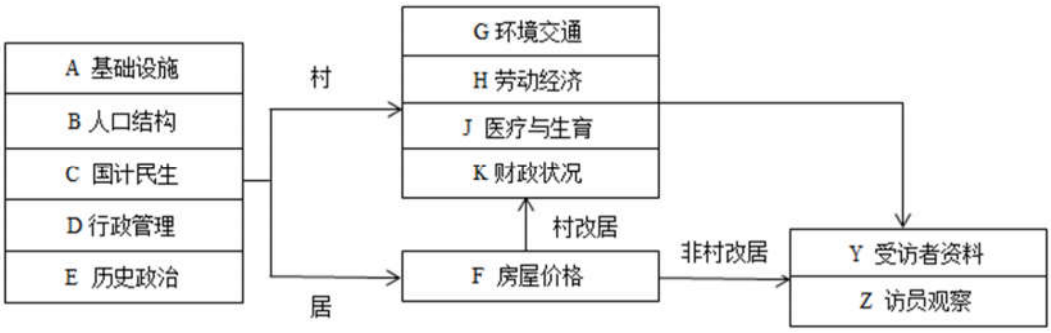


图 8. CFPS 2010 年村居问卷流程图

表 4. CFPS 2010 年与 2014 年村居问卷主要内容

模块	2010 年问卷内容	2014 年问卷内容
A 基础设施	村/居属性，受访人职务，设施，地界，行政面积，水源，燃料，高污染企业	村/居属性，受访人职务，设施保有及新增，地界，行政面积，水源，燃料，高污染企业
B 人口结构	总户数，总人口，户籍人口，常住人口，外来流动人口，年龄结构，出生与死亡，少数民族	总户数，总人口，户籍人口，常住人口，外来流动人口，年龄结构，出生与死亡，少数民族，大姓分布
C 社会保障与食品价格	低保政策，物价水平	低保政策，物价水平
D 行政	村居办公人员规模，办公条件，周边交通，选举情况	村居办公人员规模，选举情况
E 历史政治	历史变革，是否为旅游区，是否有高污染企业，最近一次村居委会选举情况	
F 房屋价格	商品房历史最高价、上个月最高价、上个月一般价	商品房历史最高价、上个月最高价、上个月一般价
G 环境、交通与资源	距最近集镇、县城、省城的距离与交通时间，矿产资源，自然灾害，土地资源	距最近集镇、县城、省城的距离与交通时间，矿产资源，自然灾害，土地资源
H 劳动力、产值与收入	劳动力结构，雇工价格，农业总产值，非农业总产值，人均纯收入，大姓分布	劳动力结构，雇工价格，农业总产值，非农业总产值，人均纯收入
J 医疗与生育	医疗点面积，医疗卫生人员数量，农村合作医疗开展情况，计划生育政策	医疗点面积，医疗卫生人员数量，农村合作医疗开展情况，二胎政策
K 财政状况	集体企业及产值，财政总收入及来源，财政总支出及支出项目	征地经历、集体财政总收入及来源，财政总支出及支出项目，债务情况
Y 受访者资料	受访者性别、年龄、政治面貌与受教育程度，村/居主任性别、年龄、政治面貌与受教育程	受访者性别、年龄、政治面貌与受教育程度，村/居主任性别、年龄、政治面貌

	度，其他受访人姓名与职务	与受教育程度,其他受访人姓名与职务
Z 访员观察	经济状况，马路整洁情况，成员精神面貌，成员同质性，建筑格局，拥挤程度，村/居类型，地貌，受访者特征	经济状况，马路整洁情况，成员精神面貌，成员同质性，建筑格局，拥挤程度，村/居类型，地貌，受访者特征

在对社区问卷进行访问时，我们建议访员尽可能找到比较了解村居、能接触到统计资料的人员，由这些人员尽量多地回答这份问卷。村/居委会工作人员，尤其是主持日常管理服务工作的村长/居委会主任，是访问的较佳人选；会计等其他社区工作人员如果因其工作内容或者服务年限而充分了解社区情况，也可以成为访问对象；另外，村/居的党支部书记、支部委员等人如果全面掌握情况，在前两类人员无法访问的情况下也可以成为访问对象。在一个访问对象无法回答所有问题的时候，我们允许第二个、第三个受访者来配合回答某些问题。但这种访问并不是集体访问，而是每个受访者到单独的房间受访。

在使用村居问卷时需要注意的一个问题是，“年末”指自然年的最后一天，如 2009 年年末指 2009 年 12 月 31 日。

CFPS 2012 将村居问卷中的物价水平部分纳入家庭问卷，没有进行单独的村居问卷访问。CFPS 2014 对原始抽样的 649 个村居进行了回访。该轮次村居问卷的基本框架与基线调查基本一致，采集 2013 年 12 月 31 日时点上的各项统计指标，并了解自 2010 年 1 月 1 日至 2013 年 12 月 31 日间村居所发生的变化。2014 年村居问卷主要内容参见表 4 第 3 列。

3.3 住宅过滤

为确保末端抽样的准确性，虽然在末端样本框的构建阶段我们的绘图员已经通过实地走访、询问邻居、与居委会人员确认等各种手段尽力将社区内空置房屋和非家庭住户排除，但在实地调查中，我们依然要求访员根据地图地址信息找到对应的备选样本，确认样本地址对应建筑是居民住宅后方可以正式开始访问。这种确认房屋类型的工作即住宅过滤。

访员根据我们提供的地图找到样本地址后，先确认地址是否有效，即我们提供给访员的地址是否实际存在。在地址确认有效后，通过询问住宅内的住户或其他知情人员，判定样本地址上建筑物的类型。经实地确认为住宅的样本会直接进入住户过滤环节，对无效地址、非住宅和空置房屋的样本则终止调查。对于难以确认建筑物类型的样本，我们会要求访员多次尝试。

3.4 住户过滤问卷

在确定样本地址上的建筑符合我们定义的住宅条件后，调查便进入了住户过滤环节。住户过滤问卷的主要目的是甄别出该住宅地址内符合访问条件的住户。住户过滤问卷主要有以下环节：

首先，界定出样本地址上的经济独立单元的个数。如，父母和子女住在一起，如果是一个经济共同体，则判定为一个经济独立单元；如果是两个经济共同体，则判定为两个经济独立单元。界定的范围不仅仅局限于房屋所有者，也包括在访问时拥有抽样住宅的全部或部分居住权的成员。

其次，在经济独立单元中过滤出符合条件的家庭户。以下两种情况不符合条件：[21] 第一，如果该经济独立单元由一人构成，而此人在别处与两位或两位以上的家人属于一个经济共同体，那么，该经济独立单元不符合我们“家庭户”的条件，而是按照“少数服从多数”的原则，被认为是别处的多人家庭户中的一名家庭成员。第二，我们要求“家庭户”必须是中国大陆地区的家庭户。具体判断标准是该家庭户中至少有一名成员拥有中国国籍（不含港、澳、台）。

最后，如果经过上面两个环节我们界定出在一个住宅地址上存在多个符合条件的住户，这时计算机系统会自动运行随机抽样的程序，抽取其中的一户作为受访家庭。不过，对于所有符合条件的住户，我们在住户过滤问卷中都会询问其在本村/居以及国内其他地方的房屋拥有情况。

3.5 家庭成员问卷

3.5.1 基线调查家庭成员身份界定

在满足“同灶吃饭”[22] 的前提条件下，CFPS 2010 年基线调查的家庭成员含两类，第一类是所有的直系亲属，第二类是截至调查时在该家庭内居住时间满 3 个月的非直系亲属。我们把 2010 年这些家庭成员以及他们今后的新生血缘/领养子女全部视为 CFPS 基因成员。所有基因成员均为 CFPS 永久追踪的对象。除家庭成员外，在 2010 年我们还界定出一

[21] 最初住户过滤的设计还要求受访家庭户中至少有一名成员在抽样社区居住时间满 6 个月，但在执行过程中，这一条件被取消，实际被这一条件过滤掉的仅有极少数家户。

[22] “同灶吃饭”指经济联系在一起的家庭和非家庭成员，包括了有血缘/亲缘关系的成员以及在家里工作的非血缘/亲缘关系的成员如保姆、司机、担任保姆工作的远房亲戚等。

类长期共同居住（居住时间满 6 个月）的非家庭成员，在家庭成员问卷中收集了他们的一些基本社会人口信息。他们与家庭并没有血缘/婚姻/领养的紧密关系，因而不是我们关注的主要人群，不需要回答个人问卷。他们一旦离开了受访家庭，我们也不会对其进行追踪。家庭人员身份界定的流程图见图 9。

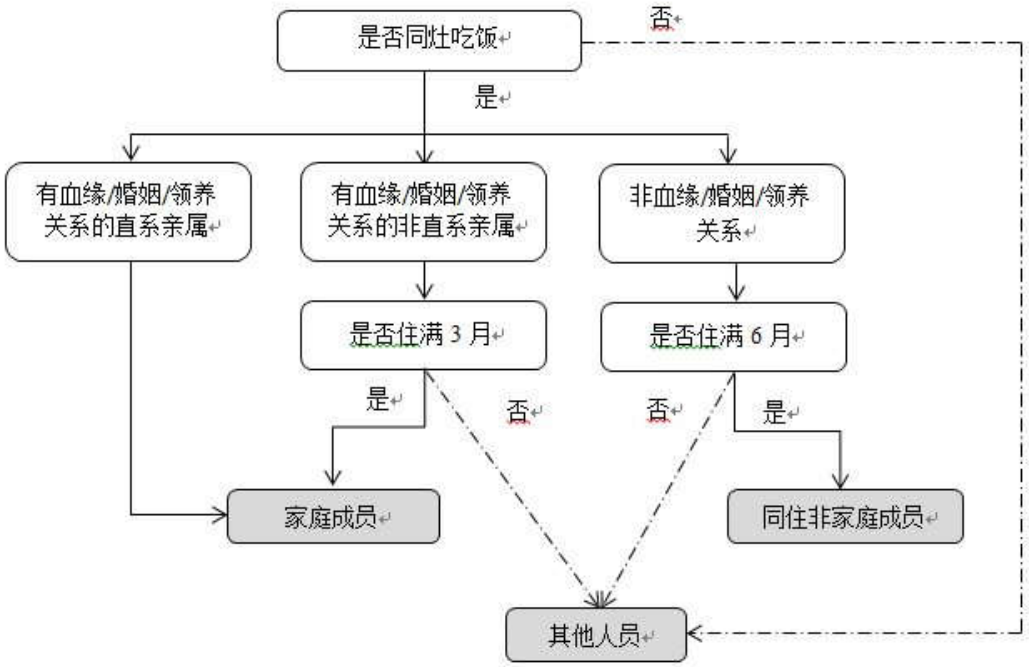


图 9. 2010 年基线调查家庭人员性质界定流程图[23]

在基线调查以后的追踪调查中，我们需要进一步考虑家庭结构的变化，并基于此对家庭成员问卷做出调整。中国家庭的复杂性增加了调查的难度，尽可能全面地收集不同家庭的变迁信息，同时保证实施过程中的可行性，是我们对家庭成员问卷进行调整的出发点。在 2010 年基线调查所确定的基本原则的基础上，我们在之后的调查中对家庭成员的界定增加了若干可操作的实施细节。我们将在下文详细介绍。

3.5.2 家庭成员问卷内容

家庭成员问卷的回答人必须满足两个基本条件：一是同灶吃饭的成员之一；二是与家庭具有血缘/婚姻/领养关系。

[23] 此图在孙妍等（2011，p.126）基础上调整。

家庭成员问卷的主要目的是界定样本家庭的内部关系网络，我们将在接下来的章节中中对 CFPS 采集家庭关系信息的具体方法进行详细介绍。同时，家庭成员问卷还在家庭层面采集所有家庭成员以及同住非家庭成员的一些重要的社会人口信息，如性别、年龄、教育程度、婚姻状况、职业、户口、居住地等，以方便研究者了解家庭的全貌。而且，多途径的信息采集也可以确保 CFPS 数据库中重要社会人口信息的完整性。

3.5.3 T 表格设计理念与操作方案

提到 CFPS 家庭成员问卷，我们必然要提及 T 表格的设计，这是 CFPS 首创的一套家庭关系调查理念与设计方案。

我们知道，在以往的社会调查中，对家庭关系信息的收集一般仅限于在个人层次的问卷中直接询问受访者本人的父母、配偶、子女的基本情况。这样做的缺陷在于：第一，这些调查通常对每户只抽取一名受访者，并以这名受访者为家庭关系的中心，提问其他人员与该受访者相对应的亲属关系。这种做法首先假定了家庭关系中只有受访者一个核心，一般会采用随机抽取一名成员或人为指定“户主”的方式来确定这一核心。但无论采取何种方法，核心的确立实际上并没有多大的意义，因为逻辑上家庭中的每一个人都可以作为家庭关系的核心，家庭关系应该是由多个核心连接起来的树状的网络(family tree)，而以往常见做法收集到的却是一个从单一核心出发（如户主或受访者）的辐射状结构，这个结构仅是树状家庭网络中的一小部分。第二，由于常见做法收集到的是一个单一核心的辐射状结构，且以户主或受访者为核心，研究者从中只能了解到每个（或几个）家庭成员/亲属与受访者之间的关系，却无法获知除受访者以外的这些家庭成员/亲属之间的关系。第三，以往调查通常只是笼统地询问受访者的父母、配偶、子女的情况，不询问他们的姓名，也没有专门的编号来进行识别。所以，即使受访家庭中有多个受访者，也无法通过姓名或编号将每个受访者填写的家庭关系联系起来。第四，以往调查主要收集的是同辈（如兄弟、配偶）或上下代（如父母、子女）的信息，由于没有收集亲属之间的关系、也没有收集亲属的姓名，因此，从中无法获得跨代的信息。[24]

CFPS 在 2010 年基线调查中使用的 T 表格设计很好地解决了这些问题。T 表格由 T1、T2、T3 三张表构成，位于家庭成员问卷的起始部分（图 10）。T1 表（同住家庭成员表）、T3

[24] 技术报告：CFPS-7。

表（不同住直系亲属表）分别记录全部家庭成员和他们不同住的父母、子女和配偶的基本社会人口特征， T2 表（“关系表”）则建立起了 T1 表中全部家庭成员之间的对应关系，以及 T1 表成员与 T3 表成员之间的对应关系。

T 表格的全部信息由家庭成员问卷回答人代答，不要求所涉及的家庭成员及亲属亲自作答。在实地访问中，访员在计算机辅助调查系统的帮助下完成对 T 表格的填写。

首先，在上文介绍家庭成员问卷内容的时候，我们已经了解到，在 T 表格回答之前，家庭成员问卷以经济联系为标准，通过一系列问题对一个家庭中的家庭成员进行了界定。经界定符合家庭成员资格的人员，采用逐人逐项提问的方式，填写 T1 表的内容，最后生成同住家庭成员列表（T1 表），以“1”开头的三位个人码进行标识。此外，T1 表中也为同住非家庭成员生成了相关纪录，以“3”开头的三位个人码进行标识。

然后，根据 T1 表中家庭成员的信息，生成 T2 表的姓名初始列表。T2 表采用“轮流坐庄”的方式，每一个人轮流作为家庭关系的核心（或“庄主”），采集其父母、子女和配偶的姓名，生成家庭成员直系亲属关系表（T2 表）。

最后，调查 T2 表中每一位“庄主”的父母、子女和配偶的基本信息。T2 表提到的父母、子女和配偶，如果已经出现在 T1 表中（即父母、子女和配偶为家庭成员），计算机辅助调查系统会将已有信息自动加载，不需要重复提问；如果没有出现在 T1 表中，系统将会自动生成 T3 表的姓名初始列表，再次采用逐人逐项提问的方式，填写 T3 表的内容，最后生成不同住直系亲属成员列表（T3 表）。T3 表中成员的三位个人码均以“2”开头。

可以看出，三张表整合起来，我们能够得到一个全面的家庭及亲属关系网络，通过这个网络，同代、上下代、隔代的关系均可以关联，同时每位成员的个人基本信息都有详细记录。T 表格的设计不仅在理念上填补了传统社会调查收集的家庭关系不明确的缺陷，为研究者掌握家庭关系的全貌以及全部家庭成员的有效信息、深入利用家庭背景信息研究相关问题提供了更为丰富全面的资料，同时，它所设计出来的具体操作方案也有效避免了重复提问，极大地提高了调查效率。它在基线调查中所界定出的完整、全面的家庭关系也为 CFPS 之后对家庭以及家庭成员的追踪奠定了良好的基础。

T 表格设计方案虽然有诸多优点，而且设计之初已经充分考虑到简便和效率，但是，由于其收集信息之多，相比常规的调查仍需花费更多的时间，操作起来也相对复杂。访员的行为规范与访问技巧直接影响到受访者的配合程度与 T 表格的数据质量，在实地调查之前需要对访员进行严格和完备的培训。

T1 表：同住家庭成员表

个人编码	姓名	出生日期/属相年龄	性别	婚姻状况	最高学历	主要工作	行政管理职务	外出人员信息
101								
102								
...								
301								
302								
...								

T2 表：家庭成员父母、配偶、子女关系表

个人编码	姓名	父亲	母亲	配偶	孩 1	孩 2	...	孩 10
101								
102								
...								
301								
302								
...								

T3 表：家庭成员不同住父母、配偶、子女列表

个人编码	姓名	出生日期/属相年龄	性别	婚姻状况	最高学历	主要工作	行政管理职务	居住地与户口信息
201								
202								
...								

图 10.　T 表格设计图

3.5.4 家庭变迁

CFPS 在个人层面的追踪对象是恒定的，个人一旦成为 CFPS 的基因成员（2010 年家庭成员或其在今后的新生/领养子女），便是 CFPS 永久的访问对象。然而，个人所在的家庭及其成员则是不断变化的：一是家庭单元中成员结构会发生变化，如原有成员去世，或增加新的家庭成员等；二是家庭单元本身会发生变化，如因家庭成员全部去世而彻底瓦解，或因为分家、婚姻而组建新的家庭等。家庭结构作为个人的重要环境信息需要持续采集。与此同时，家庭结构变化本身也具有很高的学术研究价值。

尽管究其实质 CFPS 的追踪对象是个体样本，但在实际操作过程中，CFPS 每轮追踪调查均是以以往调查的家庭为出发点，了解家庭结构的变动及人员的流向。具体而言，CFPS 采用“加减法”的方式收集基线调查之后家庭结构的变化信息。第一步是减法。以最近一轮调查界定的家庭成员为出发点，了解这些成员本轮调查时的住家状态[25]。对于不在家成员，我们根据离家的原因将其分为不需要追踪和需要追踪两类。其中，不需要追踪的成员包括客观上无法追踪的基因成员（如：死亡、离家出走），以及搬到机构居住、超出家庭访问范畴的基因成员（如：出家、服刑、参军/服役、在养老院居住）。需要追踪的成员则根据其是否已经在经济上独立于当前家庭判定其属性：经济独立判定为另组家庭的家庭成员，经济不独立则判定为当前受访家庭的外出人员。扣除不需要追踪成员及另组家庭成员，我们便得到了当前受访家庭的“原家庭成员列表”。

第二步为加法。这一步主要了解被访家庭在两次调查期间新增加的家庭成员。家庭成员判断的标准与基线调查基本相同，具体包括：（1）在家中的基因成员及其在家中的非基因父母、子女和配偶；（2）基因成员在家中的其他直系亲属且其在经济上不独立于该家庭；（3）基因成员在家中的非直系亲属，在经济上不独立于该家庭且居住时间满 3 个月。以第一步得到的“原家庭成员列表”为基础，加上第二步的新增家庭成员，我们就得到了当前受访家庭的“当前家庭成员列表”。

第三步为调整。上述两步能捕捉到绝大多数家庭结构的变化，但这种做法得到的只是基于两个时点的数据，即上轮调查和本轮调查时的家庭结构，对于在两轮调查期间新增又离开的家庭成员无法体现。如果其为上一轮调查后基因成员新生或新领养但在本轮调查中已经离开当前受访家庭的孩子，我们可能会遗漏掉这一需要永久追踪的基因成员。尽管这种情况出现的概率很低，我们依然为此做了第三步的调整。在这一步我们网罗了当前家庭成员中所有基因成员在两次调查间新生但已离开该家庭的血缘子女。对于这部分筛选的离家新增基因成员，同样根据新生孩子所在家庭与当前家庭是否经济独立来判断其个人属性。

最后，经以上三步得出的家庭成员中如果不存在基因成员，我们则对该家庭终止调查。对于另组建的家庭，我们“加减法”的起点是从原家庭分流到新家庭中的基因成员，在此基础上按照上述同样的加减流程构建另组家庭的家庭成员结构。

[25] 对于调查当时短期离家的成员，我们以三个月内会回来，并且会在家中长期居住为标准，请受访者主观判断其是否住家。

3.5.5 经济独立判断

在家庭成员问卷的访问过程中，我们需要对两类变化的人员身份进行判断：一是新进入家庭的人员是否为家庭成员；二是离家成员是否仍属于当前家庭。无论对于哪一类人员，经济联系都是定义其为家庭成员的重要标准之一。但我们对两类人员与家庭的经济联系的衡量标准并不完全一致。对于新进人员是否与当前受访家庭有经济联系的判断标准，历轮追踪调查与基线调查一致，即是否与其他家庭成员同属一个经济共同体（“同灶吃饭”）。但判断离家人员是否仍与原家庭有经济联系，亦或是已经独立成新家庭户则更加复杂。离家成员与原家庭曾属于经济共同体，在中国特殊的社会环境下，即便物理上不在原地址居住，也极有可能存在金钱或供养等各式各样的联系。我们经历了反复的探索和尝试，希望构建客观标准判断离家人员家庭属性。

在 2011 年的样本维护调查中，我们尝试以统一的客观标准来衡量离家人员与原家庭的经济联系：如果平均每年有 1000 元以上的实物或现金往来，则视为经济不独立；反之，则视为经济独立。执行过程中，我们发现经济往来的数额与受访家庭的经济收入以及当地的社会经济发展水平存在密切的联系，并不能真实反映离家人员与家庭的经济依存关系。

在 2012 年追踪调查时，我们放弃了客观金钱数额这一判断标准，转而引入离家原因作为判断依据。有一部分离家原因明显体现了离家人员组成了新家庭，如嫁出、离婚、分家，我们将此类离家人员划定为“另组家庭成员”。另一部分离家原因则不足以判断离家人员是否组成新家庭，如外出读书、工作。我们继而综合了其是否结婚、是否有配偶和孩子留在家中、是否正式工作、是否养家、是否被家里养这几个条件来进行判断。从理论上来说，这似乎是一个综合考虑了中国实际情况的有效区分手段。但是，由于家庭的复杂性，我们的设计依然存在缺陷。由于这种做法是对离家成员逐一进行判断，因此很容易把一个居住共同体上的离家人员割离，并导致本具有相同身份的人员被赋予不同的属性。比如我们把离开原家庭在外正式工作、已经结婚、配偶孩子没有留在原家庭、不养家也不需要家里养的人员界定为另组家庭成员。但假设外出工作的夫妻双方只有丈夫一方供养原家庭，而妻子作为原家庭的媳妇并不直接供养婆家，那么这对原本属于一个家庭的夫妇一人会被判断成原家庭成员，一人会被判断成另组家庭成员，这显然不符合常情，也不利于实际的操作。

基于以上问题，在 2014 年追踪调查中，我们又做了新的尝试。首先，我们不再以单个人，而是以外出居住的地址作为基本的判断单元。我们把每个地址上的人群视为一个整体，调查其整体与原家庭是否已经彼此经济独立，这解决了以往人为割离家庭的缺陷。其次，在

2012 年追踪调查中，离家成员的属性由原家庭的受访者判断，但在 2014 年的调查中，我们引入了双判断的模式，即在原家庭访问时，由当前受访者判断离家人员属性；当访问到离家人员所在的居住单元时，由离家人员自己根据其与原家庭的经济联系再做判断。双判断的模式有助于在访问过程中与受访者达成共识，有利于家庭结构的构建以及后期家庭经济数据采集范围的界定。

3.6 家庭问卷

家庭问卷的主要目的是在家庭层面上收集样本家庭的日常生活、社会交往与经济活动方面的信息。2010 年基线调查的家庭问卷的主要内容见下表 5。之后的追踪调查也基本沿用了这些内容。

表 5. CFPS 2010 年家庭问卷主要内容

模块	问卷内容
A 地理交通	最近的公交、医疗点、高中、商业中心
B 生活条件	用水，燃料，电，卫生间条件，垃圾处理，保姆/小时工雇佣
C 社会交往	春节拜访，送礼，族谱/家谱，祭祖/扫墓，邻里交往，亲友交往
D 住房情况	房屋所有权，自建/购买，租房来源，建筑面积，入住时间，房屋市值与租金，房屋结构，其他房产情况，住房困难情况
E 经营状况[26]	U 外出工作模块（外出人员，工作地址，时间投入，假期是否回家，转移支付情况，家庭是否因其外出而雇佣/增加帮工），政府补助，致贫原因，V 非农经营模块（非农产业类型、数量、参与者、总资产、家人拥有股份、雇佣人数，营业额，税后纯利润），房屋出租，土地与其他生产资料出租，财物出卖，拆迁，土地征用
F 家庭收入	存款，金融产品，离退休金/社会保障金/低保收入，工资/奖金/补贴/红利等收入，非工资性/农业生产收入，礼金/礼品折现
G 家庭资产	保险可赔偿额，他人欠款，收藏品价值，其他资产现值
H 家庭支出	最贵消费品花费，借贷款，家庭各项日常支出（食品、出行、通信等），家庭各项特殊支出（家电、医疗保健、教育、商业保险等），捐赠，总支出
J 耐用品	汽车，摩托车，拖拉机，电视
K 农业生产	土地类型，土地数量，农业收支状况，农林作物类型、产量、销量、收入，家畜与渔业类型、产量、销量、收入，家畜饲养条件
Z 访员观察	问卷回答人，家庭住房条件，家庭整洁度，家庭成员精神面貌，家庭成员间关系，长幼关系，性别间关系，受访者个人特征

[26] 模块 E 中镶嵌了 U 外出工作、V 非农经营两个模块。家庭如有外出工作人员或从事非农产业，则转入相应的 U 模块或 V 模块，回答完该模块问题后，再回到 E 模块，继续回答其它问题。

家庭问卷由家庭中最了解相关家庭事务的人员回答，因而可以是由一名对家庭各方面情况都非常了解的受访者独立完成问卷，也可以由多名受访者轮流完成相应模块的问题，如由农业活动的管帐人回答农业活动相关的问题，由家庭私营企业的管账人回答家庭私营企业相关的问题，由食品采购人负责回答家庭支出相关的问题，等等。尽管家庭问卷的内容在历轮调查中基本稳定，但我们对具体的测量方法做过几次调整。变动较大的是收入和支出部分。CFPS 在不同年份对家庭收入与支出的分项设计分别如表 6 和表 7 所示。总的来说，CFPS 2010 采用了较为粗略的分项设计。为了对数据的采集更为完整，CFPS 2012 采用了非常精细的分项设计。CFPS 2014 则重新整合了一些过于细分的收支项目，同时对若干收支项目的提问进行了优化。

在家庭收入上，CFPS 2012 在以下几个方面对 2010 年的问卷做了调整：（1）增加了 2010 年调查中遗漏掉的部分收入项目。例如，2010 年没有直接提问供自家消费的农产品的价值、个体经营的利润、农业打工收入以及在学者的奖助学金和实习/兼职收入等。CFPS 2012 明确提问了这些项目。（2） 在 2010 年基础上对收入大类别下的具体内容进行了细化。以农业生产收入为例，2010 年调查只是笼统地询问农户从事农、林、牧、副、渔业的全部毛收入和总成本，而 2012 年调查则分别询问了农户从事种植业和养殖业的销售收入、自家吃用部分的收入以及各生产环节的成本。又如，在调查家庭的政府转移性收入时，2010 年调查只用了一道题笼统地询问家庭离退休金/社会保障金/低保收入的总额，而 2012 年调查则对该类收入细化，先让受访者列举收到的政府补助项目，再逐项提问家庭从该补助项目中得到的收入。这些对收入内容的列举与细化有助于受访人回忆和明确需要回答的收入内容，避免遗漏。（3）对多项收入增设分级展开的逼近式提问法(unfolding brackets)。该方法在受访人无法回答或拒绝回答某项重要收入的具体金额时，会进一步要求受访者选择其在该项上的收入所属的级别（即区间）。这一设计能够降低提问的敏感性，减少数据缺失。CFPS 在 2010 年调查收入时仅对工资性总收入使用了分级展开提问，而 2012 年调查对农业总收入、个体经营及办私营企业收入、个人的工资性收入均增设了分级展开提问。我们对分级展开提问所获得的数据的处理方法是对相应收入区间的上限和下限取均值，以此替代未作答的具体收入金额。4) 调整工资性收入的提问方式。2010 年调查是在家庭层面上提问全家每一位家庭成员的工资性收入，即由家庭问卷的回答人先逐一回答每一位家庭成员的工资性收入或收入区间，再回答所有家庭成员的工资性收入的总额或总额的区间。2012 年调查则将工资性收入分散到每一位成员的个人问卷中提问。因此，整个家庭的工资性收入需要通过加总每一份个人问卷

的工资性收入来得到。

2012 年对收入项目的设计尽管可以采集到更多的信息，但却增加了访问时长与受访者回答的难度。我们因此在 2014 年又对一些过细的分项进行了合并。与此同时，我们调整了少量收入指标的提问时间区间。关于收入的数据清理过程及变量信息，请参考 7.4 节。

表 6. CFPS 家庭年收入分项设计

2010	2012	2014
Ⅰ.经营性收入 1. 农业收入[27] - 农林牧副渔纯收入 ■ 农林作物净收入 ■ 畜牧渔业净收入 2. 每一项私营企业的净利润	Ⅰ.经营性收入 1. 农业收入（产品净收入与自家消费） - 种植业和林产品 - 畜牧和水产品 2. 每一项个体经营或私营企业的净利润	Ⅰ.经营性收入 1. 农业收入（产品净收入与自家消费） - 农副产品总值 2. 全部个体经营或私营企业的总净利润
Ⅱ.工资性收入 1. 工资性收入（含工资、奖金、补贴、分到个人名下的红利等） 2. 外出打工者寄钱[28]	Ⅱ.工资性收入 个人问卷采集 1. 个人农业打工（农活/打散工）收入 2. 个人从事每一份受雇非农工作的收入 工作的税后收入 工作的实物形式福利 3. 个人接受正式教育期间勤工助学/实习/兼职收入	Ⅱ.工资性收入 1. 帮其他农户干农活/打工收入 2. 外出打工者寄钱[29] 3. 受雇非农工作全部工资性收入（含工资、补贴、奖金、实物福利）
Ⅲ.转移性收入 1. 家庭全部离退休金/社会保障金/低保等收入 2. 政府补助总收入（含现金与实物）	Ⅲ.转移性收入 1. 退休退职人员的养老金/退休金（个人问卷采集） 2. 政府补助 - 低保 - 退耕还林补助	Ⅲ.转移性收入 1. 家庭全部养老金（退休金） 2. 政府补助总收入（含现金与实物）

[27] 2010 年问卷没有采集自家农产品消费收入，但在数据清理过程中对这一项进行了估计。详情参见 7.4 节。

[28] 2010 年数据清理时并未将此项额外加进家庭收入的计算。

[29] 2014 年数据清理时并未将此项额外加进家庭收入的计算。

	- 农业补助 - 五保户补助 - 特困户补助 - 工伤人员供养直系亲属抚恤金 - 救济金、赈灾款（含实物形式） - 其他政府补助 3. 捐助或补偿 - 社会捐助（包括现金和实物） - 征地补偿金 - 住房拆迁补偿金 4. 接受教育期间的奖学金/助学金	3.捐助或补偿 - 社会捐助（包括现金和实物） - 征地补偿金 - 住房拆迁补偿金（包括现金和房产等）
Ⅳ.财产性收入 1. 房屋出租总收入	Ⅳ.财产性收入 1. 房屋租金收入 - 自家正在居住房屋每月出租收入 - 其他房产每月出租收入	Ⅳ.财产性收入 1. 房屋出租总收入
2. 出租土地或其他生产资料的总收入	2. 出租土地收入 - 出租自家集体分配土地 - 转租已租用土地	2. 出租土地收入 - 出租自家集体分配土地 转租已租用土地
3. 出租家里其他东西总收入	3. 出租其他家庭资产（如设备等）收入	3. 出租其他家庭资产（如设备等）收入
4. 出卖财物（家里东西）的总收入		4. 投资收入[30]
V.其他收入 1. 收到的礼金/礼品收入及其他收入	V.其他收入 1. 私人性经济支持或赠与 - 不同住亲戚的经济支持和赠与 - 其他人的经济支持和赠与	V.其他收入 1. 私人性经济支持或赠与 - 不同住亲戚的经济帮助（现金与实物） - 其他人的经济帮助（现金与实物） - 重要事件人情礼收入

[30] 数据清理时并未将此项加入家庭收入。

支出模块在历轮调查中的调整与收入模块类似。CFPS 2010 的调查问卷采用表格的形式汇总了家庭的各项支出，包括家庭日常支出（问卷中以月计的数值换算为以年计）和特殊支出（问卷中以年计）。鉴于 2010 年的调查项目过于笼统，CFPS 2012 在维持 2010 年问卷中支出大类不变的情况下，对支出大类别下的具体内容进行了细化，并通过列举的方式帮助受访人回忆，如，在提问家庭购买日用品的支出时提示受访者日用品包括洗衣粉、香皂、肥皂、牙膏、牙刷等。CFPS 2014 采集家庭支出数据的问卷内容与 2012 年基本一致，但在个别项目的提问方式上进行了优化，如合并了一些过于细碎的支出项目，修改了部分项目的回忆时段，改进了软检查的值域设置，补充提问了 2012 年遗漏的重大事件支出项目，以便受访者更好地回忆和作答。

表 7. CFPS 家庭支出分项设计

2010	2012	2014
Ⅰ.生产与经营支出 1.农林牧副渔经营总成本	Ⅰ.生产与经营支出 1.种植业与林业投入 - 种子化肥农药 - 雇工役畜 - 机器租赁与灌溉 - 其他 2. 畜牧水产 - 种畜鱼苗 - 雇工役畜 - 饲料 - 其他	Ⅰ.生产与经营支出 1.种植业与林业投入 - 种苗化肥农药 - 雇工役畜 - 机器租赁 - 灌溉 - 其他 2. 畜牧水产 - 种畜鱼苗 - 雇工役畜 - 机器租赁 - 饲料 - 其他
Ⅱ.食品支出（上个月） - 食品支出	Ⅱ.食品支出(过去 1 周) - 外出就餐（含请客吃饭） - 购买自家消费的烟酒 - 购买自家消费的其他食品 - 消费自家产的农副产品的价值	Ⅱ.食品支出(过去 12 个月平均每月) - 总伙食费（含购买自家消费的零食饮料烟酒） - 外出就餐
Ⅲ.日常生活支出（上个月）	Ⅲ.日常生活支出（过去 1 月）	Ⅲ.日常生活支出（过去 12 个月平均每月）

- 通信支出	- 邮寄、通讯支出（含电话、手机、上网、邮寄等）	- 邮寄、通讯支出（含电话、手机、上网、邮寄等）
	- 水费、电费	- 水费 - 电费
	- 燃料费	- 燃料费
- 出行支出（含养车费）	- 本地交通费（含汽车油费）	- 本地交通费（含公交费、汽车和摩托车油费）
- 日常用品	- 日用品	- 日用品
- 房租	- 房租	- 房租
- 雇佣保姆、小时工	- 雇佣保姆、小时工、佣人	
	- 文化娱乐	
	- 购买彩票	
- 赡养支出		
- 住房按揭		
- 车辆按揭		
- 其它按揭		
IV.长期生活支出（过去一年）	IV.长期生活支出（过去一年）	IV.长期生活支出（过去 12 个月）
- 衣着	- 衣着鞋帽	- 衣着鞋帽
- 文化娱乐休闲		- 文化娱乐
	- 旅游	- 旅游
- 居住支出（取暖、物业等）	- 集中供暖	- 集中供暖
	- 物业费（含车位费）	- 物业费（含车位费、卫生费）
- 购房建房（不含房贷）		- 房贷 - 住房维修、装修
	- 购买汽车	- 购买、保养、维修汽车
	- 购买、维修其他交通、通讯工具及配件	- 购买、维修其他交通、通讯工具及配件
- 家电	- 购买可办公类电器	- 购买及维修家具、电器其他耐用消费品
- 家庭杂项商品、服务支出	- 购买家具及其他耐用消费品	
- 教育支出	- 教育支出	- 教育支出
- 医疗保健	- 医疗支出	- 医疗支出
	- 保健费用	- 保健费用
	- 美容支出	- 理发、美容支出
- 购买商业性保险	- 购买商业性医疗保险	- 购买商业性保险
	- 购买商业性财产疗保险	
	- 给不同住亲戚的经济支持和赠与	- 给不同住亲戚现金或实物经济帮助

	- 给其他人的经济支持和赠与	- 给其他人经济帮助
- 现金与实物社会捐助	- 现金与实物社会捐助 - 税费与杂费 - 租用土地 - 租用了其他家庭资产（如设备等）	- 现金与实物社会捐助 - 租用土地
- 其他支出	- 其他支出	- 其他支出
V 重要事件支出 - 自家婚丧嫁娶支出 - 亲朋好友人情礼		V. 重要事件支出 - 自家宴请与办仪式总花费 - 亲朋好友人情礼
去年家庭总支出确认		过去 12 个月总收入与总支出确认

3.7 个人问卷

3.7.1 设计原则

CFPS 将 16 岁以下的人群定义为少儿，16 岁及以上的人群定义为成人。少儿问卷与成人问卷是分别为这两类人群设计的个人访问问卷。

在介绍个人问卷内容之前，我们需要先介绍如下设计原则：

首先是年龄的计算方法。由于 CFPS 以 16 岁为区分点为少儿与成人分别设计了不同的问卷，而少儿问卷本身也根据不同的年龄段设计了不同的模块，因此，年龄的计算方法便尤为重要。在计算年龄时，我们只考虑出生的年份，不考虑出生的月份，具体计算方法为：调查年份-出生年份[31]。如，2000 年 10 月出生的孩子，当我们 2010 年 7 月份进行调查的时候，虽然距离他满 10 周岁还差 3 个月，但在 CFPS 调查中，我们将其算作 10 岁。这个计算年龄的规则在 CFPS 其它类型的问卷中以及今后的调查中依然适用。

其次是个人自答问卷的回答人。成人自答问卷由本人完成。少儿问卷由自答和代答两大部分组成。[32]其中，代答部分覆盖全部年龄段的少儿，由其家长完成；自答部分仅覆盖 10-15 岁的少儿，由其本人完成。“家长”指与被访少儿同住的最主要的监护人，即照顾小孩最多、最了解小孩情况的人。

自 2012 年起，除了个人自答问卷外，我们设计了个人代答问卷，代答问卷主要适用于

[31] 由于个人问卷是在家庭成员完成时候生成的，因此用来计算年龄的调查年份是指家庭成员完成时的年份，有可能与个人问卷完成时的年份不同。
[32] 在此提醒用户注意：少儿自答问卷中的代答部分有别于后文提到的少儿代答问卷。

两种情况。一是针对家庭成员中物理外出的人员，由住家成员完成一份外出人员的代答问卷，外出人员自身再提供一份自答问卷；另一种情况是由于客观条件限制而无法回答自答问卷的家庭成员（如语言或思维能力障碍等），由了解该家庭成员的家庭成员完成代答问卷。代答问卷为自答问卷的简化版本，并充分考虑了代答的性质，如，删除了无法由他人代答的主观态度题部分。代答问卷的设计有助于我们最大程度地收集到基因和核心成员的关键个人信息。

3.7.2 问卷内容

在 2010 年基线调查中，我们为 0-1 岁（不含 1 岁）、1-3 岁（不含 3 岁）、3-6 岁（不含 6 岁）和 6-16 岁（不含 16 岁）的少儿的自答问卷分别设计了相应的代答模块，具体的结构流程与访问内容见图 11 上半部分。可以看出，问题的内容随着年龄段的增大而累加。0-1 岁的少儿需要回答的问题最少。6-16 岁的少儿需要回答的问题最多：在包含了前面 0-1 岁、1-3 岁、3-6 岁的少儿需要回答的全部问题的基础上，我们还会访问该少儿的上学情况以及父母对其学习与生活的关怀情况，同时，我们还需要父母对该少儿的行为习惯做出一些判断。自 2012 年追踪调查起，我们将这些模块进行了整合，通过逻辑跳转的方式过滤掉不适应相应年龄段的问题或在之前调查中已经回答过的问题。这一做法保留了原有的问题内容，但使问卷形式更为简洁，同时也可以避免少儿从一个年龄段进入到另一个年龄段后信息的重复采集。

2010 年基线调查的少儿自答问卷的自答部分的具体内容见图 11 的下半部分。在之后的调查中，问卷的内容基本保持稳定。

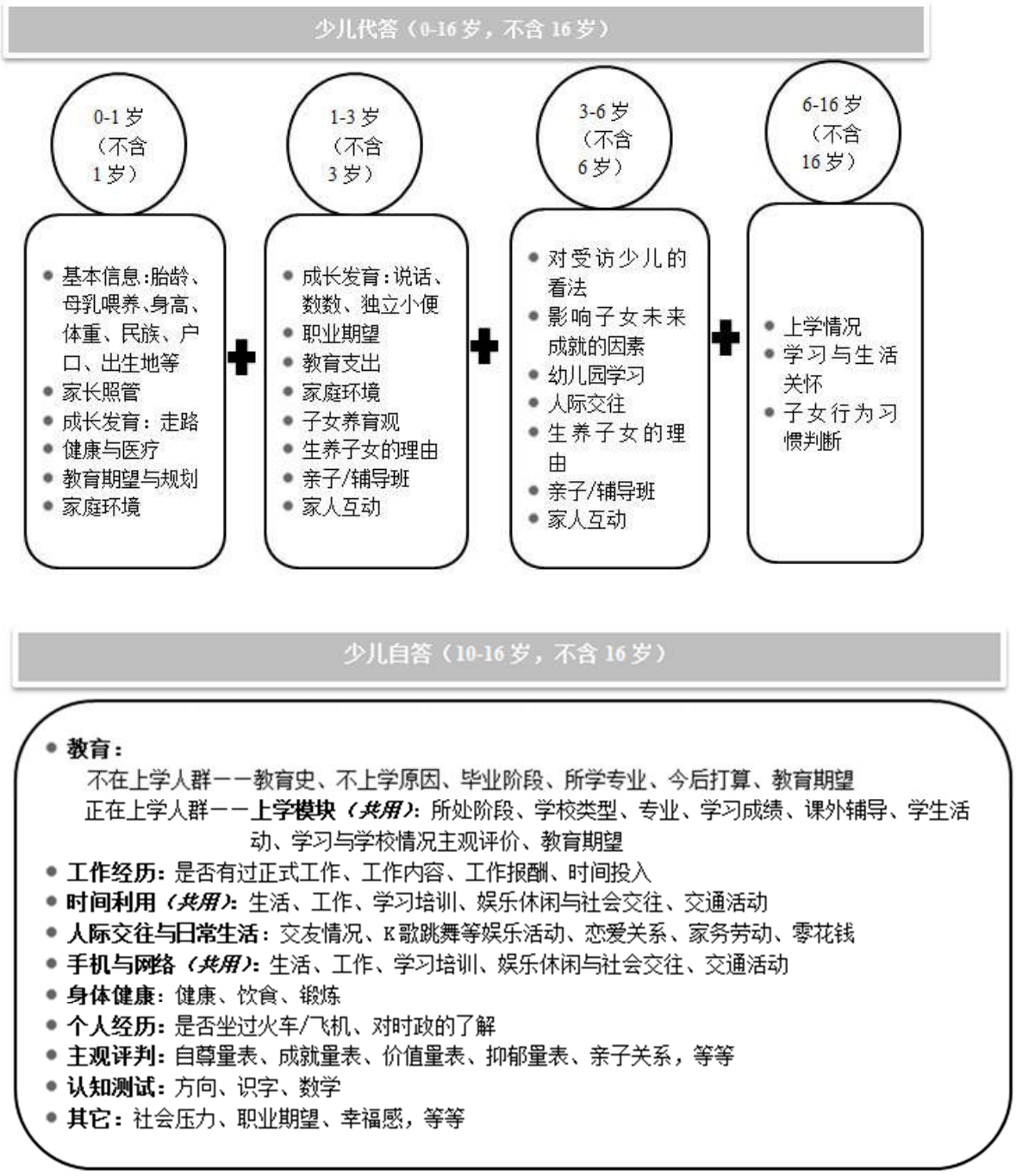


图 11. CFPS 2010 年少儿问卷结构流程与访问内容[33]

CFPS 的成人自答问卷的内容也基本保持稳定，基线调查的具体内容见表 8 第 2 列。但为了适应追踪调查数据的采集模式，CFPS 将需要在基线调查以后的调查中继续使用的问题分为基线、核心、轮替和扩展四类题组。基线题组仅适用于初次接受个人问卷访问的受访者，问题均为回溯性的客观题。核心题组为每轮调查需重复提问的问题，内容涵盖较

[33] 此图在孙妍等（2011，p.151）基础上调整。

为广泛，主要用于测量受访样本在相应变量上的变化情况。轮替题组根据既定规则选择性地用于不同轮次或不同受访者。CFPS 有两类轮替题组：一类是根据访问轮次决定提问内容；另一类则是根据受访者的基本特征，如所处年龄组，决定提问内容。扩展题组是根据调查执行当年社会及学术研究的热点在问卷中搭载的少量扩展专题研究。此类题组通常仅在一轮访问中出现。我们对成人问卷主要内容的题组划分见表 8 最后一列。此外，在 2012 年和 2014 年的追踪调查中，我们增加了少数调查问题。表 9 罗列了这些新增加的问题及其所属的相应题组。

表 8. CFPS 2010 年基线调查成人自答问卷主要内容及后期题组划分

模块	调查内容	题组类型
基本信息	出生日期，出生体重，出生地，居住地，户口，民族，文革家庭成份，政党与社团组织，3 岁以前和 4-12 岁时与父母一起居住的时间	基线题组
兄弟姐妹情况	兄弟姐妹数目、名字、出生日期、是否健在、去世年龄与原因、婚姻状况、最高学历、职业、行政/管理职务、居住地，在世父母和谁住在一起，去世父母的去世原因	（仅在 CFPS 2010 使用）
教育史	已完成的最高学历，小学至已完成的最高学历之间的各个阶段的学校类型、学习时间、结束时间、学校名称、是否毕业、学科与专业等，教育期望	基线题组
语言运用	各类语言重要程度，与家人交流的语言	核心题组
上学模块	当前正在上学所处阶段，学校类型，专业，学习成绩，课外辅导，学生活动，学习与学校情况主观评价，教育期望	核心题组
婚姻	婚姻状况（未婚/在婚/同居/离婚/丧偶），现任/前任/初婚配偶/同居对象的出生年月、结婚/同居时间、婚前同居情况、如何认识，前次/初次婚姻解体的原因与时间	核心题组
子女关系	60 岁以上受访者与子女关系评价，与子女间的交往活动	核心题组
工作	见图 12	核心题组
个人收入	非经营性收入，经营收入，亲友资助，国家政府补贴救济	核心题组
时间利用	生活，工作，学习培训，娱乐休闲与社会交往，交通活动	核心题组
娱乐休闲	闲暇活动，频率，出行方式，出国经历	核心题组
手机网络	手机使用情况，社交网络使用、邮箱使用情况，网络重要性评估，上网频率与地点	核心题组
社会关系	找人帮忙，烦恼倾诉，社会地位自评	核心题组
主观测量	价值观，社会观，成就量表，生活满意度等	核心题组
政治	遭遇偷抢威胁的经历，不公正待遇，新闻关注，政府工作评价	核心题组
健康	身高，体重，健康自评，身体不适，慢性疾病，住院经历，医疗费用，病痛处理方式，对医疗状况的满意度，中医，体育锻炼，饮食，P-ADL，吸烟喝酒经历，睡眠，记忆力，生病时主要照料人，身体机能	核心题组
心理健康	K6 量表、CESD 量表	轮替题组

认知测试	识字、数学、记忆、数列	轮替题组
个人信息与访员观察	联系信息，问卷回答人，受访者个人特征	核心题组

表 9. CFPS 成人追踪问卷新增内容

2012 年内容扩展		
去世兄弟姐妹情况	基线调查已去世兄弟姐妹最高学历、职业	（仅在 CFPS 2012 使用）
养老保险	各类养老保险参保、缴费及领取情况	扩展题组
生育意愿	理想孩子个数	核心题组
信任度	对父母，邻居等几类人的信任程度	核心题组
宗教信仰	宗教信仰及宗教活动参与频率	核心题组
第三方健康评价	对问卷设定人物健康状况的评价	核心题组
2014 年内容扩展		
父母信息	父母出生年，受访者 14 岁时父母的职业、政治面貌	轮替题组
风险偏好	风险偏好试验、金融知识模块、法律模块	扩展题组
[EHC-RESI]	EHC 迁移模块	核心题组
[EHC-Marriage]	EHC 婚姻模块	核心题组
[EHC-Job]	EHC 工作模块	核心题组
家庭决策	家庭事务谁说了算	核心题组
婚姻满意度	对婚姻/同居生活以及配偶/伴侣的满意程度	核心题组
政治选举	是否参与选举投票	核心题组
社会治安	对居住地治安状况及司法公正的态度	核心题组
阅读	过去 12 个月的阅读量	核心题组
传统观念	对亲子关系和性别分工等传统观念的态度	核心题组

3.7.3 测量方法更新

尽管 CFPS 个人访问的主体内容在历轮的调查中基本稳定，但为了适应数据采集的需要，我们对一些问题的测量方法进行了更新。下面我们将具体介绍。

3.7.3.1 工作模块更新

a. 2010 年基线调查流程与内容

图 12 展示了工作模块在 2010 年基线调查中的提问流程与内容。后期的更新以这些内容为基础展开。

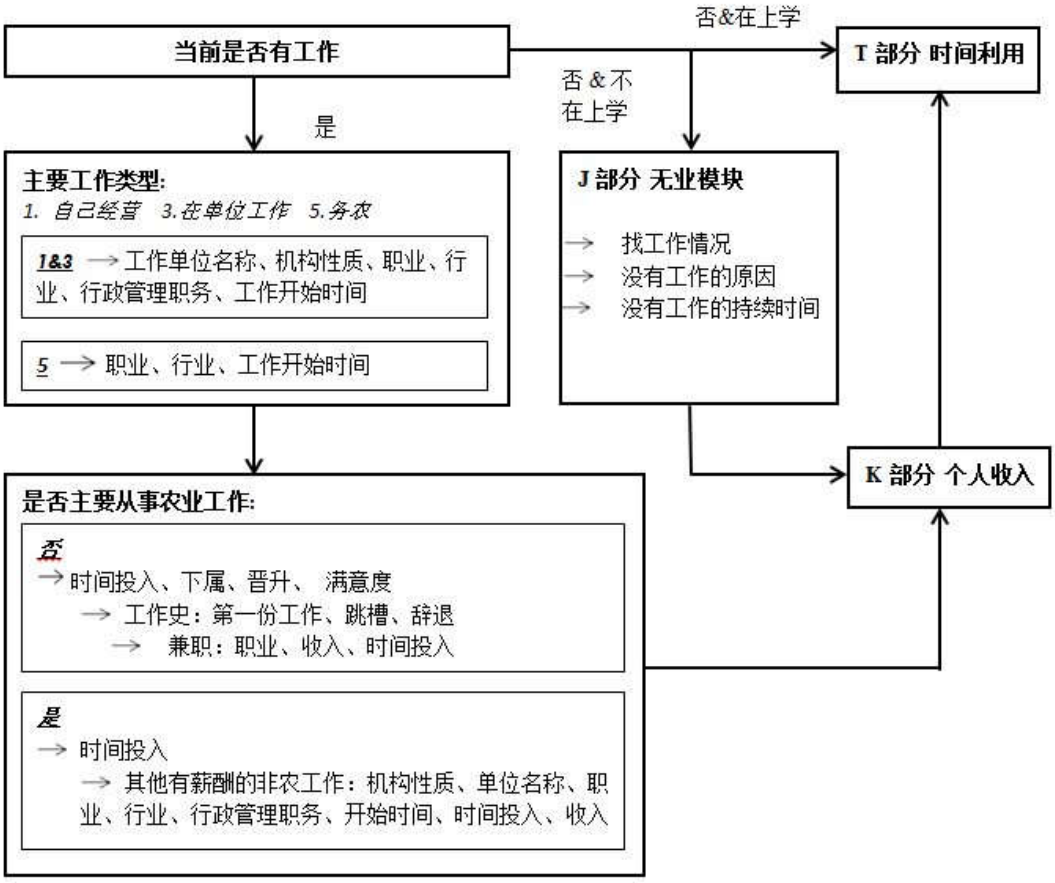


图 12. CFPS 2010 年成人问卷工作模块结构流程图

b. 就业状态界定

对受访者就业状态的测量通常有两种手段。比较简单的方法是主观测量，即直接提问受访者当前是否有工作，由受访者自行判断其就业状态。但通用的、规范的方法是客观测量，即提问一系列客观的问题以判断受访者当前的就业状态。CFPS 在 2010 和 2011 年的调查中采用了简易的主观测量方法，但从 2012 年起改用客观测量的方式。CFPS 参照了国际劳动组织（International Labor Organization，ILO）对于就业定义的标准，综合参考了 CPS（Current Population Survey）、CULS（China Urban Labor Survey）、HRS（Health and Retirement Survey）、CHARLS（China Health and Retirement Longitudinal Study）等调查的设计思路，并根据 CFPS 调查人群的特点进行了相关设计。

CFPS 将自家农业生产经营活动、农业打工、非农受雇、个体/私营/自雇工作均算作工作，但不包括家务劳动和义务的志愿劳动。界定受访者当前的就业状态的流程及判断标准参见图 13。

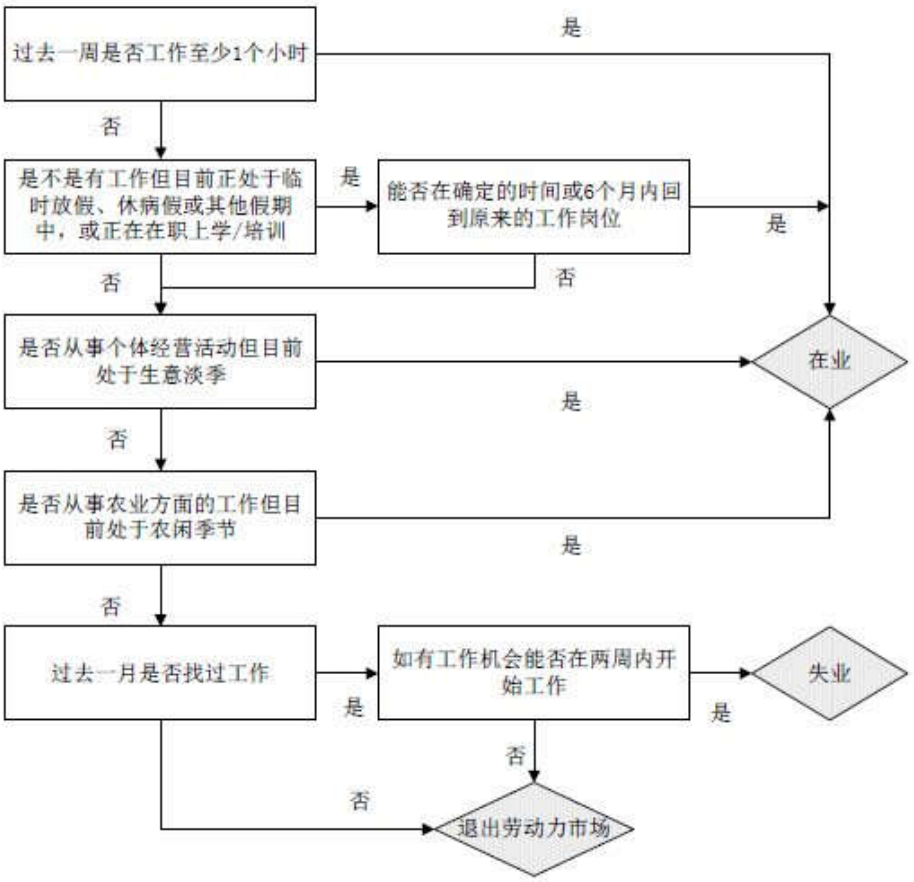


图 13. CFPS 就业状态界定流程图

c. 工作类型判断

不同类型（农业、非农业、受雇、自雇）的工作由于性质的不同，其提问的方式与内容也有很大差异。如，农业工作的收入结构很少涉及到保险、福利、奖金、公积金等，而且通常以家庭为生产单位计算；而对于受雇工作，保险、福利、奖金、公积金等则是收入结构的重要内容，且这些收入往往都是个人收入。因此，工作类型也是决定工作信息采集的一个关键变量，工作类型的错误界定会导致提问流程的错误，以及大量不适用问题的产生，不仅影响数据质量，而且影响访问的进展。

如图 12 所示，CFPS 在 2010 年仅询问当前最主要的一份工作的信息，对工作类型仅简单划分为农业工作与非农工作两类。由于对每类工作采集的信息均较粗略，且没有个性化的问题，所以这样的简单分类方式没有太大问题。

自 2012 起，CFPS 希望收集全部工作的具体情况，这对工作类型划分的精确性提出更高的要求。2012 年采用过滤方式进行提问，首先请受访者逐一判断自己是否从事以下五种

类型的工作：自家农业生产经营活动、农业打工工作、非农受雇工作、个体/私营经济、帮工活动，然后针对受访者自报的每一类工作展开提问。实际访问过程中，我们发现相当一部分受访者并不了解其工作的类型，无法在现场访问中作出准确判断，由此导致重报或漏报。

2014 年我们做了新的设计尝试。我们在内容上保留了 2012 年复杂的工作类型与个性化问题设计，但我们不再把受访者对工作类型的主观判定作为关键过滤问题，而是对每一份工作直接提问两道客观题（表 10），然后由计算机程序根据受访者的回答判定其所属工作类别，再针对其所属的工作类别展开提问。工作类型判定标准参见表 10。

表 10. 工作类型的基本判定

为自己/自家干活还是受雇于他人/他家/组织/单位/公司？	是农业工作还是非农工作？	
	农业工作	非农工作
为自己/自家干活	类型 1：自家农业生产经营	类型 2：私营企业/个体工商户/其他自雇
受雇于他人/他家	类型 3：农业打工	类型 5：非农散工
受雇于组织/单位	类型 4：农业受雇	类型 4：非农受雇

2014 年在提问雇主性质的基础上，根据受访者所从事工作本身的性质，定义了农业工作与非农工作。这一设计弥补了 2012 年调查中农业工作与非农工作的定义没有统一标准的缺陷。需要说明的是，对于受雇工作和自雇工作判断农业及非农标准上略有不同。对于受雇工作来说，农业工作与非农工作指代的是雇主的性质，比如农业打工的雇主必须是农户，但是工作内容既可以是“做农活”（指从事与农业相关的生产劳动），也可以是“打散工”（指的是做一些时间短、比较琐碎的事情）；再比如，非农受雇必须是受雇于非农户的个人、组织、企业或者单位，属于雇佣与被雇佣的关系，而受访者从事的工作本身可以是农业活动，也可以是非农业活动。对于自雇工作来说，农业与非农的界定则更加模糊，比如自家农业生产经营活动指代的是“与农业相关的生产活动以及在此基础上衍生出来的相关经营活动”，既可以是自己栽培苹果的农业活动，也可以是出售自家生产的苹果的经营活动。再比如，个体或私营经济活动既可以是“擦皮鞋的路边小摊”，也可以是“租赁土地、雇佣当地农民进行生产的方式开发苹果园并进行苹果生产经营”。

d. 工作信息扩展

自 2012 年开始，CFPS 每轮的调查均采集受访者两次调查期间所有工作的情况。但由于访问时间有限，我们不可能采集受访者每一份工作的详细信息，折衷的解决方法是对工作区分主次。我们在每轮调查中都会采集受访者从上次到当次调查期间最主要的一份工作的详

细信息，而对两次调查间的其他一般工作则只采集几个比较关键的变量信息。主要工作与一般工作在信息采集上的差异见图 14。

CFPS 参照 PSID 的方法界定当前主要工作，具体是：（1）如果受访者调查当时只有一份工作，当时的工作即作为主要工作；如果调查当时有多份工作，由受访者主观判断其中哪一份为最主要工作。（2）如果受访者调查当时没有工作，选择其最近结束的一份工作作为主要工作。如果有多份工作在最近同一时间结束，由受访者主观判断其中哪一份为最主要工作。

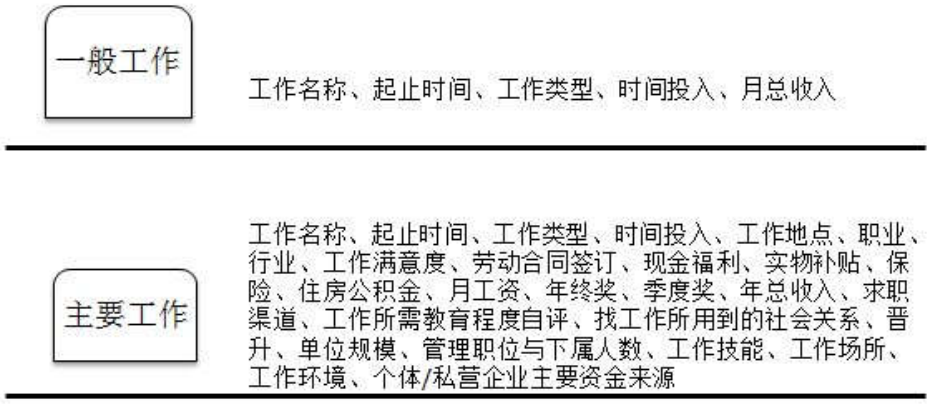


图 14. 一般工作与主要工作的信息采集内容

3.7.3.2. 事件日历记录法

事件日历记录法（Event History Calendar，EHC）是解决复杂信息采集、帮助受访者更好地回忆时间信息的一个有效工具。由于 CFPS 追访的频率从最初计划的一年一次改为了两年一次，为提高信息的准确程度，CFPS 自 2014 年起对居住地变化、工作、婚姻三个模块使用了 EHC 采集信息。图 15 是 CFPS 问卷中 EHC 的一个简单示意图。EHC 在信息采集开始前会根据受访者往期接受访问的情况定制回忆时段[34]，并在计算机屏幕上呈现出一张确定了起始日期的空白日历表，如示例图中受访者的回忆时段为 2012 年 10 月至 2014 年 6 月。通过向受访者提问，EHC 自动将采集到的信息转化为可视化的日历界面。如图中显示，受访者于 2012 年 10 月至 2013 年 6 月、2014 年 1 月至 2014 年 6 月两个时间段在地址 A 居住；

[34] 我们对追访者回忆时段的设置是从上次调查月至此次调查月，对初访者回忆时段的设置是从上次调查年的 1 月 1 日至此次调查月。

在受访者与配偶 A 的婚姻结束后，受访者结束了工作 B，并同时搬至了地址 B 居住。借助于一些线索——如结婚、生子、搬家、变换工作等重要事件，人们可以回忆出与之相关的越来越多、越来越准确的信息。根据这一记忆原理，EHC 一方面可以清晰明了地向受访者展现出事件的时间表，帮助受访者以其中一些事件作为时间线索回忆出另一些事件的时间；另一方面，EHC 在帮助受访者对事件时间构建记忆的同时，也帮助受访者更完整、准确地回忆起各类事件的相关信息。详细 EHC 设计细节可参见技术报告《

》。

	2012 年			2013 年												2014 年					
	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
	居住地																				
地址 A	■	■	■	■	■	■	■	■	■							■	■	■	■	■	■
地址 B										■	■	■	■	■	■						
	婚姻																				
配偶 A			■	■	■	■	■	■	■												
配偶 B															■	■	■	■	■	■	■
	工作																				
主要工作 A						■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
一般工作 B			■	■	■	■	■	■	■												

图 15. CFPS 事件日历记录法示意图

3.7.3.3. 认知测试

CFPS 共设计了四类认知测试，分别是识字题、数学题、记忆题和数列题，以满足多样化的研究需求，并提高对认知能力测量的全面性与准确性。为减少访问时长，CFPS 将这四类认知测试分为两组，按调查轮次更替使用。其中，A 组包含识字题和数学题，B 组包含记忆题和数列题。具体来说，2010 和 2014 年调查使用了 A 组测试题；2012 年和 2016 年调查则使用了 B 组测试题。

每类测试都包含多套等效的题组，计算机会通过加载数据判别用户以往的测试经历，依此确定在当轮调查中使用的题组。对于在往期调查中接受过该类测试的受访者，系统会按顺序自动加载该受访者上一次接受所使用题组的下一题组；对于从未进行过该类测试的用户，系统则会随机选择题组。同时，在一个家庭内部的同一期调查中，系统会尽可能地为家庭成员加载不相同的题组，以降低测试环境的干扰。关于认知测试的数据清理以及变量信息，可参考 7.3 节。

3.7.4 心理量表

CFPS 个人问卷不仅采集了丰富的社会人口、行为、认知信息，还使用了心理量表对受访者的心理因素进行测量。近年来，越来越多的研究者开始关注包括非认知能力在内的一系列心理因素对个体的影响。CFPS 心理量表所采集到的相对丰富的、具有全国代表性的关于中国城乡居民个人特质的心理数据，能够为推进相关领域的研究提供有价值的数据支持。CFPS 心理量表的测量内容主要包括受访者的个人特质、亲子关系和主观态度 3 大方面。为了保证测量的可靠性与数据的可比性，我们尽可能地引用国际或国内已有的成熟量表。同时，结合中国社会的具体情况，我们对少量量表的内容进行了调整，并自行开发了一部分量表。CFPS 在 2010、2012 与 2014 年的调查中共使用了近 20 个心理量表，分别是良好行为量表（Positive Behavioral Scale，PBS）、自控量表（Self-Discipline Scale）、控制点量表（Nowicki-Strickland Locus of Control Scale for Children）、罗森伯格自尊量表（Rosenberg Self-Esteem Scale，RSES）、凯斯勒心理疾患量表（Kessler 6 Rating Scale，K6）、流调中心抑郁量表（Center for Epidemiologic Studies Depression，CES-D）、责任感量表、父母教养方式量表（Parental Bonding Instrument，PBI）、家庭环境观察量表（The Home Observation for Measurement of the Environment Inventory，HOME）、子女价值量表（The Value of Children to Parents）、养育观念量表、与父母关系量表、成就影响因素量表、信任度量表、不平等程度量表、家庭观念量表、工作满意度量表、重要程度量表和婚姻满意度量表。关于 CFPS 2010 的 K6、CFPS 2012 的 CES-D 和 RSES 量表的详细信息可参考技术报告《2010 年综合变量（2）：受教育水平&抑郁量表（CFPS-12）》和《中国家庭追踪调查 2012 年心理健康量表（CFPS-26）》。此外，《中国民生报告 2016》的第十四章《心理量表的设计与测量》对 CFPS 的心理量表也有详细介绍。CFPS 心理量表跨轮收集的具体情况见表 11。

表 11. CFPS 2010-CFPS 2014 心理量表概况

量表名称	量表说明	问卷位置	受访人群		
			2010	2012	2014
个人特质					
良好行为量表（PBS） WE301-WE312	测量受访者的良好行为，5级评分。	家长代答问卷	3、7、11、15 岁	3、7、11、15 岁	历年数据缺失或新进的 3-15 岁
自控量表 WM701-WM712	测量受访者的自控程度，5级评分。	少儿自答问卷	×	10-15 岁	历年数据缺失或新进入的 10-15 岁
控制点量表（NLCS-C） QM4011-QM40111	测量受访者的内控与外控倾向，5 级评分。	少儿自答问卷 成人问卷	13、15 岁	×	历年数据缺失或新进入的 10-21 岁
自尊量表（RSES） QM1011-QM10113	测量受访者的自尊程度，5级评分。	少儿自答问卷 成人问卷	10 岁	10、12、14 岁	历年数据缺失或新进入的 10-21 岁
凯斯勒心理疾患量表（K6） QQ601-QQ606	测量受访者的心理健康状况，5 级评分。	少儿自答问卷 成人问卷	10 岁及以上	×	10 岁及以上
流调中心抑郁量表（CES-D） QQ6011-QQ60120	测量受访者的心理健康状况	少儿自答问卷和 成人问卷	×	10 岁及以上	×
责任感量表 WF801-WF807	测量受访者的责任感意识，5级评分。	家长代答问卷 少儿自答问卷 成人问卷	代答：6—15 岁 自答：10 岁及以上上学的受访者	代答：上学，或未上学且大于等于 4 岁 自答：10 岁及以上上学的受访者	代答：上学，或未上学且大于等于 4 岁 自答：10 岁及以上上学的受访者
亲子关系					
父母教养方式量表（PBI） WM201-WM214	测量受访者感知的父母教养方式，5 级评分。	少儿自答问卷	11 岁	11、13 和 15 岁	历年数据缺失或新进入的 10-15 岁

家庭环境观察量表（HOME） WG301-WG306 WG308	测量受访者从家庭环境中得到的刺激和支持的强度，5级评分。	家长代答问卷	1-5 岁	1-5 岁	1-5 岁
子女价值量表 WE201-WE209	测量受访者的生育动机，5级评分。	家长代答问卷	2、6、10、14 岁	2、6、10、14 岁	历年数据缺失的，或新进入的 0-15 岁
养育观念量表 WE101-WE108	测量受访者的养育观念，5级评分。	家长代答问卷	1、5、9、13 岁	1、5、9、13 岁	历年数据缺失或新进入的 0-15 岁
与父母关系量表 QM1001-QM1006	测量受访者对于和父母关系的主观评价，5 级评分。	成人问卷	×	×	16 岁及以上
主观态度					
成就影响因素量表 QM3011-QM3017	测量受访者主观评定的各项成就影响因素的重要性，11级评分。	家长代答问卷 少儿自答问卷 成人问卷	代答：4、8、12 岁 自答：12、14 岁	代答 0、4、8、12 岁； 自答：10-15 岁	自答：历年数据缺失，或新进入的 21 岁及以下
信任度量表 QN10021-QN10026	测量受访者的信任度，11 级评分。	少儿自答问卷 成人问卷	×	11、13、15 岁； 16 岁及以上；	历年数据缺失或新进入的 10-15 岁； 16 岁及以上
不平等量表 WV101-WV108	测量受访者对当今社会不平等程度的主观认知，5 级评分。	少儿自答问卷 成人问卷	16 岁及以上	10、12、14 岁	历年数据缺失或新进入的 13-15 岁
家庭观念量表 QM1101-QM1104	测量受访者对两性在家庭中扮演角色的主观认知，5 级评分。	成人问卷	×	×	16 岁及以上
工作满意度量表 QG501-QG506	测量受访者对工作满意度的主观认知，5 级评分。	成人问卷	16 岁及以上	×	×

重要程度量表 QM501-QM510	测量受访者对金钱、人际关系、家庭生活等维度的重要程度的主观认知，5 级评分	成人问卷	16 岁及以上	×	×
婚姻满意度量表 QM801-QM803	测量受访者对婚姻、同居生活的满意度，5 级评分	成人问卷	×	×	16 岁及以上

4. 执行

4.1 预调查

预调查是基线调查之前的一个重要过程。CFPS 在 2010 年全国基线调查正式开始之前，于 2008 年和 2009 年开展了两次预调查。2008 年 5 月至 9 月，CFPS 在北京市、上海市和广东省三地开展了初访的预调查。此次预调查采用了纸笔调查方式，调查主题主要为社会、经济、教育、健康等方面。2008 年预调查的设计规模为 2400 户，每个省/市 800 户，分布于 8 个区/县，每个区/县 4 个村/居，每个村/居 25 户，抽样方式同样为 PPS 抽样，最终完成了 24 个区/县 95 个村/居 2375 户家庭 7214 位个人的访问。[35]

2009 年 5 月至 9 月，我们对三地的初访调查的样本户进行了追访的预调查。这是一次对初访样本的追踪调查的测试调查，但是有两种情况没有被纳入其中：一是离开了村居的家户，二是离开了家庭的人口。在正式的调查中，这两种情况已经被纳入追踪范围。2009 年的样本户是 1995 户。与 2008 年不同的是，2009 年 CFPS 正式开始采用计算机辅助面访，这也是 2009 年预调查测试的一个重要方面。在调查过程中我们对计算机辅助面访调查技术、调查进程实时管理技术、调查进程实时技术支持系统、数据质量实时监控技术的稳定性和可靠性进行了全面测试。

两次预调查为 2010 年的基线调查打下了良好基础。同时，这两次预调查也形成了可供研究者使用的关于三省的研究数据。

4.2 2010 年基线调查访员状况

为了能够更好地控制执行成本与执行时间，2010 年基线调查主要使用样本区/县当地的访员。访员招聘按照每个访员负责调查 2 个村/居的比例，在大城市则按照 1.2-1.5 倍的比例扩招。

访员招聘主要采用了网络招聘的形式，通过简历筛选、电话面试、实地面试等环节，从 2009 年 10 月开始至 2010 年 7 月，我们共招聘到 453 名访员。访员分 14 批分别在北京大学开展为期 6 天的培训。培训从 2010 年 2 月 22 日开始，到 8 月 13 日结束。培训主要采取小班教学的方式，培训课程包括课堂讲解、分组练习、课堂模拟测试和实地入户训练等。最后

[35] 北京大学中国社会科学调查中心（2009）。

有 438 名访员通过培训和考试，成为正式访员。

2010 年的访员的基本状况可以参见下表 12。关于 2010 年基线调查访员招聘与培训的具体信息，可参考《中国家庭追踪调查 2010 年基线调查执行报告（CFPS-3）》。

表 12. 2010 年基线调查访员的基本特征[36] （访员总数：438）

特征	分类	频率	百分比（%）
性别	男	294	67.1
	女	144	32.9
婚姻状况	未婚	265	60.5
	已婚	173	39.5
年龄（岁）	18-19	10	2.3
	20-29	306	69.9
	30-39	101	23.1
	40 及以上	21	4.8
受教育程度	研究生	11	2.5
	本科	198	45.2
	专科	153	34.9
	高中及以下	76	17.4
职业	企业职员	137	31.3
	学生	109	24.9
	事业单位	50	11.4
	无业/待业	44	10.0
	计生系统	39	8.9
	教师	29	6.8
	自主经营	30	6.6

4.3 2010 年基线调查执行概况

2010 年全国基线调查共涉及 25 个省/市/自治区、162 个区/县[37]、649 个村居[38]。执行时间分为两大部分，一是调查季的大规模调查，二是后期针对调查季的调查结果所做的一些补访调查。

调查季从 2010 年 4 月开始，至 2010 年 9 月结束，共完成 600 个村/居的访问工作[39]，完成住户过滤问卷 14852 份，家庭成员问卷 14326 份，家庭问卷 14192 份，成人 32202 份，

[36] 技术报告：CFPS-3。
[37] 上海的 32 个街道/乡镇共涉及 18 个区/县。
[38] 合并村居在这里独立计算。
[39] 执行过程中的“完成访问工作”指经与访员核实所有未完成样本都经过六次联系、三次拒访后，确定访员可以离开本村居结束访问工作。

少儿 8789 份。[40] 在完成访问的 600 个村居中，有 4224 个样本未完成家庭成员问卷。[41] 未能完访的主要原因如下：[42]

⑴ 抽样框精度流失：1690

⑵ 拒访流失：1490

⑶ 六次联系不上流失：461

⑷ 不符合条件被过滤：374

后期的补访工作则主要针对以下几类情况：一是未访问村居或未达成 25 户预定目标的村居，共涉及 324 个村居，调用访员 118 名。其中，未访问村居既包括调查季拒访村居，也包括一些在调查季当时不适合进行访问的村居，如考虑到调查季上海正处在世博会期间，为确保社会安全，我们将上海地区的部分样本的调查推迟。二是经核查作弊的样本，我们需要重新抽样或者重新访问。2010 年我们总共发现 5 例访员重大作弊事件，具体情况及解决方案可参考《中国家庭追踪调查 2010 年基线调查执行报告（CFPS-3）》。为保证数据的一致性和有效性，补访同样采用了计算机辅助入户面访的方式，问卷内容及访问系统完全与调查季调查相同。

经过补访工作，我们 2010 年共完成村居问卷 635 份，住户过滤问卷 15717 份，家庭成员问卷 14960 份，家庭问卷 14798 份，成人问卷 33600 份，少儿问卷 8990 份。2010 年访问样本经 2011 年的维护调查和 2012 年的追访调查确认，已没有错访地址、错访受访人、替换访员和伪造数据的情况。

4.4 2010 年基线调查拒访与拒访逆转

拒访是导致样本流失的一个重要因素，CFPS 2010 年基线调查共产生 1000 多户拒访家庭。[43]经分析发现，居民对访问的接受程度的高低与社区的性质有很大关系，一般来说，在以区县部门领导或社区机关人员为主的社区、高档社区、老人占多数的社区，以及军人社区，受访者对调查的防范心理和抵触心理会比较强，拒访率相对较高。此外，由于我们的调查经常会借用村/居委会与村/居民进行联系，因此，村/居委会的配合力度以及村/居委会与社区的关系也直接影响到受访者对我们访问的配合程度。村/居委会的配合力度小，或对社区管理

[40] 技术报告：CFPS-3。
[41] 技术报告：CFPS-3。
[42] 技术报告：CFPS-3。
[43] 此数据为过程数据。

力度小甚至根本没有实质管辖权的社区，拒访率通常也相对较高。

对于拒访和因其他原因未完成访问的样本，首先，按照执行流程的要求，我们会要求原访员多次登门，并充分发动村居协调人和已接受访问受访者进行协助劝说，以扭转受访者对访问的态度。访员遭到受访者严重拒访必须超过 3 次方可挂起样本，并且需跟督导沟通拒访具体原因及填写未完访情况说明表。其次，我们在执行过程中采取了给拒访户邮递劝说信件、中心简报和《中国报告·民生》[44]的方式增强其对中心项目的信任感；最后，我们还会通过安排协调能力强的访员攻关、督导组队攻关、请计生委单独协调等方式进行多次攻关尝试。通过以上手段与措施，我们在拒访逆转方面取得了一定的效果。

4.5 2010 年基线调查最终联系结果

表 13、表 14、表 15 反映了 2010 年基线调查抽样单元（即家户）的分布情况，表 16 和表 17 反映了个人样本的分布情况，表 18、表 19 分别是对家庭和个人层面各类执行率的计算结果。具体的统计口径与计算方法可参考《中国家庭追踪调查 2010 年基线调查样本联系情况（CFPS-5）》。

表 13. CFPS 基线调查抽样单元最终状态分布[45]

抽样单元最终状态	居委会		村委会		整体	
	数量	百分比	数量	百分比	数量	百分比
符合访问条件						
完访(I)	5081	63.09	9879	82.79	14960	74.85
拒访(R)	348	4.32	143	1.20	491	2.46
其他原因未完成访问(O)	14	0.17	25	0.21	39	0.20
未联系(NC)	17	0.21	36	0.30	53	0.27
不符合访问条件(NE)	652	8.10	812	6.80	1464	7.33
不确定是否符合访问条件(UE)	1941	24.10	1038	8.70	2979	14.91
合计	8053	100	11933	100	19986	100

[44] 《中国报告·民生》（后改名为《中国民生发展报告》）是基于 CFPS 最新数据所写的涵盖中国社会众多热点议题的描述性报告。

[45] 技术报告：CFPS-5。

表 14. CFPS 基线调查抽样单元不符合访问条件类型分布[46]

不符合访问条件类型	居委会		村委会		整体	
	数量	百分比	数量	百分比	数量	百分比
错误地址	62	9.51	23	2.83	85	5.81
非住宅	97	14.88	37	4.56	134	9.15
空置房屋	412	63.19	628	77.34	1040	71.04
住户过滤不合格	81	12.42	124	15.27	205	14.00
合计	652	100	812	100	1,464	100

表 15. CFPS 基线调查抽样单元不确定是否符合访问条件类型分布[47]

不确定是否符合访问条件类型	居委会		村委会		整体	
	数量	百分比	数量	百分比	数量	百分比
未接触地址	7	0.36	5	0.48	12	0.40
正确地址无法联系到住户	662	34.11	484	46.63	1146	38.47
住户拒访	1194	61.51	430	41.43	1624	54.51
其他原因无法进行住户过滤	78	4.02	119	11.46	197	6.61
合计	1941	100	1038	100	2979	100

表 16. CFPS 基线调查个人样本最终状态分布[48]

样本状态	居委会		村委会		整体	
	数量	百分比	数量	百分比	数量	百分比
未联系	508	3.06	1163	2.87	1671	2.92
完访	12793	77.03	29797	73.49	42590	74.52
拒访	1752	10.55	2533	6.25	4285	7.50
其他原因未完成访问	539	3.25	1535	3.79	2074	3.63
不符合访问条件	1015	6.11	5520	13.61	6535	11.43
合计	16607	100	40548	100	57155	100

[46] 技术报告：CFPS-5。
[47] 技术报告：CFPS-5。
[48] 技术报告：CFPS-5。

表 17. CFPS 基线调查不符合访问条件个人样本分布[49]

不在家的原因	当前居住地属于CFPS 调查区县		当前居住地不属于CFPS 调查区县		整体	
	数量	百分比	数量	百分比	数量	百分比
外出读书	126	1.80	972	13.89	1098	15.69
外出工作	294	4.20	5135	73.36	5429	77.56
出家	17	0.24	59	0.84	76	1.09
探亲访友	28	0.40	246	3.51	274	3.91
服刑					19	0.27
参军/服役					75	1.07
出境（包含港、澳、台）					29	0.41
合计	465	6.64	5931	84.73	7000	100

表 18. CFPS 基线调查抽样单元各类执行率计算结果（%）[50]

类型	计算公式[51]	居委会	村委会	整体
应答率	RR3=I/(I+R+NC+O+eUE)	69.35	89.16	81.25
累积应答率	$RR_{累积}= RR_{住户过滤}* RR_{家庭成员}$	69.35	89.16	81.25
合作率	COOP1=I/(I+R+O)	93.35	98.33	96.58
联系率	CON2=(I+R+O)/(I+R+NC+O+eUE)	74.29	90.68	84.13
拒绝率	REF2=R/(I+R+NC+O+eUE)	4.75	1.29	2.67

表 19. CFPS 基线调查个人层面各类执行率计算结果（%）[52]

执行率类型	社区类型		年龄段		整体
	居委会	村委会	成人	少儿	
RR5=I/(I+R+NC+O)	82.05	85.07	82.52	90.76	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 基线调查样本维护

为了防止样本流失，保证跟踪调查的长期有效进行，CFPS 制定了详细的样本维护策略。

[49] 技术报告：CFPS-5。
[50] 技术报告：CFPS-5。
[51] e 表示已完成访问资格筛选的样本中符合访问条件样本所占的比例。
[52] 技术报告：CFPS-5。

2010年基线调查的样本维护实践分多个步骤进行。首次维护从调查季开始后的9月份开始，对象是所有的样本家户和样本村居，包含接受访问的和没有接受访问的。这次维护主要以邮寄材料的形式进行，邮寄的材料分别有致受访户的感谢信、致拒访户的信和致村居委会的感谢信。第二次维护在春节期间进行，对象为2010年接受调查的14767个样本家户。由于春节期间，调查中心对部分样本家户进行了CFPS补访调查和居民对医改满意度调查，对这部分家户借助这两个调查项目进行了实地“搭车”维护。[53] 剩余的97个村居是独立进行维护的村居，主要采用以电话维护为主、实地维护为辅的维护方式，即对有电话的样本家户采用电话维护，对没有电话以及电话维护没有成功的家户进行实地维护。在此基础上，我们通过邮寄的方式向获得有效地址的家户发送了春节贺卡和年度简报。关于2010年样本维护的具体执行方案与过程以及维护结果的统计分析，可参见《中国家庭追踪调查2010年基线调查样本维护（CFPS-18）》。

4.7 追踪调查策略

基线调查结束后，CFPS每两年对基线调查所界定出来的基因成员及其所在家庭进行追踪，目前为止已经实施了2012、2014、2016三轮追踪。此外，CFPS还在2011年实施了以样本维护为主要目标的维护调查。CFPS的追踪调查以基线调查所界定出来的基因成员为源头，根据特定追踪规则动态确定每轮的受访家庭及个人，力求保持样本的截面代表性。CFPS遵循以下追踪策略：

（1）基因成员永久追踪；

（2）核心成员与基因成员关系存续时访问，关系断裂时停止访问；

（3）除死亡自然退出外，所有样本无论前期追踪状态如何，后期追踪时均作为发放样本尝试追踪。

CFPS 2010基线调查采用计算机辅助面访（CAPI）调查模式。结合CFPS项目存在大量离家样本的特点，我们在2012年追踪调查时引入计算机辅助电访（CATI）模式，主要适用于难以实现面访的外出人员、搬迁家庭或另组家庭的追踪访问。考虑到电访的执行难度，我们从个人面访问卷中提取核心内容生成了个人电访问卷。为了尽量降低因无法追踪到离家成员带来的数据损失，从2012年追踪调查起，CFPS引入了代答-自答相结合的数据收集模

[53] 后有59个村居的1226个家户未能“搭车”维护，与97个独立维护村居一起采用了电话维护为主、实地维护为辅的维护方式。

式。在访问离家人员所在的原家庭时，邀请一位对离家人员情况最了解的成员回答一份采集其基本情况的代答问卷。在此基础上，执行团队尝试对离家人员进行追踪。若追踪成功，则邀请其本人回答个人自答问卷：若受访者选择电访访问模式，则回答电访自答问卷；若受访者选择面访访问模式，则回答面访自答问卷。若追踪未成功，我们依然有家庭成员先前的代答信息可供参考。由此，自 2012 年追踪调查起，CFPS 个人层面存在三类问卷：面访自答问卷，电访自答问卷，及代答问卷。

4.8 追踪调查执行情况

为了保证追踪调查的顺利进行，在每一轮全样本追踪调查前 CFPS 都会安排预调查或测试调查。2012 年和 2014 年预调查分别选择了外出人口较多的广东省一个区县、甘肃省一个区县和北京市八个区县进行了小规模调查。预调查的目的一方面是测试问卷系统，另一方面是评估追踪难度和根据实际困难优化执行流程。2016 年 CFPS 项目首次尝试将电访问卷与面访问卷统一，为了测试问卷的适用性，我们进行了一次方便样本的电访测试调查和一次真实样本的预调查。通过方便样本的电访测试，我们了解了电访问卷的调查时长，并结合访员和受访者的反馈进一步完善了电访问卷。真实样本的预调查从当年的发放样本库中选取了 500 多条区县层面聚集程度低的样本户作为测试样本，进一步测试电访系统，同时获取应答率等信息。

追踪调查的访员规模基本维持在 400 到 500 名，但随着混合调查模式的深入，电访访员的规模在逐轮增加。到 2016 年，电访访员已增加到 70 名左右。访员培训分批次进行，大部分培训批次在北京大学完成，少量批次在合作院校或其他合作机构完成。2010 年基线调查主要以社会访员为主，学生访员只占四分之一。在后续的追踪调查中，学生访员比例逐渐增加，2012 年增加到整个访员队伍的一半，而到 2016 年，学生访员的比例已经超过四分之三。从地域上来说，追踪调查的面访访员基本维持以本地访员为主的特征，而电访访员基本是在京各高校的学生和合作机构的职业电话访员。

由于追踪调查的样本较之基线调查更为分散，追踪调查的执行总体上分多个阶段进行。在 2012 年，整个执行工作共分为三个阶段：一是原地址回访与本地追踪访问，于 2012 年 7 月 20 日开始，至 2012 年 11 月 30 日结束；二是异地追踪以及电话访问，于 2012 年 9 月 19 日开始，至 2013 年 1 月 18 日结束；三是春节补访，于 2013 年 2 月 1 日开始，至 2013 年 3 月 4 日结束。异地追踪工作是该轮调查的重点，追踪对象主要有四类：外出家庭成员、另组

家庭、全家外出家庭、搬迁/拆迁家庭。对这四类追踪对象的追踪策略分为三步：一是就近调配访问，将需要异地追踪的样本即时调配当地的访员开展就近访问；二是追访小组访问，对于外出到CFPS样本地区范围外无法就近调配的样本，由专门的追访小组跨地区进行异地追踪，除专职访员外，中心的部分员工也参加了此次追访；三是电话访问，对于愿意接受电话访问的家户及个人，由中心组织电话访问。2012 年的执行还包括一部分纸版问卷的访问，目的是完善初访数据中所发现的家庭关系存疑的家户信息。执行管理队伍包括一名执行主管，4 名负责前期准备阶段、访员培训、调查执行和电访阶段的督导，以及 10 名分省督导。平均每个督导在实地执行期间负责 16 个左右区县的约 40 名访员的相关工作。

2014 年追踪调查工作也分为三个阶段：一是原地址回访与实地追访，于 2014 年 7 月 4 日开始，至 2015 年 6 月 7 日结束；二是电话访问工作阶段，于 2014 年 8 月 8 日持续至 2015 年 5 月 18 日结束；三是春节补访工作，于 2015 年 2 月 7 日开始，至 3 月 22 日结束。鉴于 2014 年全国追访工作范围广、追踪样本量大的特点，CFPS 设计开发了追访系统，其主要功能包括电话约访、电话访问及 2012 年另组家庭样本的样本发放与进度监控。本轮追访涉及 10 个样本类型，除完访、死亡、错误地址以外的家户样本全部进入再次追踪的范围。在工作方式上，我们不再使用以往的手工表格整理追访信息的方式，而是通过追访系统进行追访样本的调配、发放与管理工作。该系统的主要功能包括以下几个方面：一是展示往年调查中获取到的全部电话信息，在提高了受访者联系信息利用率的同时，也将所有电话信息进行了一次筛查，将所有无效的电话标示出来；二是将电话约访工作和电访工作系统化；三是展示样本的访问进展和结果。经实战验证，追访系统有效加快了样本的循环速度，提高了样本传递的准确性，从而使访问的成功率有了显著的提升。在追访规则上，对于原地址回访、本地追踪、异地追踪过程中不愿意接受面访，由访员选择相关代码将其调回中心，然后由电访访员先进行约访工作，根据受访者的意愿以及异地追踪访员的分布情况，决定是采用电访还是面访（市内聚集 3 个及以上样本可进行实地面访）。另外，对于外出至 25 个基线样本省外的样本，系统也将其自动转为电话访问。

2016 年追踪调查工作共分为两个阶段：一是原地址回访与实地追访，于 2016 年 6 月 28 日开始，至 2017 年 4 月 30 日结束；二是电话访问工作阶段，于 2016 年 5 月 13 日开始，至 2017 年 4 月 30 日结束，参与电访工作的部分人员是原面访访员，这部分访员可以在异地进行电话访问。由于电访问卷的增长，使得此轮调查的电访任务较往年更重，因此我们在执行中做了如下三项调整：一是简化电访转化流程以提高电访效率和成功率，在与受访者的约访

电话中，如果受访者选择电访则立即生成电访问卷，可以直接进行电访；二是充分调动已参加过培训的面访访员，经过技术部门的配置，这部分访员可以在异地通过网络电话进行电话访问；三是从专业电访公司雇佣了部分电访访员，提高电访样本的完成速度。

除了两年一次的全样本追踪之外，CFPS 在 2011 年还进行了一次以样本维护为主要目标的维护调查。维护调查执行分为常规访问和追访两个阶段，前期的常规访问采用面访，后期的针对外出青少年的追访首次尝试了电话访问、网络问卷和邮寄纸笔问卷相结合的混合访问模式。面访调查的执行从 2011 年 7 月 21 日开始，至 2011 年 11 月 20 日结束，混合模式调查的执行从 2011 年 12 月 29 日开始，至 2012 年 1 月 9 日。此外，我们还按照早期确定的样本维护策略，每年定期对样本开展维护工作。

4.9 家庭层面追踪结果

表 20 及表 21 分别展示了 CFPS 2012 和 CFPS 2014 家庭层面的访问结果。两轮追踪调查的截面完访率分别为 79.4%和 77.9%。根据前一轮次被访家庭的完访状态，我们将样本分为完访和未完访两类。2012 年追踪时，基线完访家庭的跨轮追踪率[54]为 85.3%，基因成员另组家庭的截面应答率为 35.9%。2014 年追踪时，2012 年完访家庭的跨轮追踪率为 89.7%，2012 年未完访家庭的截面应答率为 44.6%。

表 20. CFPS 2012 家庭层面访问结果

	所有家庭		2010 完访		2010 未完访	
	家庭数（户）	百分比（%）	家庭数（户）	百分比（%）	家庭数（户）	百分比（%）
发放家庭数	14960		14960		0	
另组家庭数	2031		0		2031	
全家去世	37		35		2	
家庭总户数	16954	100	14925	100	2029	100
未联系上	1847	10.9	845	5.7	1002	49.4
拒访	887	5.2	744	5.0	143	7.0
搬迁无法获得联系方式	469	2.8	410	2.7	59	2.9
受访者原因无法完成访问	298	1.8	202	1.4	96	4.7
完访	13453	79.4	12724	85.3	729	35.9

[54] 扣除全家去世样本后，前一轮次完访样本在本轮追踪中的完访率。

表 21. CFPS 2014 家庭层面访问结果

	所有家庭		2012 完访		2012 未完访	
	家庭数（户）	百分比（%）	家庭数（户）	百分比（%）	家庭数（户）	百分比（%）
发放家庭数	14925		12724		2201	
另组家庭	3286		729		2557	
全家去世	60		36		24	
家庭总户数	18151		13417		4734	
未联系上	1938	10.7	624	4.7	1314	27.8
拒访	1215	6.7	426	3.2	789	16.7
搬迁无法获得联系方式	589	3.2	209	1.6	380	8.0
受访者原因无法完成访问	265	1.5	124	0.9	141	3.0
完访	14144	77.9	12034	89.7	2110	44.6

家庭层面追踪情况的更详细情况可参考《中国民生发展报告 2016》第十二章《CFPS 样本流失情况分析》。

4.10 个人层面追踪结果

表 22、表 23 展示了个人样本追踪情况，2012 和 2014 两轮追踪调查的截面应答率[55]分别为 74.1%和 72.8%，跨轮追踪率分别为 80.6%和 83.8%。六个抽样框中，上海及广东的跨轮追踪率及未完访样本挽回率[56]均处在较低水平，样本损耗较大，详见表 23。纵向比较可以发现，2014 年追踪的跨轮追踪率有所提升，但样本挽回难度也随之增大。个人样本完访问卷类型如表 24 所示，随着样本分散程度的增加，电访模式占比逐渐提升。

CFPS 每轮调查基因成员的数量及完访情况如表 25 所示。经过两轮追踪，基因成员总量小幅增加 2004 人。将本人完成自答问卷界定为严格个人完访，将完成个人问卷界定为宽松个人完访，基因成员三轮调查问卷完访情况参见表 26。宽松标准下，8.9%的基因成员从未完成个人问卷；45.6%的基因成员持续三轮追踪成功。严格标准下，87.9%的基因成员本人至少接受过一次访问。

[55] 计算应答率时在分母中扣除了死亡及无需追踪的样本。
[56] 扣除不符合访问条件样本后，前一轮次未完访样本在本轮追踪中的完访率。

表 22. 个人样本追踪情况（基于上一轮访问结果）

	CFPS 2010				新增样本	整体
CFPS 2012	完访	未完访	无需追踪	县外放弃		
完访	33956	3637	43	3974	2729	44339
未完访	8185	4170	38	2377	748	15518
死亡	42	18	41	35	4	140
无需追踪	407	205	1	26	0	639
	CFPS 2012				新增样本	整体
CFPS 2014	完访	未完访	无需追踪			
完访	36856	6075	52		2722	45705
未完访	7103	9122	43		774	17042
死亡	97	75	39		1	212
无需追踪	283	246	2		0	531

表 23. 个人样本分抽样框追踪率（基于上一轮访问结果）

	前期完访		前期未完访		新增样本		整体	
	样本数	应答率	样本数	应答率	样本数	应答率	样本数	应答率
				CFPS 2012				
上海	3522	65.7%	807	31.2%	185	72.4%	4514	59.9%
辽宁	3658	80.8%	994	43.2%	279	77.0%	4931	73.1%
河南	5005	86.9%	1486	67.9%	475	85.1%	6966	82.8%
甘肃	4917	87.1%	1957	66.8%	419	76.6%	7293	81.1%
广东	4185	76.3%	2238	51.6%	336	71.1%	6759	67.9%
其他	21303	80.8%	7083	51.9%	1787	79.4%	30173	74.0%
全国	42590	80.6%	14565	53.8%	3481	78.5%	60636	74.1%
				CFPS 2014				
上海	2677	75.5%	1799	20.4%	195	62.1%	4671	53.9%
辽宁	3552	87.2%	1319	35.6%	200	75.5%	5071	73.5%
河南	5690	89.8%	1195	48.4%	569	86.3%	7454	83.0%
甘肃	5833	86.2%	1380	53.4%	416	83.7%	7629	80.3%
广东	4537	74.5%	2160	29.5%	379	66.0%	7076	60.4%
其他	22050	84.1%	7801	44.8%	1738	78.4%	31589	74.1%
全国	44339	83.8%	15654	40.1%	3497	77.9%	63490	72.8%

表 24. 个人样本完访问卷类型

	完访问卷	整体	自答		代答	
			面访长卷	电访短卷	面访短卷	电访短卷
CFPS	计数	44339	40504	1027	2806	2

2012	百分比	100.0%	91.4%	2.3%	6.3%	0.0%
CFPS	计数	45705	39450	2238	2829	1188
2014	百分比	100.0%	86.3%	4.9%	6.2%	2.6%

表 25. 基因成员追踪数量及完访情况

	完访	未完访	无需追踪	死亡	县外放弃	总量	应答率
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%

表 26. 基因成员完访类型

完访轮数	宽松标准		严格标准	
	计数	百分比（%）	计数	百分比（%）
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
合计	64133	100.0	64133	100.0

个人层面追踪情况的更详细情况可参考《中国民生发展报告 2016》第十二章《CFPS 样本流失情况分析》。

5. 调查质量控制

5.1 CFPS 质控手段与技术

5.1.1 基线调查

CFPS 2010 年基线调查采用了严格的质量控制手段以保证数据的质量，针对问卷设计不当、末端抽样不准确、访员行为不规范、数据汇总和整理过程出错等一系列可能影响数据质量的因素，通过电话核查、实地核查、录音核查、采访过程回放、数据统计分析等手段进行了监控与干预。

以数据使用者普遍关心的由于访员替换或访错住户、替换或访错个人而影响调查代表性的问题为例。质控团队在调查开始之前、之后分别选择了不同的方法进行质量控制。对于随意替换或访错住户的情况，在调查开始之前，抽样员或村居干部会把给住户的信投递到正确的地址，然后由住户打电话回中心登记必要的信息。质控部门会将这些信息与访员回传的住户信息进行核对；在调查开始之后，也会通过实地核查来核实住户抽样的准确性，实地核查员将登记并反馈地址内符合调查资格的住户数，以及访员访问时地址内各住户是否拒访或者无法联系等情况。对于随意替换或访错个人的情况，主要在调查开始之后，通过实地核查、电话核查、录音核查来进行控制。此外，对于访问错误的住户或者个人，需要找到正确的住户或者个人重新进行访问。而且，在访员培训过程中，我们也重点强调并严格考察了访员在这方面的行为规范。

此外，对由问卷设计不当产生的系统性误差，我们通过每周对问卷数据和并行数据的统计分析，识别系统偏差，并根据需要修正设计上的漏洞，更新访问系统，从而达成质量控制的目的；针对末端抽样不准确，我们主要依靠实地核查来进行质量控制；针对访员引导受访者作答，访员通过跳转模式故意回避需要长时间作答的题组等常见的访员不规范操作引发的问题，我们通过电话核查、录音核查以及采访用时的统计分析、对问卷数据无回答率的分析等方法，都能达到质量控制的目标。

由于使用了计算机辅助访问系统，CFPS 在数据汇总和整理过程出现的系统误差相比常规的纸笔调查大大减少。但是，在对原始开放型问题的编码过程中，系统误差依然不可避免。针对这一问题，我们主要借助系统化的编码表，通过三人共同给同一个题目的原始资料进行

编码的方式进行控制。同时，数据整理过程中出现的一些误差主要依靠多人多次循环检验的方式进行控制。

5.1.2 追踪调查

从 2012 年起，追访调查的数据质量已不再受到末端抽样框精度的影响，因而不再采取 2010 年初访时采用的针对末端抽样框误差的质量督导方法。追踪调查的核查手段上保留了数据核查、录音核查及电话核查。由于组织成本的问题，2012、2014 年并未开展实地核查。此外，软件回放技术因成本较高、收效较低，自 2012 年调查起被数据核查所取代。

2012 年我们对访员不规范行为进行了细化，并制定了客观的核查评价标准。数据核查方面，在单条样本访问时长分析的基础上，增设了单题访问时长核查点。在访问开始前，为问卷中每道问题设置最短访问时长，数据回收后将每道问题实际耗时情况与最短时长进行比较，由此对访问质量作初步评估，并以此为依据安排后续核查手段。根据访员访问行为界定出“臆答”、“捷径跳转”、“录入错误”、“提问不准确”、“追问不足”、“代访”等 12 种不规范访问行为，并为每个不规范行为构建了客观的评价标准。录音核查方式由全样本录音回听修正为目标问题录音评估，提高了录音核查的效率和针对性。

5.2 CFPS 质控策略

从质控的覆盖面来看，我们规定：第一，定期（每 7 天）对每一份问卷中的所有变量和并行数据中的所有变量通过统计分析的方法进行检验；第二，每个访员都要经历所有的核查方式；第三，每个访员访问的每类问卷数据都需要被核查到；第四，每个访员每种类型的联系结果（如，空访、空址、拒访、身体语言障碍等等）也都需要被核查到。

数据核查自 2012 年起覆盖 100%完访样本，核查周期由每 7 天一次缩短为每天进行。对于核查过程中发现不规范行为的样本采取了扩大核查比例的策略，对于无法核查的样本则采用了替换核查的策略，保证核查覆盖面。此外，自 2014 年起，为了提高核查资源的利用效率，实施了效率监控策略，根据累计核查结果实时调节每个访员的核查比例。

从数量上看，我们规定：第一，每个访员未能完成访问的地址，按 60%的比例进行实地核查；每个访员已完成家庭或个人问卷访问的住户，分别按 15%、25%、15%、5%的比例进行录音核查、电话核查、实地核查和采访回放核查。

核查样本抽选规则自 2012 年起作了较大的调整。核查样本由基线调查的随机抽取改为定向及随机相结合。首先，所有数据核查不通过的样本，作为“存疑样本”100%优先进行录音核查；其次，数据核查通过的样本中，随机抽取 10%的样本，分为两组，分别优先进行电话或录音核查；再次，访员完成的每类问卷的前三份优先进行录音核查用于判断访员访问技能掌握情况；最后，当一种核查手段发现不规范行为时，将启动另一种核查手段进行佐证。

从顺序上来说，数据的统计分析始终用于全面数据核查，在各类问卷样本量达到 30 以上时开始，每 7 天汇总一次所有数据，检查系统误差；录音核查优先用于全面核查，在收到回传数据的第 2 天开始进行，延续至调查期结束后的第 3 天；电话核查优先用于没有录音或没有经过录音核查的样本，在收到回传数据的第 2 天开始进行；实地核查重点用于检验末端抽样的精度，在收到回传数据的 20 天之内开始进行实地核查；采访过程回放主要针对前几种方式所发现的问题，例如，采访用时过短、用时过长等，对样本进行回放。整个质量控制过程重视时效性原则。

自 2012 年起，我们确定了数据核查始终优先于其他核查方式，利用 SAS 开发数据核查程序，每天对所有上传的完访样本的并行数据和问卷数据作统计分析，筛选访问质量存疑样本。录音核查优先用于访问前期完成的样本、数据核查不通过的样本及随机选取但无电话的样本。电话核查优先用于随机选取的样本。

5.3 核查比例与质控结果[57]

5.3.1 基线调查

录音核查方面，实际核查的家户数占全部访问成功的家户数的 28%，实际录音核查的问卷总量达全部问卷总量的 16%。

电话核查方面，成功接受电话核查访问的家户数占全部访问成功的家户数的 19%。

实地核查方面，在调查季实地核查的家户数占全部家户数的 25%。此外，在 2010 年 12 月至 2011 年 2 月、2011 年 7 月至 11 月，调查中心又分别对部分样本和全部样本进行了两轮实地核查。

[57] 此部分数据统计结果来自技术报告：CFPS-4。

采访过程回放方面，核查的样本数占全部有效样本数的 3%。[58]

数据的统计分析方面，包含采访用时、无回答率、离群值、态度量表题目内部一致性信度系数等内容的分析报告每周都将向质量督导部门和调查实施部门提交。

从核查结果来看，在所有接受了电话核查、实地核查、录音核查的家户中，没有发生访错受访户的情况。成人问卷有 81 份问卷存在代答情况，涉及 21 名访员，其中有 59 份属于同一访员的作弊行为，有 7 份问卷属于受访者不在家而被家人代答的情况，其余 15 份属于受访者无回答能力或者被抢答的情况。少儿自答问卷有 22 份问卷存在代答情况，涉及 20 名访员，其中有 7 份属于孩子在回答过程中被家长抢答的情况。除此之外，经实地核查我们还发现了其它几例严重作弊行为，在本手册的第 4 章已经谈及，此处不再赘述。

此外，绝大多数访员在采访过程中严格遵守了采访规范，经质量控制发现的访员的不规范访问之处主要表现在由于语速过快或问题漏问而导致采访用时过短这个方面。我们将总采访用时过短，并且采访用时过短的题目数占监测题目总数的比例超过 50%的样本定义为“采访用时过短样本”。这类样本共有 1051 份（包括家庭和个人问卷），共涉及到 50 名访员。在被抽查的问卷中，发现至少有 1 个漏问题目的问卷共有 7914 份。[59] 在调查过程中，我们针对问题漏问这一现象采取了干预手段，取得了比较显著的改善效果。

CFPS 在 2010 年不仅通过严格的质量控制手段有效保证了数据的质量，同时也在质量控制方法与策略方面积累了很好的经验，具体实施方案与过程可参考《中国家庭追踪调查 2010 年基线调查质量督导报告（CFPS-4）》

5.3.2 追踪访问

2012 年数据核查覆盖 52545 条完访样本，核查不通过样本 4896 份，不通过比例为 9.32%。质控小组成功完成 18515 条完访样本录音核查，核查比例为 35%；不通过样本 1030 条，占比 5.56%。我们成功对 3062 个完访家庭成功进行了电话核查，核查比例为 26.54%，不通过样本数为 566 条，占比 18.48%。排除仅因为酬金数额问题被查出质量不合格的 352 条样本后，修正后的电话核查不通过率为，占不通过样本的 6.99%。

[58] 在调查执行的中后期，采访过程回放由于时间成本过高且收效不佳，最后被问卷监测题目采访用时的统计分析替代。

[59] 问题漏问并不能简单地用来判断访问质量。在漏问问题中，除少数问题属于故意漏问外，绝大多数漏问问题都是由于访员根据观察或者已经获得的信息能够直接判断出问题的答案，如，访员看见受访人在打手机，在问到 “您是否有手机”这个问题时，则没有询问受访人而自行填了答案。

2014年数据核查覆盖67482条完访样本，核查不通过样本2531份，不通过比例为3.75%。录音核查覆盖15928条样本，占完访样本总量的20%。其中，成功核查15484条，核查成功率为97%。成功核查样本中，不通过样本579条，占比3.7%。质控小组共抽选了4904个完访家庭进入电话核查，核查比例为35%；其中，成功对3745个家庭成功进行了电话核查，核查成功率为76%。成功核查的家庭中，不通过样本数为207条，占比5.57%。

具体质量督导实施方案及核查结果参见将发布的质量督导报告。

6. 数据库与数据清理

6.1 数据库基本情况介绍

CFPS 2010 年基线数据库包含村居问卷数据库、家庭关系数据库、家庭问卷数据库、成人问卷数据库和少儿问卷数据库五个类型，分别对应了村/居问卷、家庭成员问卷、家庭问卷、成人问卷和少儿问卷的内容。CFPS 后续追踪调查数据库总体结构与基线数据库类似，但由于 2012 年没有对村居开展访问，因此目前只有 CFPS 2010 和 CFPS 2014 数据库包含了村居问卷数据。从 CFPS 2012 开始，我们在基本问卷数据库之外，还创建了跨年个人状态库，用以记录所有曾经进入 CFPS 的个体样本的基本信息及每轮的访问状态。在上文抽样部分我们已经介绍过，CFPS 共有 6 个抽样框，分别代表 6 个子总体。在发布的数据库中，我们用指示变量 subpopulation 对不同的子总体进行标示，subpopulation=1，2、3、4、5、6 分别代表上海、辽宁、河南、甘肃、广东和其他省市。此外，我们增加了一个二分指示变量 subsample，subsample=1 指再抽样得到的全国性样本。

全国完全样本包括所有 CFPS 数据，由代表 6 个子总体的 6 个子样本组成。通过加权，可以代表全国。全国再抽样样本是将 5 个“大省”采用和“小省”一致的比例再抽样后而得到的具有全国代表性的样本。权数的使用参考第 9 章。

表 27 展示了 CFPS 2010 年基线调查各个问卷数据库中的变量数以及各抽样框的样本量。数据用户可以通过此表大体了解 CFPS 数据库的概况，其他轮次调查的情况请具体参考相关数据库，此处不再一一列举。

表 27. CFPS 2010 年数据库变量总数与样本量

	变量总数	样本量						
		全国完全	全国再抽样	上海	河南	甘肃	辽宁	广东
村/居问卷数据库	221	635	417	58	64	65	63	64
家庭关系数据库	355	57155	36964	4329	6491	6874	4652	6423
家庭问卷数据库	624	14798	9661	1405	1506	1537	1478	1394
成人问卷数据库	1493	33600	21812	3162	3732	3704	3129	3070
少儿问卷数据库	968	8990	5944	360	1273	1213	529	1115

注：变量的具体内容参见编码手册。

除以上公开使用的数据库之外，CFPS 项目组还创建了限制性使用[60]的区县数据库，包括 CFPS 基线样本区县的一系列宏观变量（区县 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 到 19 岁人口性别比例、非农业户口人口比率）。该区县库不包含区县国标码，但用户可以通过该库中的区县顺序码与 CFPS 公开使用数据库（如家庭库和个人库）链接。目前区县数据库中的宏观变量主要来自 2010 年国家统计年鉴数据。为了进一步保护 CFPS 受访者隐私，区县数据库中的数值均经过模糊化处理，并非这些区县相关变量的原始数值。有关模糊化处理的过程，请参见 CFPS 技术报告《中国家庭追踪调查区县数据库模糊方法（CFPS-23）》。

6.2 数据清理

6.2.1 家庭关系数据库清理

CFPS 2010 家庭关系数据库的主要内容是 T1、T2、T3 三张表格。上文提到过，2010 年 CFPS 在家庭成员问卷中借助 T1、T2、T3 三张表格收集家庭成员关系及家庭成员的社会人口信息。T 表格收集到的这些信息为我们后期的数据清理提供了大量的依据和有效的帮助。

2010 年整个家庭关系数据库的清理工作分多个阶段进行。第一阶段的基础清理主要致力于更正执行过程中出现的错误，如更正访员录入错误的信息、处理各类废卷（如访员作弊问卷、重复问卷等）、更正样本编码调用错误的户编码等。基础的清理工作主要借助于执行团队反馈回来的实地信息完成。

第二阶段的基础匹配清理则主要对家庭关系库中的个人样本与个人库样本之间的对应关系进行核查和调整。按照 CFPS 的设计规则，家庭关系数据库中个人编码以“1”开头的家庭成员需要回答个人问卷。所以，在此阶段，如果联系结果显示某家庭成员回答了个人问卷，我们需要能够在个人库中找到其对应的个人问卷。同时，我们要进一步确定根据其家庭关系库中的个人编码在个人库中所找到的相对应的个人问卷是其本人的个人问卷，不能存在张冠李戴的错位现象，我们主要是通过姓名、出生年月等关键变量来检测家庭关系库与个人库中同一编码的样本的一致性。对于核查中发现的错误，进行了逐一确认和更正。

[60] 限制性使用数据(restricted-use data)与公开使用数据(public use data)相对应。在 CFPS 项目中，后者的使用只需在线申请，审核通过后，用户可自行从网上下载。而限制性使用数据需要使用者另外提交申请报告，说明研究目的以及限制性数据的具体用途。

第三个阶段主要是对家庭关系数据库的深度清理。在前两个阶段的工作完成后，我们进行了两方面的匹配工作：[61] 一是将家庭关系库中 T1、T3 表中所有家庭成员及其直系亲属按照 T2 表中的关系索引进行了一一匹配，经匹配后发现了少数样本存在逻辑错误或者疑点，如父亲性别为女、母亲性别为男、夫妻之间不互认、父母与子女之间年龄相差太小或太大，等等；二是将家庭关系库与个人库进行匹配，经匹配后也发现了一些逻辑错误或者疑点，如，初婚配偶所填报的孩子总数不一致，家庭关系库中配偶不健在但是个人库中的婚姻状态为已婚或者同居，等等。

对第三阶段的清理，我们首先通过系统版本的整体迁移，解决了少数由于计算机系统程序错误导致的家庭成员编码错误的问题。除此之外的其他错误均由手工清理完成。手工清理同时参考了 2010 年家庭关系库、2010 年个人库、2011 年个人库和 2011 年家庭关系库等多方面的信息，经多方佐证得到可靠信息之后，才进行修正。若无可靠参照信息，对于绝对的逻辑错误，如父亲性别为女、配偶双方性别相同，一般处理成缺失；对于一些可能但不确定出错的疑点，如父母与子女年龄相差太小，则不进行处理。手工清理的详细过程与方法可参考技术报告《中国家庭追踪调查 2010 年家庭关系数据库清理（CFPS-7）》。

2010 年基线调查时 CFPS 通过独特的三个 T 表设计构造出了详细的家庭关系网络，而在后续追踪调查时，家庭关系问卷主要采集家庭结构的变化：如家庭的分裂、成员流动、新进成员与原有成员的家庭关系等。数据清理时我们主要通过以上这些变量的信息，恢复 T1 和 T2 表。由于成年子女经济独立或是夫妻离婚等原因，2010 年基线调查界定的家庭在后续调查中会出现分裂，产生两个或以上家庭。我们将其中一个称为原家庭，延续以前家庭的家户号[62]，其它家庭称为另组家庭，产生新的家户号。

追踪调查年的家庭关系库只包含当年成功访问的家庭，不包含前期成功访问但在调查年流失的家庭。[63]家庭关系库的构建步骤主要有两步：一是以往期家庭关系库为基础，添加调查当年新进的家庭成员及其家庭关系，更新原家庭关系；二是加入另组家庭的家庭关系。与基线家庭关系库的清理工作一样，在基础关系库构建完成后我们会进行一系列针对年龄、性别以及婚姻状态的核查和更新。除此之外，我们还会对相关联家庭间和不同数据库间的样

[61] 具体的匹配方法参见技术报告：CFPS-6。
[62] 原家庭和另组家庭的区分带有一定的随机性。从执行上来说，第一位受访成功的人员所在家庭为原家庭（通常是在原地址上的单元），其它家庭为另组家庭。
[63] CFPS 中还会出现少量全家死亡的家庭，这些家庭的信息不会出现在家庭关系库中，但在跨年个人状态库中会有所体现。

本一致性进行核查。

从最终形成的数据库来说，追踪调查年的家庭关系与基线数据相比具有如下特点：

（1）从原家庭分离出来的另组家庭成员会在原家庭和另组家庭中均有一条记录，形成部分个人样本在家庭关系库中存在多个记录的现象。这样的数据库设计是为了反映出家庭成员在不同家庭间的动态流动过程。用户可以通过 co_aXX_p（XX 代表调查年，如 12，14 等）变量来判断该名成员经济上应该归属的家庭，其中 co_aXX_p = 1 代表该名成员经济上归属这个家庭，而 co_aXX_p = 0 则表示该名成员已从这个家庭中分离出来。当我们把数据只限定于 co_aXX_p = 1 的观测时，只保留了各家庭中“同灶吃饭”的成员，这时每位成员只存在唯一一条记录，即每位成员在一轮调查中只能经济上归属一个家庭。

（2）由于家庭的分裂和重组，追踪调查年的关系库中不仅有个人调查当年所属家庭的家户号，还包含该成员在之前调查年所归属的原家庭的家户号。我们在家户号变量名后加上年份以区别成员在不同年份所归属的家庭（如 fid10, fid12, fid14）。

（3）根据家庭的分裂、人员的流动以及存殁，我们对基因成员进一步细分，在追踪调查年家庭关系库中提供 genetype 变量来指征基因成员的具体类型（如，在家基因成员、新进基因成员、外出基因成员、死亡基因成员等）。追踪调查年家庭关系库的创建和使用的详细信息可以参考技术报告《中国家庭追踪调查 2012 年家庭关系原始库的分解与重构（CFPS-33）》。

6.2.2 其它数据库清理

其它数据库的清理同样分为几个阶段完成。

第一阶段的清理主要是针对个人问卷数据库的逻辑关系清理，这与家庭关系数据库的深度清理是一个同步进行、相互辅助的过程。一方面，在清理家庭关系数据库的过程中，经过家庭关系库和个人库的匹配，我们同时也发现了个人库的一些问题，如本人填写的结婚年龄与其配偶所报的结婚年龄出入太大、家庭关系库中配偶不健在但是个人库中的婚姻状态为已婚或者同居等问题，若经多方验证发现属于个人库有误，则对个人库进行修正。另一方面，通过个人库本身的信息我们也发现了一些逻辑错误或疑点，如离婚时间早于结婚时间，结婚时间早于出生时间，结婚年龄小于 16 岁，等等。对此，我们借助家庭关系库的辅助信息进行调整与修正。处理的原则与家庭关系库一致，即如果有可靠的信息来源则进行处理，如果

没有可靠的信息来源，则将出现逻辑错误的变量处理为缺失，对存在疑点但不确定出错的变量不进行处理。此外，对于若干重要的变量，我们没有对原始的取值进行修正，而是在原始取值的基础上，另生成一组“最佳变量”保存经清理和综合各方面信息后得到的我们认为最为合理的取值。在后面的综合变量部分，我们将对最佳变量进行详细介绍，此处不再赘述。

第二阶段的清理则是对家庭关系数据库之外的其它所有数据库的变量清理。与第一阶段的工作相比，第二阶段的清理工作是一个精益求精的过程，从第一阶段清理结束开始，一直持续到数据发布。数据团队经过对所有变量的逐一核查，删除了数据库中的冗余变量，对数据处理过程中丢失的变量进行了补充，对有误的变量名、变量名标签和变量值标签进行了调整，同时，结合各方面信息修正了一些明显错误的取值[64]，等等。不过，对于一些疑似不合理的极值，我们虽然重点关注并进行了原因分析，但是在绝大多数情况下我们会保留原值，除非有特别可靠的依据进行修改。比如，在后期的数据清理过程中，我们发现财产、收入类数据中存在一些过大或过小的数值。针对这些奇异值，CFPS 数据人员对于有录音的样本采用了录音回放的方式进行了后期核查，对确认有误的样本进行了修正。修正的样本大部分属于访员在记录的过程中忽略了单位而造成数值录入错误的情况。有关财产类数据清理的详细过程，可以参考技术报告《中国家庭追踪调查 2012 年和 2010 年财产数据技术报告（CFPS-29）》。从 CFPS 2016 开始，我们将录音回放提前到与实地调查同期进行，范围也从一开始的财产、收入类数据扩大到大部分连续变量。

此外，我们的数据清理工作还包括对不同类型问卷数据的整合。我们在前文提到过，从 2012 年追踪调查开始，CFPS 在面访版本的基础上增加了电访问卷和代答问卷。这两类问卷的初始数据均为独立数据库，但为了用户能更加方便地使用数据，我们将其与面访及自答问卷数据整合在一起，以降低用户遗漏样本的可能性。在整合不同问卷数据的过程中，我们遵循以下两条基本原则：

（1） 对于面访和电访问卷中的相同变量[65]，合并库采用同一变量名；如果面访和电访问卷的问题不完全相同，合并库同时保留面访和电访版本，并将该样本没有接触到的访问模式所对应的变量设为缺失。绝大部分样本只有面访数据或者电访数据中的一种，少量样本既存在面访又存在电访问卷数据，在这时我们只保留面访数据，并将访问模式(Iwmode)记录为面

[64] 如，某家户在 fe3“您家是否参与经营或完全经营非农产业”一题的取值为“-8”，但在非农经营模块中有有效数据，经过清理，将 fe3 的变量值修改为“1”（是）。

[65] 相同变量是指问题的题干和选项均完全相同。

访。

（2）对于自答和代答问卷中的相同变量，合并库采用同一变量名；如果自答和代答问卷的问题不完全相同，合并库同时保留自答和代答版本，并将该样本没有接触到的问卷版本对应的变量设为缺失。当自答和代答问卷数据同时存在时，只采用自答数值（除非自答数据中该变量为缺失）。

6.2.3 后期实地信息采集与数据更正

后期实地信息采集也是我们进行数据清理的一个重要途径。由于成本较高，我们仅对少部分存在关键信息缺失或错误且无法通过常规途径进行补充修正的家户或个人采用了这一方法。到目前为止，大部分的后期实地信息采集均针对 2010 年基线数据中对于后期追踪调查的开展至关重要的一些变量，以下为具体介绍。

⑴ 回访

对于 2010 年家庭关系库中存在逻辑错误但由于信息不足无法进行清理的问题家户，我们在 2012 年对其进行了回访。回访主要针对三种错误类型：一是由于信息不足无法做出明确更正的问题家户，如父母与子女之间年龄差距不合理（小于 15 或大于 50），原因可能有：①现实情况确实如此；②父母年龄填错；③子女年龄填错；④孩子非亲生导致的不规范匹配；⑤将其他同辈、上辈或者隔辈的家庭成员误填在了孩子的位置上导致的错误匹配。如果在数据清理阶段我们从多处信息来源中依然无法获得准确信息，那么，这一部分家庭将进入我们的回访名单。二是确认重组家庭中的孩子是亲生子女还是继养子女。根据 CFPS 最初的设计原则，T 表中父母-子女关系应该遵循血缘关系（含领养），而不应该包括继养。但在实地操作中，该原则并没有规范执行，使得 T 表中父母-子女关系的标准并不统一。因此，在此次回访中，我们把全部的重组家户列入了回访清单，对其子女的属性进行筛查，以便今后对血亲和继养的子女进行明确区分。三是家庭关系完全缺失的个人。这一部分人或者独立组成了单户家庭，但在现实情况中直系亲属全部没有的情况甚是少见；或者虽然属于受访的多人家庭户中的一员，但却与家中其他家庭成员之间没有任何关系，我们对这类人的身份也持怀疑态度，他们有可能不符合家庭成员资格，或者可能是漏填家庭关系。为了重新找回这些个人的家庭关系，以便对基线调查的家庭成员有更严谨的界定，我们在 2012 年对这类人群所在的家户进行了回访。

由于每个家庭的错误类型、待确认的情况以及家户资料各不相同，且需要访员在访问过

程中有一定程度的主动沟通或者追问，我们无法采用标准化的问卷和访问流程。基于此，回访采用了纸笔的访问形式，设计了个性化问卷（case-by-case questionnaire），针对具体家庭设计具体的问题、提问方式以及填答方式，用以确认和补充信息。回访与 2012 年的调查同步进行，问题家户在完成 2012 年的全部调查后，还需要完成为其特别设计的个性化问卷。

在 2012 年的访问工作完成后，我们使用回访收集到的信息，对 2010 年的相关数据进行了更新。对于第一种错误类型和第三种错误类型，我们对原有的家户关系进行了补充和修正。对于第二种错误类型，我们在每一个庄主的所有孩子的位置上新生成一组变量 bio_cN（N=1，2，…10），该变量表示该庄主对应的孩子是否与其具有血缘关系。其中，取值“1”表示有血缘关系，“0”表示没有血缘关系，“-8”表示不适用的情况，“-9”表示信息缺失。[66] bio 变量可以帮助研究者了解重组家庭中子女与父母双方的亲生及继养关系，也一定程度上解释了数据中父母-子女不互认、父母-子女年龄相差过大或过小等现象。

关于回访的更多信息还可参考《中国家庭追踪调查 2010 年家庭关系数据库清理（CFPS-7）》。

⑵ 信息回顾、确认与补充

除回访外，我们在 2012 年的调查问卷中也设计了一些问题用以对 2010 年数据中的一些变量进行了信息确认。信息确认主要针对 2010 年数据清理过程中发现的一些由于信息不够充足而难以修正的重要变量，以及由于 2010 年问卷设计的缺陷或执行上的不规范操作而受到影响的一些关键性的研究变量。具体如下：

① 受访者本人的性别与出生日期。

② 受访者 2010 年的婚姻状态以及婚姻史中的一些重要时间信息，如，结婚的日期，配偶的出生日期，婚姻解体的时间，等等。

③ 2010 年访问时婚姻状态为离婚、丧偶的受访者的上一任配偶的受教育程度。2010 年 T 表格的设计不能获取离婚、丧偶的受访者的上任配偶的教育信息。考虑到这些缺失的信息对研究家庭和婚姻的学者有用，我们在 2012 年的问卷中对这些信息重新进行了收集。

④ 教育史相关信息。CFPS 2010 年和 2011 年的设计仅仅收集了 16 岁及以上成人的教育史信息，而对于 16 以下的少儿，仅仅提问了其最高学历（未在上学者）或正在上学的阶

[66] 对于没有再婚经历的庄主，其孩子位置上的 bio 变量没有赋值。

段（正在上学者）的相关信息，并没有收集详细的教育史信息，为此，对于 2010 年和 2011 年都没有回答教育史相关问题的受访者，我们在 2012 年补充收集了相关信息。此外，2012 年的调查问卷还重新确认了 2010 年上学的状态以及正在上学的阶段。

⑤ 所有已离校受访者的离校阶段。2010 年问卷设计中对于受教育程度的提问方式是："您已经完成（毕业）的最高学历是什么？"这实际上忽视了中途离校/辍学的受访者最后一个阶段的学校教育，使得他们的受教育程度被低估。为了准确估算受访者的受教育程度，2012 年对所有已经离校的受访者重新收集了这方面的相关信息，关于受教育程度具体使用的题目是："您从哪个阶段离开学校？"

⑥ 受访者的父母信息，含父母的出生日期、职业、教育与政治面貌。CFPS 2010 没有收集已经去世父母的上述信息，为了弥补 T3 表数据上的缺陷，2012 年我们重新收集了所有人的父母信息。

在 2012 年调查结束后，我们利用采集到的以上信息对 2010 年的数据进行了更新。研究者在后期发布的 2010 年更新数据版本中将可以使用到更新后的数据，也可以根据研究需求参考 2012 年的相关数据自行调整与补充。

7. 综合变量与编码变量

7.1 受教育程度（2010）

个体的受教育程度是社会科学研究中经常使用的变量。为尽可能地避免数据的缺失，CFPS 从不同的维度对该变量的信息进行了采集， 包括：1）样本成员个人自答的受教育程度信息。2）样本成员作为其他家庭成员的关系人而由其他家庭成员代答的受教育程度信息，如家庭成员问卷回答人对所有家庭成员的代答，监护人对 10 岁以下少儿的代答，成人受访者对当前配偶的最高学历的代答，以及受访家庭成员对不能访问到的样本成员的代答。通过家庭关系的匹配，这些代答的信息可以用来插补和校正被代答样本的缺失和错误信息。3）对样本成员受教育程度变化的追踪，以及通过追踪进一步补充先前缺失或者错误的受教育程度信息。我们综合以上信息来源对样本成员在调查年份的受教育程度进行了综合的评估、补充和修正，并基于此生成了方便研究者使用的综合变量。

对于 2010 年的个人问卷受访者的受教育程度，我们为研究者提供了已完成的最高学历、上学/离校阶段[67]与受教育年限三个综合变量（见表 28）。具体来说，这三个变量综合了来自 2010 年家庭成员问卷代答值、2010 年个人问卷自答值与在 2012 年个人问卷自答值基础上逆推得到的 2010 年受教育程度（简称“逆推值”）三方面的信息。其中，2010 自答值和 2012 逆推值亦是整合了各方面信息得来的综合结果。图 16 详细展示了我们在生成受访者 2010 年受教育程度综合变量时所使用到的具体信息。

表 28. CFPS 2010 受教育程度最佳变量列表

变量名称	变量标签	变量类型
cfps2010edu_best	最高学历最佳变量	分类变量
cfps2010sch_best	上学/离校阶段最佳变量	分类变量
cfps2010eduy_best	受教育年限最佳变量	连续变量

在生成的这三个综合变量中，与已完成的最高学历不同，离校阶段这一变量整合了中途离校/辍学的受访者未完成的最后一个阶段的学校教育。同样，受教育年限的计算也考虑了这一因素：在已完成的最高学历相对应的年数（见表 29）的基础上，我们对在学/中途离校

[67] 对于正在上学的受访者，该变量记录的是受访者正在上学的阶段；对于已经离开学校的受访者，该变量记录的是受访者离开学校时所在的阶段。

/辍学的受访者加入了其未完成的最后一个阶段的受教育年数。

关于受教育程度的三个综合变量的具体生成方式、初步统计结果与数据评估请参考即将发布的技术报告《中国家庭追踪调查2010年受教育程度变量收集、清理与评估（CFPS-21）》。

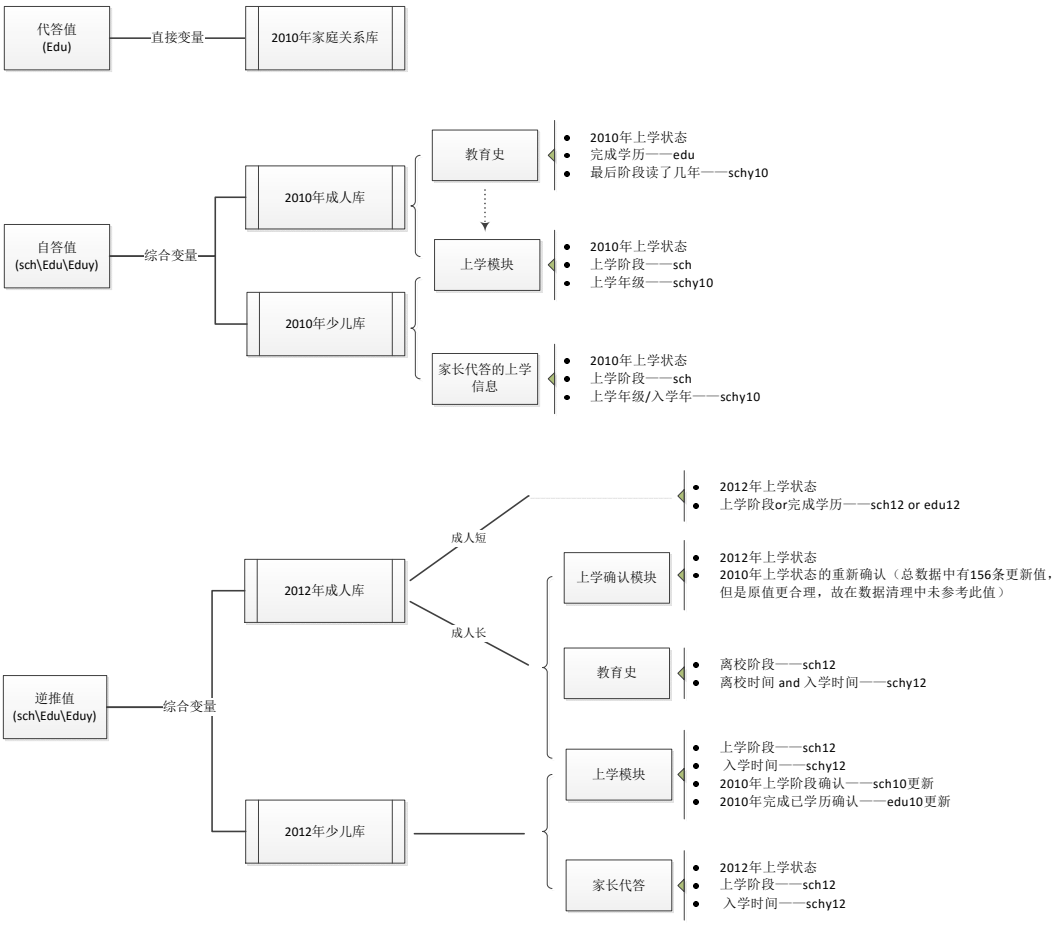


图 16.　CFPS 2010 受教育程度综合变量信息来源

表 29.　受教育年限转换表

	已完成的最高学历	年数
1	文盲/半文盲	0
2	小学	6
3	初中	9
4	高中	12
5	大学专科	15
6	本科	16
7	研究生	19
8	博士	22

7.2 抑郁程度（2010）

CFPS 2010 在成人问卷和少儿问卷采用了同一组量表来测量个人的精神状态（表 30）。我们使用因子分析输出的因子得分作为个人抑郁程度的得分，变量名为 fdepression。depression 则是 6 道题的加总得分。关于这两个综合变量的具体情况请参考《中国家庭追踪调查 2010 年综合变量（2）：受教育水平&抑郁量表（CFPS-12）》。

表 30. CFPS 2010 年抑郁量表

成人题号	少儿题号	问题
Q6	N4	下面有一些对人们精神状态的描述，请根据您最近 1 个月内的情况选择。1. 几乎每天 2. 经常 3. 一半时间 4. 有一些时候 5. 从不
Q601	N401	最近 1 个月，您感到情绪沮丧、郁闷、做什么事情都不能振奋的频率？
Q602	N402	最近 1 个月，您感到精神紧张的频率？
Q603	N403	最近 1 个月，您感到坐卧不安、难以保持平静的频率？
Q604	N404	最近 1 个月，您感到未来没有希望的频率？
Q605	N405	最近 1 个月，您做任何事情都感到困难的频率？
Q605	N406	最近 1 个月，您认为生活没有意义的频率？

7.3 认知水平

CFPS 2010 年基线调查使用了一组识字题和一组数学题来测试和评估所有需要自答个人问卷的受访者（即 10-15 岁少儿与所有成人）的认知水平。

识字题（少儿自答问卷 X2，成人问卷 X1）总共有 8 组难度水平相当的题目，每组 34 个字，按由易到难的顺序排列。在访问时，由计算机随机选择一组供受访者回答。受访者在 T1 表中填答的学历决定了答题的起点：如果为“1-2”（小学或以下），从第一个字开始顺序提问；如果为“3”（初中），从第 9 个字开始顺序提问；如果为“4-8”（高中或以上），从第 21 个字开始顺序提问。提问采用出示卡片的形式，访员请受访者念出卡片上的字，如果受访者有连续 3 个字都不认识或已经提问至第 34 个字，提问终止。

数学题为加、减、乘、除、指数、对数、三角函数、数列、排列组合等运算，共四组，组与组之间难度水平相当，由计算机随机选择一组题供受访者回答。每组 24 题，按由易到难的顺序排列。与识字题的答题规则类似，受访者答题的起点取决于他/她在 T1 表中所填答的学历：如果为“1-2”（小学或以下），从第 1 题开始顺序提问；如果为“3”（初中），从第 13

题开始顺序提问；如果为“4-8”（高中或以上），从第 19 题开始顺序提问。

考虑到以后的调查还会再次使用这两套测试题，为了不影响测试的效果，我们没有公开具体的题目。我们计算出了受访者在这两组题上的得分供用户使用，CFPS 2010 个人数据库中相应的变量名分别为 wordtest 和 mathtest。评分的基本规则是：以受访者答对的最难的一道题的题号作为他/她的最终得分；如果受访者一道题也没有答对，则以他/她的起点题的前一题的题号作为他/她的最终得分。具体的变量生成方法及初步的统计分析结果可参考技术报告《中国家庭追踪调查 2010 年综合变量（1）：字词与数学测试（CFPS-11）》。

CFPS 2014 问卷沿用了这两组测试题，但对原设计中固定三级起点的做法进行了调整：受访者的答题初始起点依然由其教育水平决定，但在初始点的第一道问题就回答错误时，起点下调到更低一级，直至退到初始起点。由于在设计上进行了调整，CFPS 2014 个人数据库中有两组相应综合变量，其中 wordtest14 和 mathtest14 的计分方法依然假设固定起点，与 2010 年变量可比；而 wordtest14_sc2 和 mathtest14_sc2 不假设固定起点。在使用这两组变量的时候应该注意的是，这是两个未经标准化处理的变量，并不代表随年龄增长或受教育水平提高而逐渐增强的识字/数学水平。用户需要根据具体的研究需要进行相应的处理。

CFPS 在 2012 年和 2016 年使用的是另两套认知测试题：记忆测试和数列测试。这两套测试题的原型来自于美国健康与退休调查（Health and Retirement Study，HRS）。在记忆测试中，访员给受访者读出 10 个在生活中常见的单词（如山、米饭、河流等），受访者在听完全部 10 个单词后立即回忆访员读出的单词，第一次尝试时一个单词都没有回忆出的受访者可以有第二次尝试的机会。此次回忆所得分数称为即时记忆得分。即时记忆测试过若干分钟后，访员会要求受访者再次回忆刚才听到的 10 个单词，此次回忆所得分数称为延迟记忆得分。测试的得分为受访者答对的单词的总数，对顺序不做要求。在 2012 个人数据库中，IWR1、IWR2 以及 IWR 分别代表受访者第一次尝试、第二次尝试以及综合两次尝试所得的即时记忆测试的得分；DWR 代表受访者延时记忆测试的得分。

数列测试使用了两阶段的适应性测试。在第一阶段，受访者回答三道数列题，由此得出其第一阶段回答正确的题目的数目（0 到 3）。在第二阶段，系统根据该数目从四组题中选择相对应的一组题发放给受访者。这四组题难度呈梯度分布，第一阶段答对题数越多的个体在第二阶段将会接受更难的测试。这一设计的基础是现代测量学理论，目标是在尽短的时间内更准确地测量出个体的真实能力。适应性测试的设计非常简洁，但它对计分提出了较高的要求。显然，传统的计分方法（按答对题数计算总分）并不适用，因为不同人群拿到的试题难

度有系统性的差别。适应性测试的计分一般基于现代测量学理论的“项目反应理论”模型计算。在该模型中，每道题都有自己的特征参数（难度、区分度等），这些特征参数不随样本人群的改变而改变；每个被试者也有自己的能力分数，这个能力分数决定了个体答对每道题的概率。“项目反应理论”的一个重要应用是受访者的得分不受其回答的具体问题的影响。由于这种计分方法较为专业，为了省去用户自行处理的麻烦，我们在发布数据库中直接提供了由Rasch 模型计算出来的得分，该分数在 CFPS 2012 中的变量名为 NS_W。同时，我们还生成了变量 NS_WSE，用来表示与 NS_W 相应的标准误。

在数据清理的过程中，我们发现 CFPS 2012 的数列测试数据存在较大程度的缺失，主要原因是在开始数列测试前，受访者如果表示他们不理解数列测试的两道例题，则停止测试。此外，我们还发现有一道题存在内容上的错误，由于这道题是两阶段适应测试第二阶段的最后一道题，我们在计算分数的过程中对这道题的答案进行了插补。在 2016 调查中，我们修正了以上错误。关于数据测试分数计算的具体信息，用户可参考技术报告《中国家庭追踪调查 2012 年数列测试题（CFPS-31）》。

7.4 收入

CFPS 收入部分的详细设计见第 3.6 节。CFPS 2010 年基线调查只询问了出售农、林、牧、副、渔等农产品所得的收入，却没有调查农村家庭自产自销部分的价值。由于中国的农村家庭会将相当一部分农产品用于自家消费，所以，CFPS 2010 已有的数据并不能准确地估算农业家庭的全部农业生产收入——对农业生产收入的计算如果忽略了自产自销的部分，则会低估农村家庭从事农业生产的真实所得，尤其会低估农产品市场化程度低的地区的农民收入，以及贫困家庭的收入。

为了弥补 2010 年调查的这一缺陷，我们利用问卷中已有的信息，设计了调整从事农业生产的家庭的农业收入的方法，生成了调整后的家庭农业生产总收入与纯收入，变量名分别为 inc_agri 和 net_agri。调整主要依据的是 2010 年家庭问卷对从事农业生产的家庭所采集的关于其具体种植/饲养的几类主要农产品的产量、销售量、销售收入与净收入的信息（家庭问卷 K6、K7 部分）。我们根据各项农产品的总产量与销售量的差值计算出农村家庭用于实物消费的农产品数量，并根据市场价格将这部分消费折算为收入，然后与已知的出售农产品所得的总收入或纯收入相加，得到家庭农业生产的总收入或纯收入。关于农业生产收入调整的具体内容可参考《中国家庭追踪调查 2010 年农村家庭收入的调整办法（CFPS-14）》。

表 31 比较了调整前后农业生产纯收入的均值、标准差、中位值的变化。[68]计算的对象是“去年”从事农业生产的家庭。可以发现，经调整，每户家庭农业纯收入的均值提高了 2469.7 元/年，中位数提高了 2275 元/年。

表 31. 调整前后家庭农业生产纯收入的比较

	均值（元）	标准差	中位数	样本量（户）
调整前	4889.0	12796.6	2105.0	7586[69]
调整后	7358.7	14044.3	4380.0	7586

由于经济情况调查的复杂性，CFPS 2010 年基线调查所设计的家庭与个人收入方面的提问项目繁多。为了给用户提供方便，我们同时还就个人收入与家庭收入生成了一系列综合变量供用户参考使用。相关的变量名与变量标签见表 32。其中，个人收入的计算方法是：首先使用自报的个人收入进行赋值；如果缺失，则用收入区间的平均值进行替代。如果依然缺失或者小于 100 元，则用分项加总的收入进行替代。[70]总收入是指没有扣除农业生产成本的收入，纯收入是指扣除农业生产成本后的收入。人均总收入和人均纯收入是用总收入或纯收入除以家庭规模得到的平均收入，计算所使用的家庭规模是 T1 表同住成员的人数。家庭总（纯）收入是五个部分收入的加总，即工资性收入、经营性总（纯）收入、财产性收入、转移性收入和其他收入。其中，工资性收入包括了工资、奖金、补贴、外出打工收入和分配到个人名下的红利。经营性收入包括了农业生产和非农经营两类收入，农、林、牧、副、渔收入归为农业生产收入，经营其他非农产业的收入归为非农经营收入。财产性收入包括了土地或其他生产资料出租收入、房屋出租收入、其他租金收入和出卖财物收入。对于存款利息、股票、基金和债券等金融资产收入，由于 CFPS 只提问了受访家庭在年底持有的本金和市值，但这些金融资产可能是受访家庭多年前购买的，所以没有计入当年的收入之中。转移性收入包括了政府补贴、离退休金、低保等政府补助。其他收入包括了亲友馈赠和受访家庭汇报的其他收入。

[68] 技术报告：CFPS-14。

[69] CFPS 2010 家庭样本中共有 7798 户家庭回答去年从事过农业生产。此处的 7586 户是家庭农业生产纯收入调整前后均无缺失值的样本，其中包含从事农业生产但居住在城市社区的家庭。

[70] 分项加总的收入包括了工资、奖金、年终奖、单位发放的福利和实物、第二职业或兼职收入、其他劳动收入、离退休金、个体经营收益、从家人和亲友得到的经济帮助、从村委会或居委会得到的经济帮助、从政府或工作单位得到的帮助或补贴。

表 32. CFPS 2010 收入综合变量及其标签

变量名	变量标签	变量名	变量标签
Income	个人收入	faminc_net_old	未调整的家庭纯收入
Firm	非农经营收入	faminc_old	未调整的家庭总收入
Finc	工资性收入	faminc	调整后的家庭总收入
Fproperty	财产性收入	faminc_net	调整后的家庭纯收入
Welfare	转移性收入	indinc	调整后的家庭人均总收入
Felse	其他收入	indinc_net	调整后的家庭人均纯收入
foperate	调整后的经营性总收入（包括农业生产和非农经营）	foperate_net	调整后的经营性纯收入（包括农业生产和非农经营）

在 3.6 节我们提到，2012 年对收入部分的设计进行了四项调整，包括弥补 2010 年部分收入项目的遗漏，对部分内容进行细化，对缺失值展开逼近式提问，以及将家庭层面的个人收入采集分散到个人问卷中。这四项调整中的前三项无疑均有助于提高 2012 年 CFPS 收入数据的质量，而第四项却存在弊端。一般来说，每个人自答的工资性收入会比由他人代答的收入更准确，这是 2012 年调查将工资性收入分散到个人问卷提问的设计出发点。但是，获得准确的全家工资性收入的前提是每一位有工资性收入的家庭成员都回答了个人问卷，且在个人问卷中都回答了个人收入。然而，实际调查难免会出现个体无应答或项目无应答的情况。我们在汇总 2012 年的家庭工资性收入时发现，有一些家庭成员从事受雇工作，却没有回答他们的工资性收入；还有一部分成员因为外出务工，没有回答个人问卷。这些无回答均会造成家庭工资性收入的低估。由于工资性收入是城乡家庭收入最重要的组成部分，这些缺失值的存在会严重低估受访家庭的总收入。为了降低数据缺失造成的影响，我们对 2012 年的工资性收入进行了插补和调整，具体的方法请参见技术报告《中国家庭追踪调 2012 年家庭收入的调整办法（CFPS-27）》。

按照收入的来源，CFPS 2012 汇总的家庭总收入仍分为工资性收入、经营性收入、转移性收入、财产性收入和其他收入五部分。但正如前面提到的，2012 年调查涵盖的收入内容更全面（见 3.6 节表 6）。基于 2012 年调查涵盖的收入项目，我们对这五个部分的收入定义进行了更新。其中，工资性收入是指家庭成员从事农业打工或从事非农受雇工作挣取的税后工资、奖金和实物形式的福利。经营性收入是指家庭从事农、林、牧、副、渔的生产经营扣除成本后的净收入和由自家生产并供自家消费的农业产品的价值，以及家庭从事个体经营或开办私营企业获得的净利润。转移性收入是指家庭通过政府的转移支付（如养老金、补助、

救济）和社会捐助获取的收入。财产性收入是指家庭通过出租土地、房屋、生产资料等获得的收入。其他收入包括亲友的经济支持和赠予、礼品礼金等。

另外，2012 年调查与 2010 年调查涵盖收入项目的差别一定程度上也造成了 2012 年与 2010 年家庭收入的不可比。考虑到追踪调查各轮次间数据比较的重要性，我们仔细比对了两期调查问卷的收入项目，生成了一组与 2010 年数据可比的 2012 年家庭收入变量。得到可比收入的具体做法包括：（1）从 2012 年家庭收入中剔除个人农业打工收入、个人实习和勤工俭学收入、奖助学金收入、征地补偿金和住房拆迁款这几项 2010 年收入调查中没有涉及或由于提问方式不一而缺乏可比性的项目；（2）剔除 2012 年家庭非农经营收入，因为 2010 年的家庭非农经营收入只提问了私营企业，没有问及个体经营，而 2012 年同时包含了这两部分收入且无法拆分。这一方法也可以通过对比表 6 中 2010 年调查与 2012 年调查收入构成明细的差别得到。因此，2012 年的家庭收入综合变量包括两个版本：一个是根据 2012 年家庭问卷的内容完整计算的家庭收入，另一个版本是与 2010 年保持一致的可比家庭收入。我们建议当使用 2012 年单期数据时，最好使用完整计算的家庭收入变量，但如果需要比较 2010 年和 2012 年的家庭经济状况时，最好使用与 2010 年可比较的收入变量。

我们就 2012 年个人收入[71]与家庭收入生成的一系列综合变量的的变量名与变量标签见表 33。变量名带“_1”后缀的是指完整计算的家庭收入，即包含表 33 中所有的收入项目；而带“_2”后缀的是指与 2010 年可比的家庭收入，仅包含与 2010 年调查一致的收入项目。此外，变量名带“_adj”后缀的表示的是经过调整的收入，但我们也保留了根据原始数据生成的收入变量。用户可以自行选择是否采用我们的调整。具体来讲，各变量所做的调整如下：

（1）家庭工资性收入 wage_1_adj、wage_2_adj：对受雇且工资性收入为 0 或缺失的个人进行了插补；对外出务工没有接受访问的家庭成员的工资性收入进行了插补。

（2）家庭纯收入 fincome1_adj、fincome2_adj：对受雇且工资性收入为 0 或缺失的个人进行了插补；对外出务工没有接受访问的家庭成员的工资性收入进行了插补；对家庭纯收入为 0 或缺失的家庭用家庭消费进行替代。

（3）家庭人均纯收入 fincome1_per_adj、fincome2_per_adj：由调整后的家庭纯收入除以家庭规模得到。

[71] CFPS 2012 的个人收入计算方法与 2010 年调查相同，即首先使用自报的个人收入进行赋值；如果缺失，则用收入区间的平均值进行替代；如果依然缺失或者小于 100 元，则用分项加总的收入进行替代。

（4）家庭人均收入分位数 fincperadj_p：由调整后的家庭人均纯收入计算得到。

表 33. CFPS 2012、2014 收入综合变量及其标签

变量名	变量标签	来源
income	个人收入	2012/2014
income_adj	个人收入（调整后）	2012/2014
wage_1	工资性收入	2012
wage_2	工资性收入（与 2010 年可比）	2012
fwage_1	工资性收入	2014
fwage_2	工资性收入（与 2010 年可比）	2014
wage_1_adj	工资性收入-调整	2012
wage_2_adj	工资性收入-调整（与 2010 年可比）	2012
foperate_1	经营性收入	2012/2014
foperate_2	经营性收入（与 2010 年可比）	2012/2014
ftransfer_1	转移性收入	2012/2014
ftransfer_2	转移性收入（与 2010 年可比）	2012/2014
fproperty_1	财产性收入	2012/2014
fproperty_2	财产性收入（与 2010 年可比）	2012/2014
felse_1	其他收入	2012/2014
felse_2	其他收入（与 2010 年可比）	2012/2014
fincome1	全部家庭纯收入	2012/2014
fincome2	家庭纯收入（2010 可比）	2012/2014
fincome1_adj	全部家庭纯收入-调整	2012
fincome2_adj	家庭纯收入-调整（2010 可比）	2012
fincome1_per	人均家庭纯收入	2012/2014
fincome2_per	人均家庭纯收入（2010 可比）	2012/2014
fincome1_per_adj	2011-2012 人均家庭纯收入-调整	2012
fincome2_per_adj	2011-2012 人均家庭纯收入-调整（2010 可比）	2012
fincper_p	家庭人均收入分位数	2012
fincome1_per_p	人均家庭纯收入分位数	2014
fincperadj_p	家庭人均收入分位数-调整	2012

CFPS 2014 的调查问卷在经营性收入、转移性收入、财产性收入和其他收入四个方面沿袭了 2012 年问卷的结构，使得两年调查的四项收入基本可比（见 3.6 节表 6）。但是如前所述，2012 年的家庭工资性收入通过加总家庭中每一份个人问卷中的工资性收入得到，由于实际调查中不可避免的会发生个体无应答或项目无应答的情况，这一做法会造成家庭工资性收入的低估。为了避免这一弊端，在 2014 年，我们将对家庭工资性收入的提问又重新放回了家庭问卷中，由家庭问卷的回答人回答家庭整体的工资性收入。值得注意的是，对于农村样本，家庭的工资性收入主要来源于外出打工者，而家庭问卷的回答人并不一定了解他们在

外的收入情况，因而只能以其寄回家中的钱物为基础估计，这又可能造成收入的低估。因此，在实际数据清理过程中，我们以家庭问卷报告的工资性收入为基础，如果遇到缺失值、0 值或者农村家户个人问卷加总的工资性收入大于家庭问卷报告的收入的情况，则用个人问卷加总的工资性收入进行插补。

2014 年调查与 2012 年调查的收入项目完全可比，但与 2010 年调查的收入项目存在差别。因此，我们同样生成了一组与 2010 年数据可比的 2014 年家庭收入变量。我们同样建议当使用 2012 年和 2014 年单期或两期数据时，最好使用完整计算的家庭收入变量，但如果需要与 2010 年的家庭收入数据进行比较时，最好使用与 2010 年可比较的收入变量。CFPS 2014 收入相关的综合变量名与变量标签见表 33。

7.5 家庭支出

CFPS 在家庭问卷中采集家庭的整体开支情况，我们在本报告第 3.6 节介绍了 CFPS 历轮调查中详细的支出项目（见第 3.6 节表 7）。总的来说，这些项目可以分为四大类别：一是消费性支出，指家庭衣食住行等日常开销，具体包括食品、衣着、居住、家庭设备及用品、交通通讯、文教娱乐、医疗保健、其他消费性支出八项子类；二是转移性支出，指家庭对非同住亲戚和朋友的经济支持、家庭的社会捐助，以及家庭重大事件中礼金或礼物的支出；三是保障性支出，指家庭购买各类商业保险的支出；四是购房建房支出，含偿还按揭的支出。各类别支出项目的具体构成参见表 34。

表 34. CFPS 2010/2012/2014 家庭支出项目构成明细

家庭支出项目[72]	CFPS 2010	CFPS 2012	CFPS 2014
消费性支出			
1.食品	家庭食品支出	自家消费香烟、酒水	
	自家消费食品（插补）	自家消费食品	自家消费食品
		消费的自家生产农产品价值	消费的自家生产农产品价值
2.衣着	家庭衣着支出	衣着消费	衣着消费
3.居住	家庭租房支出(不含住房按揭）	房租	房租
	家庭居住支出（如物业、取暖等，不含住房按揭及房租)	水电费	水电费
		燃料费	燃料费

[72] 表 7 中的生产经营支出不计算在家庭支出中。

		集中供暖取暖费	集中供暖取暖费
		物业费（含车位费）	物业费（含车位费）
4.家庭设备及用品		购买汽车支出	购买汽车支出
		购买及维修汽车外的其他交通工具支出	购买及维修汽车外的其他交通工具支出
		购买可办公类电器支出	
	家庭家电支出	购买家具和其它耐用消费品支出	购买家具和其它耐用消费品支出
	家庭日常用品支出	家庭日用品支出	家庭日用品支出
	家庭杂项商品、服务支出	雇佣保姆小时工	
5.交通通讯	家庭通信支出	邮电、通讯支出	邮电、通讯支出
	家庭出行支出(含养车费用)	本地交通费（包括汽油费）	本地交通费（包括汽油费）
6.文教娱乐	家庭教育支出	教育支出	教育支出
	家庭文化、娱乐、休闲支出	文化娱乐支出	文化娱乐支出
		旅游支出	旅游支出
7.医疗保健	家庭医疗保健支出	直接支付的医疗费用	直接支付的医疗费用
		保健及健身费用	保健及健身费用
8.其他	自家婚丧嫁娶支出		
	家庭其他支出	其他生活消费支出	其他生活消费支出
转移性支出	给机构/个人捐赠过钱物总价值	上交给政府部门的税费杂费	
		社会捐助（现金及实物）	社会捐助（现金及实物）
		给非同住亲戚的经济支持	给非同住亲戚的经济支持
			给亲属其他人的经济帮助
			重大事件（结婚、生小孩、升学等）礼金礼物支出
保障性支出	购买商业保险类支出	商业性医疗保险支出	商业性医疗保险
		商业性财产保险（含汽车险）支出	商业性财产保险支出
		缴纳各类养老保险	
购房与建房按揭支出	住房按揭支出	住房按揭支出（插补）	住房按揭支出

我们为用户生成了家庭总支出的综合变量，该综合变量的取值为上述四大类支出的加

总，如果家庭没有发生某项支出，则该项支出记为0。除了提问分项支出外，CFPS 2014还专门提问了过去12个月家庭的总支出或者总支出的区间。在生成家庭总支出的综合变量时，我们以分项支出的加总为基础，仅当分项支出的加总小于100或缺失（所有支出项目的回答都为不适用/拒答/不知道）时，才采用受访人回答的家庭总支出作为插补取值。

由于不同类型支出发生的频率不同，CFPS使用了不同的回忆时段采集相关信息，包括过去1周、过去1个月和过去12个月。家庭总支出的综合变量以过去12个月为基准，如果某项支出的回忆时段为过去一周，则将其换算成52周[73]的支出（周支出*52周）；如果某项支出的回忆时段为过去1个月， 则将其换算为12个月的支出（月支出*12月）。少数受访者由于不适应回忆时段在不同项目之间的切换而使用了错误的回忆时段，如误将 1 个月的开支回答为过去12个月的开支。因此，在构造分项开支的综合变量时，我们借助收入分位数来甄别严重脱离收入水平的开支水平，并对奇异值做了调整。[74]

尽管CFPS支出问卷的明细内容及提问形式在不同调查轮次间的变动（见3.6节表7）在一定程度上影响到了测量的精度，但总的来说，支出的大类基本一致，跨年份的消费性支出和总支出的综合变量基本可比。CFPS 2010、2012、2014支出相关的综合变量名与变量标签见表35。

表35. CFPS 2010/2012/2014支出综合变量及其标签

变量名	变量标签
pce	居民消费性支出-加总
food	食品支出
dress	衣着支出
house	居住支出
daily	家庭设备及日用品支出
med	医疗保健支出
trco	交通通讯支出
eec	文教娱乐支出
other	其他消费性支出
eptran	转移性支出
epwelf	福利性支出
mortage	建房购房贷款支出
expense	家庭总支出

[73] 一年为52周。

[74] 例如，某家庭回答过去1个月的分项支出水平超过了其所在收入分位数家庭在该分项平均支出水平的12倍，则我们推断该家庭很可能是按照过去12个月为回忆时段来回答这项收入，对此，我们会将其该项支出水平除以12作为调整。

7.6 家庭财产

在 CFPS 2010 和 CFPS 2012 的家庭问卷数据库中，家庭净资产用 total_asset 变量表示，为家庭总资产与家庭总负债之差。其中，家庭资产包括土地、房产、金融资产、生产性固定资产和耐用消费品；家庭负债包括住房负债和非住房负债。土地的价值通过估算得到：如，家庭农业总收入的 25%来源于土地，而土地的收益率为 8%，土地的价值便可以由此推算得出（McKinley，1993）。房产包括现住房和其他房产总价值。由于无法确定部分产权的比例而且家庭具有永久使用权，在计算房产价值时，我们对于部分产权房的价值也是按全部产权计算的。金融资产包括存款、股票、基金、债券、金融衍生品、其他金融产品及借款，其中 2010 年数据不包含债券、金融衍生品以及其他金融产品的价值。生产性固定资产包括经营性企业资产、农业使用机械等。耐用消费品包括汽车、电视、电脑、冰箱等家庭常见消费品。住房负债来源于调查时受访者自答的“连本带息尚未还清的房贷”。非住房负债来源于教育、医疗等方面的债务。表 36 陈列了与财产相关的发布变量，这些变量的详细清理方法请参考技术报告《中国家庭追踪调查 2012 年和 2010 年财产数据技术报告（CFPS-29）》。

表 36. CFPS 2010/2012 财产类综合变量及其标签

变量名	变量标签	来源
land_asset	土地价值（元）	2012/2010
houseasset_gross	总房产-未减房贷（元）	2012
resivalue_new	现居住房子市场总价（元）	2012/2010
houseprice1_best	现住房当前市场总价（元）	2012
otherhousevalue	其他房产总价值（元）	2012/2010
houseprice2_a_1_best~ houseprice2_a_6_best	离您家渐远的 N 房产当前市场总价（元）	2012
house_debts	总房贷（元）	2012/2010
house1_debts	现住房房贷（元）	2012
houseother_debts	其他房产贷款总额（元）	2012
fixed_asset	生产性固定资产（元）	2012
company	公司资产（元）	2012/2010
agrimachine	农业器械价值（元）	2012
finance_asset	金融资产（元）	2012
savings	现金和存款总值（元）	2012/2010
govbond	政府债券（元）	2012
stock	股票（元）	2012/2010
funds	基金（元）	2012/2010

derivative	金融衍生品（元）	2012
otherfinance	其他金融产品（元）	2012
debit_other	别人欠自家的钱（元）	2012/2010
nonhousing_debts	非房贷的金融负债（元）	2012/2010
bank_debts	除房贷外的银行贷款（元）	2012
ind_debts	欠非金融机构贷款（元）	2012
durables_asset	耐用消费品价值（元）	2012
total_asset	家中净财产（元）	2012
valuable	家中收藏品（元）	2010
otherasset	家中其他资产（元）	2010

7.7 职业编码

CFPS 2010 年基线调查采用了访员实地编码和编码员事后集中编码两种方式对受访者的职业和行业信息进行了编码。从 2012 年起，CFPS 取消了访员的实地编码，集中由编码员对上传的问卷数据进行后期编码。[75]表 37 总结了职业编码所涉及到的所有职业变量。

在基线调查时，访员实地编码的题目为成人问卷中的 G307、G308、H405、H406 四题。访员可以在 CAPI 系统下根据 CFPS 职业和行业代码字典，对受访人的职业和行业进行现场分类编码。[76] 编码系统的设计采用了简单的查询法，访员首先确认受访人的职业属于哪一大类，然后可以逐级点击，最后确认四级代码为最终代码。编码系统的访问界面见图 17。

除上述四题外，其它职业和行业编码题目均为开放性题型，在调查结束后由经过专业训练的编码员进行集中的手工编码。在编码流程上，我们采用了双向独立验证并判定（Two-way Independent Verification with Adjudication）的质控方式。第一轮编码由两个编码员分别单独对每一个受访者的职业信息进行编码，若结果一致，则确定为最终职业代码；若不一致，则由另一位经验较为丰富的编码员对这些不一致的条目独立进行第二轮编码。第二轮编码的结果如果与前一轮编码结果中的一个保持一致，则确定该编码为最终职业代码；若三人编码结果均不一致，则由专业的研究人员根据编码员编码、访员实地编码以及相关的辅助信息进行判定，从而形成最终的职业代码。

[75] 我们对 CFPS 2010 的访员实地编码数据与编码员后期编码数据进行比较，集中的编码员后期编码质量较 高，因此从 CFPS 2012 开始只采用编码员后期编码的方式。详细信息可以参考《2010 年职业行业编码（CFPS-8）》。

[76] 由于计算机辅助编码技术仍处于研究性尝试阶段，加之中国职业特征多样、职业分类复杂给技术带来的难度，为确保万无一失，我们同时也采集了详细的职业信息（G306）。

表 37. 职业行业编码题目及相关辅助题目

问卷	题号（变量名）	题目
CFPS 2010		
家庭成员问卷	B5（tb5）	“家庭成员姓名”的主要工作是？
家庭成员问卷	D6（td6）	“不同住直系亲属姓名”的主要工作是？
成人问卷	B309（qb309）	“兄弟姐妹姓名”的职业是？
成人问卷	G303（qg303）	您现在主要是在哪个机构工作？
成人问卷	G304（qg304）	您现在工作单位的名称？
成人问卷	G305（qg305）	您现在主要工作的机构属于？
成人问卷	G306（qg306）	您的职业是？
成人问卷	G307（qg307code）	您的职业属于哪一类？
成人问卷	G308（qg308code）	您工作属于哪个行业？
成人问卷	G601（qg601）	您的第一份工作是什么？
成人问卷	G701（qg701）	您的第二职业是什么？
成人问卷	H404（qh404）	您非农工作的职业是？
成人问卷	H405（qh405code）	您的非农工作属于哪类职业？
成人问卷	H406（qh406code）	您的非农工作属于哪个行业？
少儿问卷	J2（wj2）	你干过的最主要的正式工作内容是什么？
CFPS 2012		
成人问卷	E209B（qe209bcode_best）	您现在配偶的具体职业：
成人问卷	V103（qv103code_best）	您 14 岁时，父亲的具体职业是
成人问卷	V203（qv203code_best）	您 14 岁时，母亲的具体职业是
共用模块	G410（qg410code）	您/“受访者姓名”这份工作的单位主要是做什么的，也就是，他们制造什么产品或者从事什么活动？
共用模块	G411（qg411code）	您/“受访者姓名”在这份工作中具体做什么事情？
共用模块	G509（qg509code）	您主要做什么生意，即生产什么产品或者从事什么经营活动？
共用模块	G510（qg510code）	您具体从事什么工作？
共用模块	KS8	您/你将来最希望从事什么类型的职业：
共用模块	S801（ks801）	您/你将来最希望从事的具体职业是什么呢？
少儿问卷	wd1	您希望孩子长大后从事什么类型的职业？
少儿问卷	D101（wd101）	您希望孩子长大后具体从事什么职业呢？
CFPS 2014		
成人问卷	B303（qb303code）	“兄弟姐妹姓名”生前占用时间最多的主要职业是？
成人问卷	EA203（qea203code）	“配偶/同伴称呼”的具体职业是：
成人问卷	EB4022（eeb4022code）	对方的具体职业是：
成人问卷	G302（qg302code）	您/你这份工作的单位主要是做什么的，也就是，他

		们制造什么产品或者从事什么活动？
成人问卷	G303（qg303code）	您/你在这份工作中具体做什么事情？
成人问卷	GA4（qga4code）	过去 12 个月，你干过的最主要的一份实习/兼职的单位主要是做什么的，也就是，他们制造什么产品或者从事什么活动？
成人问卷	GA401（qga401code）	您在这份实习/兼职工作中具体做什么事情？
成人问卷	G1401（qg1401）	您/你的行政/管理职务是什么？
共用模块	KS8	您/你将来最希望从事什么类型的职业
共用模块	KS801（ks801code）	您/你将来最希望从事的具体职业是什么呢？
少儿问卷	wd1	您希望孩子长大后从事什么类型的职业？
少儿问卷	D101（wd101code）	您希望孩子长大后具体从事什么职业呢？
少儿问卷	GA4（wga4code）	过去 12 个月，你干过的最主要的一份实习/兼职的单位主要是做什么的，也就是，他们制造什么产品或者从事什么活动？
少儿问卷	GA401（wga401code）	您在这份实习/兼职工作中具体做什么事情？

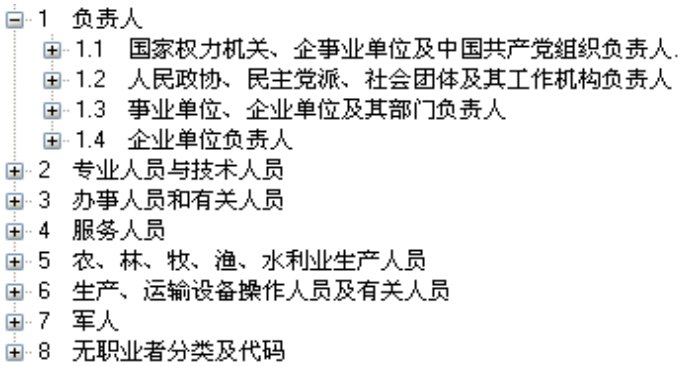


图 17. CFPS 2010 年访员实地编码系统界面显示[77]

由于后期经研究和比较分析发现编码员的编码结果与访员实地采访时的编码结果差异较大，G307、G308、H405、H406 四题最后还是采用了编码员集中编码的最终结果。

由于在对基线调查的问卷进行设计时，GB/T 6565-2009 版本的国家标准职业分类标准尚未发行，所以，2010 年问卷采用的职业编码系统是在 GB/T 6565-1999 版本的国家标准职业分类代码体系的基础上，借鉴了 CGSS（“中国综合社会调查”）的职业和行业分类标准进行修订后的职业代码表，包括了 8 大类，共计 595 个职业代码。行业分类采用了国家统计局的标准，共 20 类。

[77] 转引自技术报告：CFPS-8。

CFPS 的所有事后集中编码改用了最新的国家标准职业分类与代码表，即 GB/T 6565-2009 版本。CFPS 完全采用了该编码表的分类和排序，只是对其代码进行了重命名。

关于 CFPS 2010 年基线调查的职业编码的具体技术方法与质量评估可参考技术报告《中国家庭追踪调查 2010 年职业行业编码（CFPS-8）》。

如前文所述，CFPS 自 2012 年起开始收集调查期间全部工作的情况。由于 2012 年问卷设计的特点，在 2012 年的调查中，如果受访者在过去一年从事了农业、非农、个体经营等多种类型的工作，数据用户需要自行从中鉴别和生成当前最主要的职业的变量。考虑到工作模块的流程比较复杂，CFPS 工作人员综合了 2012 年调查的相关数据，为用户生成了 2012 年当前主要职业综合变量。具体的整理方法可以参考技术报告《中国家庭追踪调查 2012 年当前主要职业综合变量的建构（CFPS-30）》。

CFPS 2014 对两次调查期间没有全职工作经历的人和有过全职工作经历的受访者分别采取不同的策略采集数据。对于两次调查期间没有全职工作经历的人，CFPS 2014 采集了受访者的实习兼职工作的相关信息，并且对于最主要的实习兼职工作，生成了相应的行业编码变量和职业编码变量。而对于两次调查期间有过全职工作经历的人，CFPS 2014 采集了受访者在此期间全部工作的信息，并且设计了专门的主要工作模块来采集受访者最主要工作的相关信息，生成了最主要工作的行业编码变量及职业编码变量；除最主要工作之外的一般工作信息在 EHC-Job 模块中采集，一般工作没有采集相应的行业及职业信息。

7.8 职业代码转换（2010）

CFPS 2010 年基线调查的职业分类采用的是国家标准职业分类（Chinese Standard Classification of Occupations，CSCO）的代码体系（GB/T 6565-2009）。为了给用户提供方便，我们对 CFPS 的职业代码进行了转换，创建了如下几个与职业相关的综合变量：

⑴ 与 CFPS 职业分类代码相对应的国际标准职业分类（International Standard Classification of Occupation，ISCO-88）代码。该综合变量的命名方式是在原变量名后加“_isco”后缀。

⑵ 依据 ISCO-88 职业代码建构了两套职业社会经济地位指标：国际标准职业社会经济指数（International Socio-Economic Index of Occupational Status，ISEI）和标准国际职业声望

量表（Treiman’s Standard International Occupational Prestige Scale，Treiman’s SIOPS）。这两类综合变量的命名方式分别是在原变量名后加“_isei”、“_siops”后缀。

⑶ 成人问卷受访者职业的 EGP（Erikson and Goldthorpe’s Class Categories）职业分类代码。该综合变量的命名方式是在原变量名后加“egp”后缀。

表 38. 职业类综合变量命名示例

	父亲的主要工作（家庭成员问卷 B5）	兄弟姐妹 1 的职业（成人问卷 B309）	本人职业事后编码（成人问卷 G307）
CFPS 职业变量	tb5_code_a_f	qb309_occu_1	qg307code
职业 ISCO-88 代码	tb5_isco_a_f	qb309_isco_1	qg307isco
职业 ISEI 得分	tb5_isei_a_f	qb309_isei_1	qg307isei
职业 SIOPS 得分	tb5_siops_a_f	qb309_siops_1	qg307siops
职业 EGP 分类	——	——	qg307egp

除了对受访者本人的职业进行代码转换外，我们还对家庭问卷中其他家庭成员和成人问卷中兄弟姐妹的职业进行了代码转换，具体的命名规则可参考表 38。此外，我们还对用户提供从 CFPS 职业代码（CSCO）至 ISCO88、ISEI、Treiman’s SIOPS 和 EGP 值的 Stata 转换命令。

关于职业类综合变量创建的具体方法与内容可以参考技术报告《中国家庭追踪调查 2010 年职业社会经济地位测量指标构建（CFPS-10）》。

7.9 方言编码

CFPS 基线和后续追踪调查问卷分别通过受访者自答以及访员观察采集了受访者平时交谈使用的语言，以及访问过程使用的语言。 CFPS 方言编码的主要依据是《中国语言地图集》，编码由六位数字构成，分别是语系 1 位、语族 1 位、大区 (Supergroup or Group)1 位、区片（Group or Subgroup）1 位、片（Subgroup）2 位。由于这里的编码主要考虑到汉族的语言分布，因此六位编码的前两位总是 11（汉藏语系中的汉语语族）。剩余的四位编码中，第一位代表大区或者区（官话大区，晋语区、吴语区，等等）；第二位代表区或者片（比如，东北官话，或者晋语区中的并州片）；第三位和第四位代表官话区中的片（比如东北官话中

的吉沈片）。编码时，编码员通过被访者填写的文字信息，并结合其所在区县，按照《中国语言地图集》进行编码。具体的方言编码规则可以参考技术报告《中国家庭追踪调查方言编码（CFPS-28）》。表 39 为 2010、2012、2014 年问卷中有关方言的问题及相关变量名。方言变量属于限制性使用的数据，用户如果需要使用方言变量，可以联系 CFPS 项目组进行申请。

表 39. 方言编码题目及变量名

年份	问卷类型	题号（变量名）	题目	类型
2010	成人问卷	D2（QD2）	您平常与家人交谈主要使用什么语言	选择题
2010	少儿问卷	K2（WK2）	您平常与家人交谈主要使用什么语言	选择题
2010	公共模块	S3（KS3）	您在学校与同学日常交谈主要使用什么语言	选择题
2012	家庭成员问卷	Z103（KZ103）	访问过程中主要使用以下哪种语言	选择题
2012	家庭成员问卷	Z104	具体是什么方言	开放题
2012	家庭问卷	Z103（KZ103）	访问过程中主要使用以下哪种语言	选择题
2012	家庭问卷	Z104	具体是什么方言	开放题
2012	成人问卷	D201（QD201）	您平常与家人交谈主要使用什么语言	选择题
2012	成人问卷	Z103（QZ103）	访问过程中主要使用以下哪种语言	选择题
2012	成人问卷	Z104	具体是什么方言	开放题
2012	少儿问卷	Z103（KZ103_B_1）	访问过程中主要使用以下哪种语言	选择题
2012	少儿问卷	Z104	具体是什么方言	开放题
2012	少儿问卷	K2（WK2）	你平常与家人交谈主要使用什么语言	选择题
2012	少儿问卷	Z103（KZ103_B_2）	少儿访问过程中主要使用以下哪种语言	选择题
2012	少儿问卷	Z104	具体是什么方言	开放题

2012	共用模块	S3M（KS3M）	您/你在学校与同学日常交谈主要使用什么语言	选择题
2014	家庭成员问卷	Z103（KZ103）	访问过程中主要使用以下哪种语言	选择题
2014	家庭成员问卷	Z104	具体是什么方言	开放题
2014	家庭问卷	Z103（FZ103）	访问过程中主要使用以下哪种语言	选择题
2014	家庭问卷	Z104	具体是什么方言	开放题
2014	成人问卷	Z103（QZ103）	访问过程中主要使用哪种语言	选择题
2014	成人问卷	Z104	具体是什么方言	开放题
2014	少儿问卷	Z103（KZ103_B_1）	访问过程中主要使用哪种语言	选择题
2014	少儿问卷	Z104	具体是什么方言	开放题
2014	少儿问卷	Z103（KZ103_B_2）	少儿访问过程中主要使用以下哪种语言	选择题
2014	少儿问卷	Z104	具体是什么语言	开放题
2014	公用模块	S3M（KS3M）	您/你在学校与同学日常交谈主要使用什么语言	选择题

7.10 最佳变量

在数据清理的过程中，我们对部分需要修正的变量没有进行直接修改，而是在原始变量的基础上生成一组新的变量来保存修改后的值，以供用户参考，原变量同样保留并发布。由于新生成的变量是我们参考数据中各方面的信息得出的最合理取值，所以我们将其命名为“最佳变量”，变量名的命名规则为：在原变量名之后加上后缀“_best”。用户需要注意的是：“最佳”并不意味着是一定正确的取值，而是我们在已有填答的基础上，结合多个信息来源与逻辑关系判断在目前所能得到的最合理取值。用户可以自行决定是使用原始值还是使用最佳变量。

每一轮的发布库都可能存在若干最佳变量，除 7.1 节提到的受教育程度最佳变量以及财产最佳变量外，我们以下对 2010 年的几个最佳变量加以介绍：qa1y_best，qe605y_best，qe606y_best，qe1_best。这四个最佳变量全部位于成人问卷数据库中。

qa1y 是描述成人出生年份的变量。2010 年基线调查中，对于 T1 表成员，我们最多可以从三个来源获知其出生年份（见表 40）。其中，来源 1 是家庭成员问卷代答的信息；来源 2 是受访者个人自答的信息；来源 3 是将夫妻双方匹配之后，从受访者配偶的回答中间接得到的受访者本人的信息，但前提是受访者的现任配偶（可能为初婚配偶）同样也回答了个人问卷。

表 40. 出生年份信息来源

	问卷	题号	题目
来源 1	家庭成员问卷	B1	“家庭成员姓名”的出生日期？
来源 2	成人问卷	A1	请问您的出生日期是？
来源 3	成人问卷	E606/E302/E211	请问您初婚配偶/现在的同伴/现在的配偶的出生年份是？

我们经过统计分析发现有些受访者三个来源的信息并不一致，这将给用户在出生年份变量的使用上带来不便。因此，我们使用手工更改与计算机程序更改两种方式，通过对出生年份与其它多个生命事件的逻辑关系的检验，排除了不合理的取值，并综合了 2011 年调查中的年龄信息，产生了最佳出生年份的取值，即 qa1y_best。我们之所以没有覆盖原始的取值，一是因为有些取值虽然是经多重途径综合考虑后的最合理取值，但我们并不能 100%确认其一定正确；二是因为原始取值是个人问卷产生的依据，我们必须将其保留，否则用户将在问卷生成规则上产生不必要的误解。

qe1 是成人问卷中提问访问当时婚姻状况的变量。同时，在家庭成员问卷中，我们也有代答的婚姻状况信息。同样，我们发现二者存在信息不一致的问题。此外，在前面提到的家庭关系库的清理中，我们检验出部分逻辑错误，如，婚姻状况为未婚，但是在 T2 表中有能够匹配的配偶，或者个人为离婚、丧偶，但是 T2 表中有可以与之匹配的健在配偶，等等。这些逻辑错误有少数是因为婚姻状况填错导致。为此，我们生成了最为合理的婚姻状况 qe1_best 供用户参考。与 qa1y 一样，我们也没有直接修改原始取值，因为婚姻状态直接影响受访者在婚姻部分的答题流程，原始的 qe1 的取值是婚姻部分不同模块选择的依据，一旦更改，也会给用户理解数据带来不必要的麻烦。

CFPS 在婚姻史模块使用了回顾（retrospective）的方式收集受访者一生中重要的婚姻事件及其发生或者变化的时间，以满足研究者研究婚姻以及婚姻与其他事件关系的需要。但是，回顾的方式以及对诸多敏感问题的提问，给数据的准确性提出了挑战。由于受访者记忆的偏差，或者有意识地隐瞒某些事实，我们在对相关数据进行清理的过程中发现了逻辑不合理（如，第二次婚姻的结婚时间早于第一次婚姻的解体时间）、同一信息在不同的来源填答不一致（如夫妻双方所报的结婚时间不一致、受访者本人所报的配偶出生年份与其配偶自报的出生年份不一致）、违背常识（如，初婚年龄在 16 岁以下）等问题。我们挑选了最重要的两个变量——初婚年份 qe605y 和初婚配偶的出生年份 qe606y——生成了其对应的最佳变量，分别为 qe605y_best 和 qe606y_best。这两个最佳变量产生的原则是，在经过家庭关系数据库清理排除掉匹配错误后，结合不同来源的信息（如，适婚的年龄区间、婚后生育第一个孩子的年份等），排除掉不合理取值，保留合理取值，依此生成相对应的最佳变量。我们保留原有的变量，一方面是因为我们同样并不能确认我们给出的值一定是准确无误的，而只能说是我们认为的最为合理的取值，用户应该根据自己的需要进行选择；另一方面，我们也认为有些填报不一致的情况本身也具有研究价值，比如说，夫妻双方对于结婚时间的填报不一致，夫妻双方对于配偶出生年份的记忆不一致，等等。

关于这四个最佳变量的具体的生成规则与方法可以参考技术报告《中国家庭追踪调查 2010 年综合变量（3）：年龄、婚姻最佳变量（CFPS-13）》。

7.11 特殊数据处理说明

CFPS 严格保护受访者的个人信息，出于保密的需要，我们对以下变量进行了相应处理：

（1）不发布所有姓名变量。我们会提供相应的个人编码供用户使用，这不会对研究产生影响。

（2）不发布所有出生日期变量。对于出生时间，我们会提供年份与月份的变量，但不会提供日期变量。

（3）对于样本的地址，仅提供省级层面的具体地址信息。对于区/县和村/居层面的地址信息，我们仅提供具有区分度、但无实际具体含义的虚拟代码，用户无法将其定位到具体的区/县和村/居。

（4）2010 年成人问卷 G1 题选项 5 和 6 合并为“其它”，G103 不公开。

（5）2010 年成人问卷 M 706 不发布。2010 年成人问卷 M 706 不发布。

（6）受合作机构委托的搭载模块数据，可能延缓发布。

7.12 个别变量使用说明

表 41 对部分在使用时需要特别注意的变量以及在以上介绍中未涵盖的综合变量进行了简要说明：

表 41. 个别变量使用说明

内容	数据库	题号（变量名）	简要说明
问卷变量			
60 岁以上的受访者与其子女的交往及关系	成人问卷	F1-F3（qf1-qf3）	此处加载的是家庭成员问卷 T 表格中的子女信息。在上文数据清理部分提到，原始的 T 表格存在一些匹配错误。因此，此题中加载的子女不一定正确，这一定程度上会影响此题的使用价值。2012 年我们加载了经清理后正确的子女信息后，对相关内容进行了重新提问。[78] 有需求的用户可以使用 2012 年的数据。
地址类题目，举例：出生地、户口所在地	成人问卷	A102、A201(qa102acode、qa201acode)	仅公布到省级层面的地址信息，但是我们会生成相对的地址变量，说明该地区是否与调查所在地是同一个省/市、区/县、乡镇/街道、村/居。其它地址类题目同。
职业期望	成人问卷	S8、S801	由于系统设置跳转错误，这两道问题没有收集到信息，故删去。
职业期望	少儿问卷	D101(wd101)、M601(wm601)	已经编码。编码方法参见《中国家庭追踪调查职业期望编码（CFPS-9）》。
与人交谈使用什么语言	成人&少儿问卷	成人问卷 D2(qd2)、少儿 K2(wk2)、共用问卷 S3(ks3)	已经编码。编码方法参见《中国家庭追踪调查方言编码（CFPS-20）》。
单位的使用，举例：你家离最近的高中的距离有多远？___米/里/公里	家庭问卷	A6(fa6)	CFPS 在调查中提供多个单位供受访者使用。但在后期数据整理中，我们已将单位进行了统一，统一后的单位可从变量标签得知。

[78] 需要注意的是，2012 年对这一系列问题的重新提问，虽然加载的是 2010 年的子女信息，但调查的是 2012 年的情况，并不是对 2010 年信息的回溯。

多选题的存储，举例：最近一个月，你食用下列哪些食物？	少儿问卷	L5（wl5）	CFPS 2010 年对多选题的存储方式并不是生成针对每个选项的 0、1 变量。而是生成了一组变量 w15_s_1-wl5_s_9，其中，wl5_s_1 的值代表受访者选的第 1 个选项，wl5_s_2 的值代表受访者选的第 2 个选项，依此类推。如果受访者总共选择了 3 选项，则从 wl5_ s_4 开始，一律存储为"-8"，其它多选题同。
"其他"类选项，举例：请问您目前没有工作的原因是什么？ 77. 其他【请注明】_____	成人问卷	J101（qj101）	半开放性的问题中，受访者在"77"选项中所注明的文字信息不发布。其它半开放性问题同。
综合变量			
发布版本	各库	releaseversion	各库均包含一个发布版本变量，用户可通过项目网站的"CFPS 数据和文档更新"页面查看各库的最新版本号，确认自己所用数据是否为最新版本。
家庭规模	关系库	familysize	家庭成员的人数（包含住在家里的和有经济有联系的外出成员）
代际数	关系库	Generation	根据家庭成员构成及家庭关系，计算出的家庭代际数
家庭成员的判断	关系库	co_aXX_p	0-1 变量，表示 XX 轮调查该个人是否是家庭成员（或是否同灶吃饭），是判断家庭成员及家庭规模的依据。
父母背景变量	关系库 成人库 少儿库	fbirthy、feduc、foccupcode 、foccupisoc 、fparty、mbirthy 、meduc、moccupcode 、moccupisoc 、mparty、fbirthy12 、feduc12、mbirthy12 、meduc12、	根据 CFPS 2010、2012 采集的信息，在 2010 年数据库中分别生成父亲、母亲的出生年份、最高学历、主要职业两类编码、政治面貌的综合变量。在 2012 年数据库中生成父母出生年份和最高学历的综合变量。
经济问卷关联家户号及相关类型	家庭库	overlapfidX ，overlapfidXtype	CFPS 2014 年及后续调查年变量，X=1,2,3,4.代表家庭问卷与该家庭有重叠成员的家户，以及重叠的类型，详情可参考《中国家庭追踪调查 2014 年数据库介绍及数据清理报告（CFPS-34）》
城乡分类变量	成人、少儿库	urban urban12 urban14	urban 、 urban12 、 urban14 分别代表样本在 CFPS2010、CFPS2012、CFPS2014 调查时所在村

			居的城乡状态（国家统计局定义）。每个变量只存在于相应调查年的数据库中。
有无自答问卷	成人、少儿库	selfrpt	此观测是否有自答问卷数据
自答问卷模式	成人、少儿库	self_IWmode	此观测的自答问卷数据是面访还是电访模式
是否为自答中断	成人、少儿库	Interrupt_SF	此观测的自答问卷数据是否为中断样本
有无代答问卷	成人、少儿库	proxyrpt	此观测是否有代答问卷数据
代答问卷模式	成人、少儿库	proxy_IWmode	此观测的代答问卷数据是面访还是电访模式
是否为代答中断	成人、少儿库	Interrupt_PR	此观测的代答问卷数据是否为中断样本

8. CFPS 2010 年基线调查数据初步统计分析和评估[79]

8.1 性别年龄分布

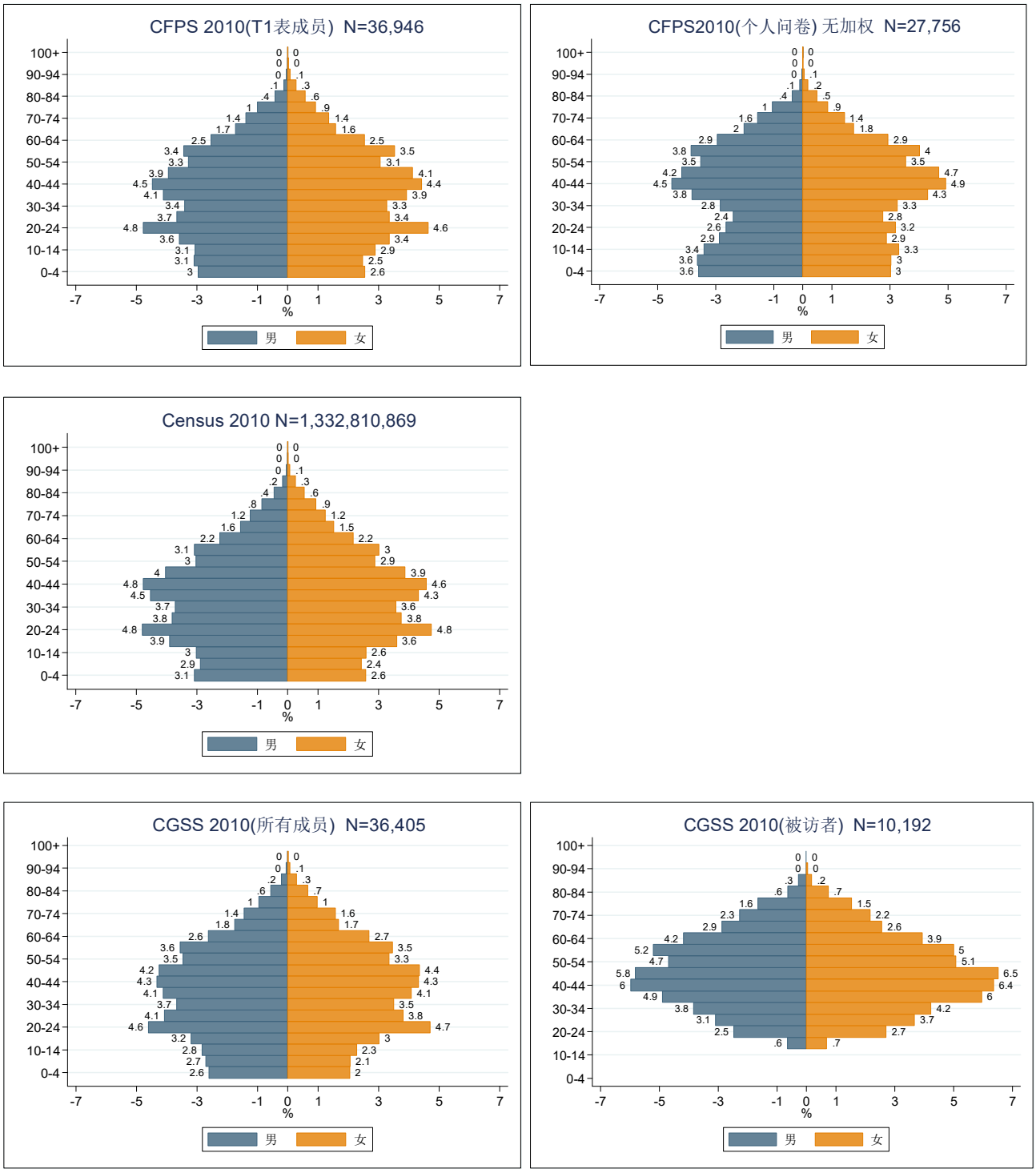


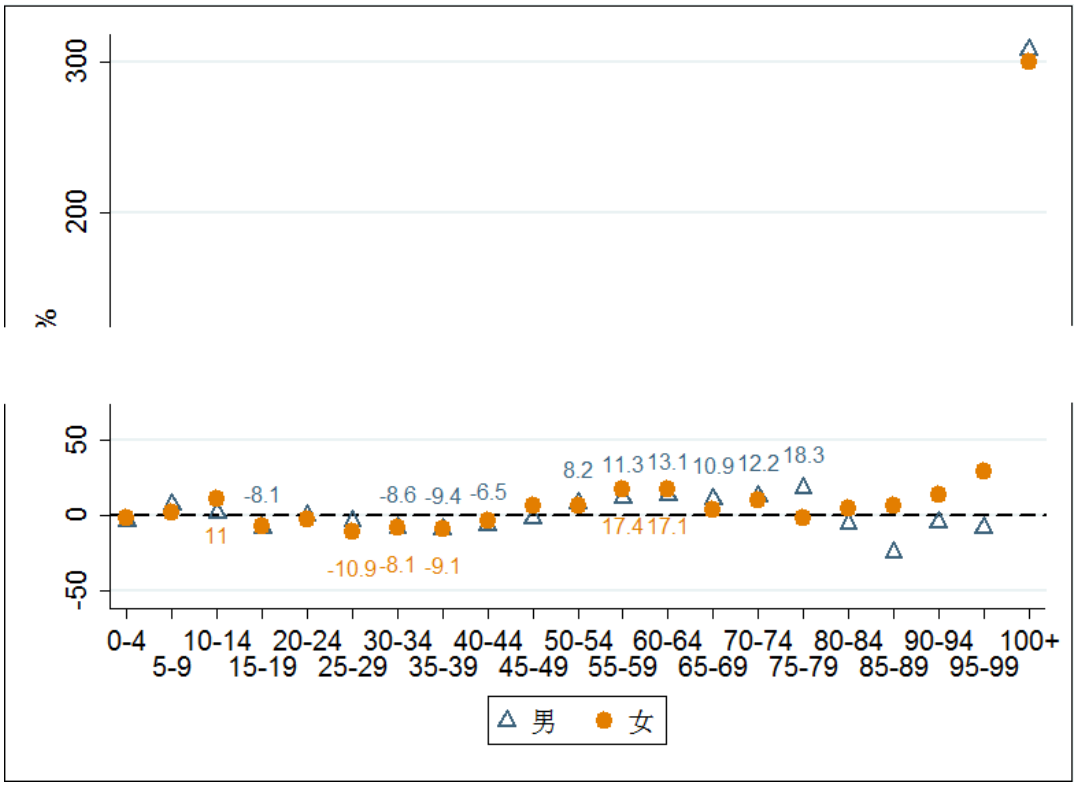
图 18. CFPS 2010 年基线调查、2010 年全国人口普查、CGSS 2010 的性别-年龄结构金字塔

[79] 第 8 部分中的 CFPS 数据使用的均是全国再抽样数据。

图 18 是根据 CFPS 2010 年基线调查再抽样数据、2010 年全国人口普查汇总数据[80] 和 2010 年中国综合社会调查（CGSS 2010）数据分别绘制的性别-年龄结构金字塔。在性别-年龄结构金字塔中，我们从 0 岁到 100 岁以上，以每 5 岁为一个年龄组，分别统计了各年龄组男性和女性人数在总人口（或总人数）中的比例（见图中标注的数字）。对 CFPS，我们既描述了受访家户中 T1 表成员（或所有同住家庭成员）的性别-年龄结构，也描述了所有已回答个人问卷（成人问卷和少儿问卷）的受访者的性别-年龄结构。根据调查设计，同住的家庭成员均是个人问卷的受访对象，但在实际的调查过程中，由于外出或拒访等原因，只有部分家庭成员填答了个人问卷。因此，基于 T1 表成员的性别-年龄结构反映的是对总人口进行概率抽样后样本人口的特征，而基于个人问卷样本的性别-年龄结构反映的是个人问卷受访者的特征；同样地，我们对 CGSS 也分别描述了受访家户中所有家庭成员的性别-年龄结构和最终受访者的性别-年龄结构，前者反映的是样本对总体的代表性，后者反映的是受访者的性别年龄构成。

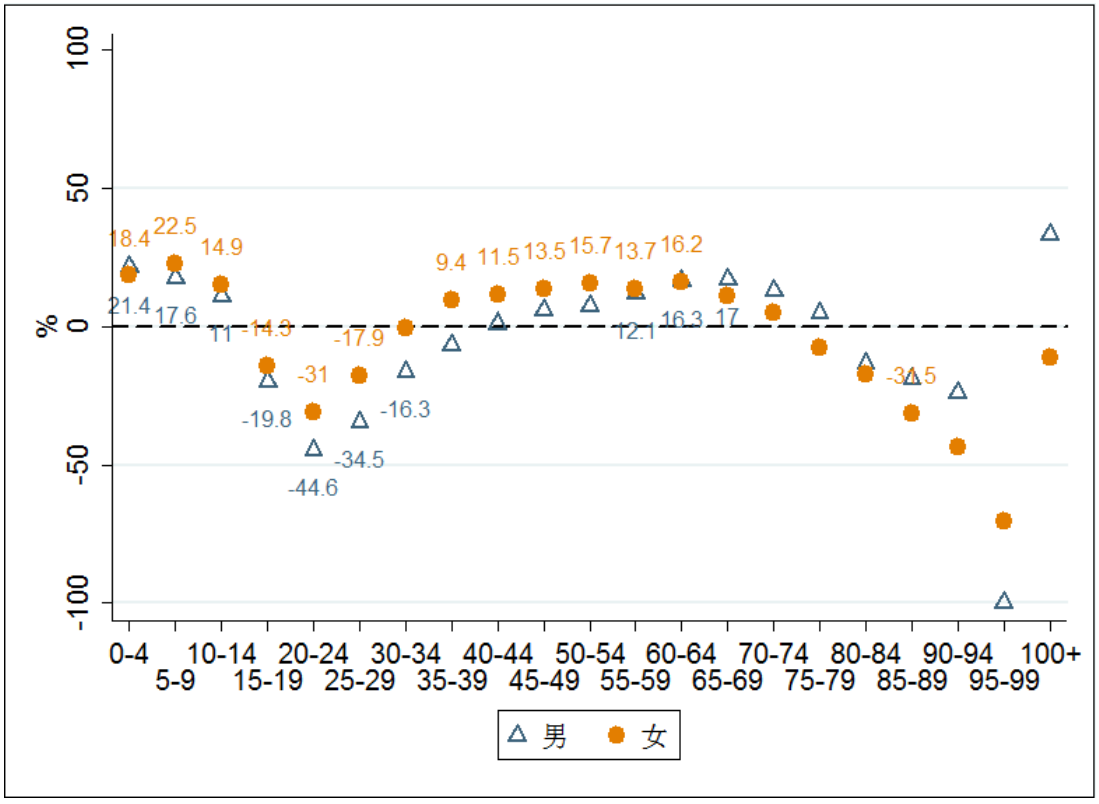
我们将 CFPS T1 表成员和 CGSS 所有家庭成员的性别-年龄结构分别与 2010 年普查的总人口性别-年龄结构比较后发现，两个调查数据的性别-年龄结构金字塔与普查数据的性别-年龄结构金字塔形状基本相似，均在 20-24 岁、40-44 岁年龄组人数较多，在高龄组和低龄组人数较少。从各性别年龄组人数占总人数的百分比上看，CFPS 的数值与普查数值更接近。为了检验三个数据中性别-年龄分布的差异是否显著，我们使用对数比率（log rate）模型，以普查数据各性别年龄组的频次作为暴露期（exposure），检验调查数据（CGSS 和 CFPS）中各性别年龄组的频次分布是否有别于普查的性别-年龄分布。该方法假定普查数据是总体，CFPS 和 CGSS 调查数据均是通过概率抽样从普查数据中获取的样本，如果样本结构如实反映了总体结构，则各性别年龄组的入样概率相等，不存在个别性别年龄组被多抽取或少抽取的情况。我们将加入性别、年龄及其交互项的完全模型与零模型比较，得到模型检验的卡方值。CFPS T1 表成员性别-年龄结构与 2010 年普查性别年龄结构相比较的卡方检验值为 253.70（自由度 41，p=0.0000），CGSS 所有家庭成员性别-年龄结构与 2010 年普查性别年龄结构相比较的卡方检验值为 529.85（自由度 41，p=0.0000）。两个调查数据与普查数据相比性别-年龄结构的差异均显著，不过 CFPS 较之 CGSS 与普查数据的性别-年龄结构更相近，卡方值较小。

[80] 数据取自 2010 年第六次全国人口普查数据光盘汇总表“T3-01 全国分年龄、性别的人口”。



注：图中标记数字的组别与普查数据的差异均在 0.01 水平上显著，未标记数字则表示差异不显著。

图 19.　CFPS 2010 T1 表成员各性别年龄组与 2010 年普查相比的偏误情况



注：图中标记数字的组别与普查数据的差异均在 0.01 水平上显著，未标记数字则表示差异不显著。

图 20.　CFPS 2010 个人问卷受访者各性别年龄组与 T1 表成员相比的偏误情况

图 19 展示了 CFPS T1 表成员各性别年龄组频次分布与普查数据相比的偏误情况，图中的三角和圆点分别代表男性和女性各年龄组与普查数据各年龄组相较偏误率的大小和方向。假定在没有偏误的情况下，偏误率为 0。大于 0 表示 CFPS 某性别年龄组抽样比例高于普查数据中该组别对应的比例，小于 0 表示 CFPS 某性别年龄组抽样比例低于普查数据中该组别对应的比例。对在 0.01 水平上显著的偏差，我们在图中标记了该组别偏误率的大小，不显著的偏差则不标记。我们看到，CFPS 对 10-14 岁的女性抽样比例相对偏高，对 15-19 岁的男性抽样比例偏低，对 25-44 岁的男性和女性抽样比例均相对偏低，对 50-79 岁的男性及 55-64 岁的女性抽样比例均相对偏高。但总体上看，CFPS T1 表成员的性别-年龄结构与 2010 年普查基本相符。

从个体受访者的性别-年龄结构上看（图 18），由于 CGSS 的受访对象限于 18 岁以上的家庭成员，且每一户只抽取一名受访者，因此，CGSS 在 0-14 岁各低龄组没有受访对象，40-60 岁之间的受访者比例较高。相比之下，CFPS 基线调查的个人问卷受访者的性别、年龄分布更广泛，在各年龄层均有受访样本。CFPS 在 30 岁以上各年龄组的分布较接近于普查数据，但在青年组（15-19 岁，20-24 岁，24-29 岁）的分布比例明显低于普查数据，这是由于该年龄段外出就学、务工的人数较多，造成个人问卷的样本流失较大。我们检验了 CFPS T1 表成员的性别-年龄结构和个人问卷受访者的性别-年龄结构的差异，以个人问卷受访者各性别年龄组的分布频次作为因变量，以 CFPS T1 成员各性别年龄组分布频次作为暴露期，使用对数比率模型，并将完全模型与零模型相比较，卡方值为 1059.86（自由度 41，p=0.0000），说明实际受访者的年龄-性别结构有偏于调查设计预期的受访者性别-年龄结构。图 20 展示了 CFPS 个人问卷受访者各性别年龄组与 T1 表成员各性别年龄组相比偏误率的大小和方向：0-14 岁各年龄组的男性和女性受访比例均相对偏高，15-29 岁的女性及 15-34 岁的男性受访比例均相对偏低，35-64 岁的女性及 55-69 岁的男性受访比例均相对偏高，85-89 岁的女性受访比例相对偏低。受访者性别年龄结构的偏差会影响统计推断的准确性，这一问题可以通过对单元无应答（unit nonresponse）调整加权的方式予以修正。我们在下文权数计算一节将简要介绍无应答调整权数的计算方法，需要了解更多详细信息的用户也可以参考技术报告《中国家庭追踪调查 2010 年基线调查权数计算（CFPS-17）》。

8.2 家庭规模和家庭户类别

在 CFPS 2010 全国再抽样样本中，根据 T1 表家庭成员计算的平均家庭规模为 3.8 人，

如果只包括 T1 表中在家的家庭成员，平均家庭规模为 3.3 人。与 2010 年人口普查相比，CFPS 的家庭规模更大，而且 T 检验的结果在 0.001 的水平上是显著的。分城乡来看，无论是以 T1 表全部家庭成员计算，还是以 T1 表在家的家庭成员计算，CFPS 的家庭规模都高于普查（p=0.001）。图 21 还列出了 CGSS 2010 调查的平均家庭规模，无论是就全国来看，还是分城乡来看，CGSS 与 CFPS 按照 T1 表成员计算出来的结果都非常接近。

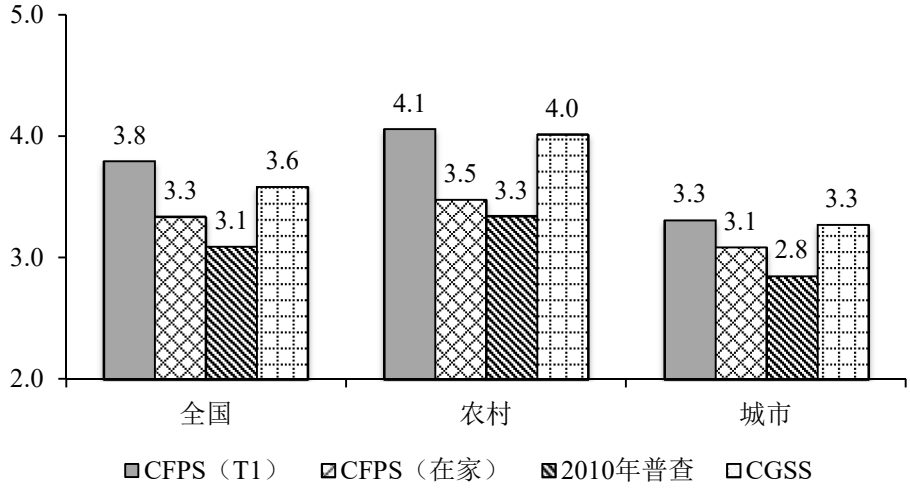


图 21. 分城乡平均家庭户规模

图 22 显示，在 CFPS 2010 的全国再抽样样本中，如果按照 T1 表家庭成员计算，一代户的比例为 20.2%，二代户的比例为 48.7%，三代及以上户的比例为 31.2%。如果只包括 T1 表中在家的家庭成员，一代户的比例为 29.3%，二代户的比例为 42.8%，三代及以上户的比例为 28.0%。与普查相比，CFPS 调查中三代及以上户的比例更高，而一代户的比例更低，而且卡方检验的结果在 0.001 的水平上显著。相比之下，CGSS 调查的家庭代数与 CFPS 根据 T1 表的家庭成员计算出来的结果比较接近。

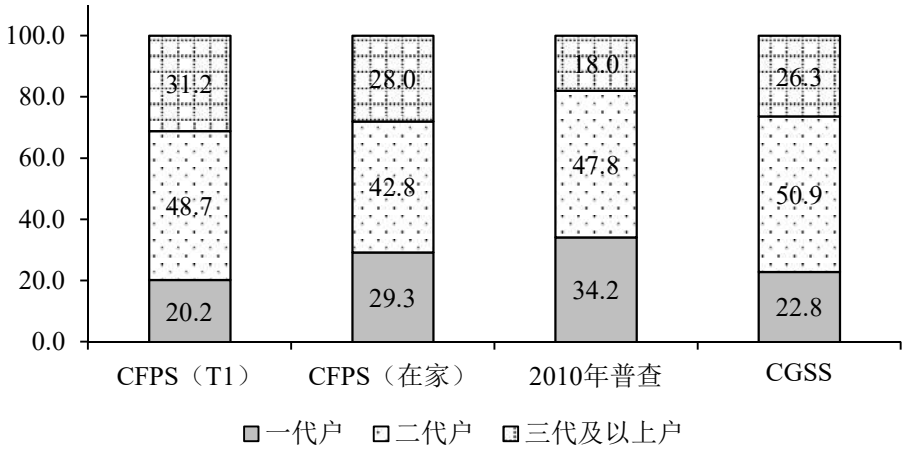


图 22. 全国不同规模家庭户类别

8.3 家庭收入

表 42 给出了 CGSS、CFPS 和 CHFS（中国家庭金融调查）这三个全国代表性样本 2010 或 2011 年的家庭收入均值及基尼系数。其中，CGSS 和 CFPS 的结果较为接近。对于城市家庭，由 CGSS 得出的基尼系数更高一些。然而，无论是收入均值还是基尼系数，由 CHFS 得出的结果都显著高于前两个数据集。其中，CHFS 显示城市样本收入均值为 87071 元，接近 CFPS 数值的两倍。无论是农村样本还是城市样本，由 CHFS 数据得出的基尼系数均超过 0.65。

表 42. CGSS、CFPS 和 CHFS 样本量、家庭收入均值和基尼系数

		CGSS 2010	CFPS 2010 (调整后)	CHFS 2011 (24 省)
农村	样本量	5313	5883	4162
	收入均值	22125.8	28826.4	32285.8
	基尼系数	0.495	0.498	0.675
城市	样本量	3849	3248	3619
	收入均值	53494.0	44917.6	87071.4
	基尼系数	0.535	0.470	0.655

注：CHFS 2011 数据包括全国 25 个省，比 CFPS 2010 多了青海省，少了福建省。为了提高和 CFPS 2010 的可比性，以上分析样本去除了青海省的数据。此外，此样本不包含所有收入为负数或零的家庭。

表 43 和表 44 分别显示了这三个样本中农村和城市家庭收入的具体分布。如前文所述，CFPS 2010 的收入包括调整前和调整后两个版本，调整前的收入是根据原始数据计算出来的，调整后的收入指的是考虑了农村家庭自留消费的农产品的价值以后计算出来的农村家庭总收入。[81] 从表 43 和表 44 可以发现，无论是农村还是城市，由 CFPS 和 CGSS 得出的收入分

[81] 详细的调整办法参见技术报告：CFPS-14。

布都非常类似。然而，CHFS 给出的收入分布呈现出两极分化。首先，在低收入部分（25%及以下分位数），CHFS 的结果明显低于其他两个样本，这一点在农村样本中尤甚。相反，在高收入部分（75%及以上分位数），CHFS 的结果却显著偏高，这一点在城市样本中更加突出。

表 45 和表 46 分别就农村和城市样本描述了这三个数据中处在不同分位数的家庭的收入占所有家庭总收入的比重。可以发现，在 50%分位数以下，CFPS 和 CGSS 的结果比较接近，而 CHFS 的结果相对较低。尤其在农村样本中，CHFS 显示收入较低的一半家庭的总收入只占所有家庭总收入的 10%。此外，在 5%和 10%分位数下，CHFS 的结果也显著低于其他两个数据集，这一点和表 42 和表 43 的结果是一致的。在较高分位数上（75%以上、90%以上、95%以上），CGSS 和 CFPS 的结果在农村样本上非常接近，城市样本中 CGSS 给出了相对更高的收入集中度。但是，无论是农村样本还是城市样本，CHFS 都给出了最高的收入集中度。尤其在农村样本中，CHFS 显示收入最高的 5%的人获得了 43.1%的总收入。

表 43. CFPS、CGSS 和 CHFS 农村家庭总收入分布（单位/元）

分位数	CGSS 2010	CFPS 2010 (未调整)	CFPS 2010 (调整后)	CHFS 2011 (24 省)
5%	2000	1940	2300	720
10%	3240	3300	4210	1750
25%	8000	9000	10000	5220
50%	15000	18000	20000	13250
75%	28880	32000	34205	32249
90%	45000	53600	56121	61025
95%	60000	74000	77628	93000

注：CHFS 2011 数据包括全国 25 个省，比 CFPS 2010 多了青海省，少了福建省。为了提高和 CFPS 2010 的可比性，以上分析样本去除了青海省的数据。此外，此样本不包含所有收入为负数或零的家庭。

表 44. CFPS、CGSS 和 CHFS 城市家庭总收入分布（单位/元）

分位数	CGSS 2010	CFPS 2010 (未调整)	CFPS 2010 (调整后)	CHFS 2011 (24 省)
5%	6000	5000	5000	1700
10%	10000	9640	10000	5170
25%	20000	17800	18000	19001
50%	30000	30200	30300	38450
75%	55000	55000	55000	78000
90%	100000	86000	86100	169000
95%	150000	120000	120000	262500

注：CHFS 2011 数据包括全国 25 个省，比 CFPS 2010 多了青海省，少了福建省。为了提高和 CFPS 2010 的可比性，以上分析样本去除了青海省的数据。此外，此样本不包含所有收入为负数或零的家庭。

表 45. CGSS、CFPS 和 CHFS 农村家庭不同分位数下的家庭收入占总收入的比重（单位：%）

分位数	CGSS 2010	CFPS 2010（未调整）	CFPS 2010（调整后）	CHFS 2011（24 省）
5%以下	0.3	0.2	0.2	0.0
10%以下	0.9	0.7	0.8	0.2
25%以下	5.5	4.1	4.6	2.0
50%以下	18.0	16.3	17.5	10.0
75%以上	59.6	61.2	59.7	72.5
90%以上	36.0	38.5	37.1	53.4
95%以上	24.1	27.1	25.9	43.1

表 46. CGSS、 CFPS 和 CHFS 城市家庭不同分位数下的家庭收入占总收入的比重（单位：%）

分位数	CGSS 2010	CFPS 2010（未调整）	CFPS 2010（调整后）	CHFS 2011（24 省）
5%以下	0.3	0.3	0.3	0.1
10%以下	1.4	1.1	1.4	0.4
25%以下	7.7	5.7	6.2	3.6
50%以下	16.9	19.0	19.1	14.0
75%以上	63.0	57.6	57.6	65.6
90%以上	39.2	35.0	35.0	43.9
95%以上	30.1	22.8	22.9	33.2

8.4 城乡分布

我们用 CFPS 2010 数据和 2010 年普查数据、CHFS 2011 数据进行了城乡分布的比较。由于 CGSS 采用了农村和城市分层抽样，对城市进行了过度抽样，因此不能和 CFPS 在这方面进行比较。

从表 47 可以看出，就国家统计局划分的农村-城镇分类而言，CFPS 的 T1 表与个人受访者的样本的频次分布非常接近，虽然二维列联表的卡方检验结果显著（$\chi^2(1) = 5.0157$，$p = 0.025$）。与 2010 年人口普查数据中城乡分布基本持平（居住在城镇地区的比例高出农村地区 0.6%）的格局相比，CFPS 数据中的居住分布呈现为农村比例高出城镇近 10%。以农村为参照组，以普查数据为暴露期的对数比率模型分析（$\chi^2(1)=462.83$，$p=0.000$），以及二维列联表的卡方检验结果（$\chi^2(1)=462.03$，$p=0.000$）均显示 CFPS 的农村-城镇分布与普查数据

存在显著差异。究其原因，既可能是 CFPS 的抽样、受访者回答率等有偏差，也可能是我们笼统地将普查数据中的“镇”划归入“城镇”，而忽略了组成“镇”的行政区划、地理单元中可能包含农村。在缺少更微观的普查数据的情况下，我们暂时无法更进一步地探求其中的原因。相对而言，CHFS 数据的城乡样本分布比例更接近于普查数据，但两者之间依然存在显著差异（$\chi^2(1)\approx 150000$，$p = 0.000$）。

就村委会-居委会的划分而言，CFPS 的 T1 表与个人受访者的样本的频次分布非常接近，虽然二维列联表的卡方检验结果依然显著（$\chi^2(1) = 10.8773$，$p = 0.001$）。CFPS 中大约 70% 的受访者所居住的社区隶属于村委会。

CFPS 数据中的农业-非农业户口分布与普查的数据相当接近，基本呈现出 7 成农业户口，3 成非农业户口的分布特点。二维列联表的卡方检验结果虽然显著，但是卡方值仅约为 96.6。CHFS 数据在频次分布上也比较接近于普查数据，但也与普查数据存在着显著差异（$\chi2(1)\approx150000$，$p = 0.000$）。

表 47. 全国分城乡人口分布（单位：%）

	CFPS 2010		2010 年人口普查	CHFS 2011[b]
	T1 表	个人问卷受访者		
国家统计局划分				
农村	55.4	54.5	49.7	48.7
城镇[a]	44.7	45.5	50.3	51.3
样本量	36571	27444	1332810869	528808705
村/居委会划分				
村委会	69.6	68.4	—	—
居委会	30.4	31.6	—	—
样本量	36571	27444	—	—
户口				
农业	—	73.6	70.9	63.3
非农业	—	26.4	29.1	36.7
样本量	—	27204	1319046434	440104614

注：[a] 本表中将 2010 年人口普查数据中的“城市”与“镇”这两个类别合并为“城镇”。

[b] CHFS 数据经过了加权。

8.5 受教育程度

表 48 显示，CFPS 2010 年基线调查的 T1 表中受访者的最高学历分布与 2010 年人口普查数据的分布非常相近。两者之间相对较明显的差异表现为 T1 表中初中学历的比例较普查数据中的比例略低（大约 5%）。除此以外，其他各学历水平的比例差异基本不超过 2%。二维列联表的卡方检验结果显著（$\chi^2(6)=493.1084$，$p=0.000$）。以“文盲/半文盲”为参照组，以普查数据为暴露期的对数比率模型分析结果显示，显著的差异来源于初中、高中、大学专科以及研究生这几个组。

表 48. 6 岁及以上人口中已完成的最高学历（单位：%）

	CFPS 2010		2010 年人口普查	CGSS 2010
	T1 表成员	个人问卷受访者		
文盲/半文盲	24.31	29.7	22.9	22.7
小学	21.54	23.6	19.9	19.0
初中	32.05	26.2	37.6	26.3
高中	14.35	13.2	11.8	18.3
大学专科	4.65	4.3	4.7	7.7
大学本科	2.93	2.9	2.8	5.5
研究生	0.2	0.2	0.3	0.6
样本量	29974	23219	111601269	10173

我们同时用 CFPS 与 CGSS 进行了比较，发现 CFPS T1 表中的受访者最高学历为初中的比例偏高（大约 6%），高中、大学专科、本科、研究生的比例偏低。为了评估 CFPS 数据的质量，我们对这一差异进行了统计检验：将 CGSS 与普查数据做对比，二维列联表的卡方检验结果显著，且卡方值约为 1200，远远高于 CFPS 与普查数据对比得到的卡方值；对数比率模型的分析结果显示，除小学组以外，CGSS 中的其他各学历组相对于参照组的差异均显著。所以，我们推测存在差异的来源可能是 CGSS 数据中高学历者被过度抽样，而低学历群体的样本代表性不足。

与 T1 表相比，CFPS 中有个人问卷者的学历水平较低。产生这一差异的原因最主要是由于个人问卷受访者的样本选择性：学历水平较低的家庭成员更容易留在家中，因此个人问卷采访到这些人的可能性更大；而学历水平较高的家庭成员外出的可能性更大，很多人没有接受访问并完成个人问卷。产生差异的另外一个较次要的原因是 T1 表的代答者倾向于高报其家人的学历水平。[82]

[82] 更详细的分析参加技术报告：CFPS-21。

8.6 婚姻状态

我们同样比较了 CFPS、CHFS、CGSS 和 2010 年人口普查数据的 15 岁以上人口的婚姻状况分布，以进一步检查 CFPS 婚姻数据的质量。表 49 列举了这一系列比较的结果。CFPS 分 T1 表家庭成员和个人问卷受访者两个群体来描述， CHFS 数据测量了所有家庭成员的婚姻状态。CGSS 数据仅包括 18 岁以上受访者的婚姻状态。数据显示，CFPS T1 表成员的婚姻状态分布与普查数据非常接近，二者的分布几乎完全一致。不过，普查数据超大的样本规模导致了二维列联表的卡方检验（$\chi^2(3)=23.5398$，$p=0.000$）和以普查数据为暴露期的对数比率模型分析结果都呈现出显著差异（$\chi^2(3)=24.15$, $p=0.000$）。

由于 CFPS 个人问卷受访者中年轻人口分布比例明显低于普查数据，个人问卷受访者的婚姻分布状态与普查数据存在差异（$\chi^2(3)=669.7521$，$p=0.000$），主要表现为未婚人口比例低于普查数据，但离异、丧偶人口比例与普查数据没有差异。这表明 CFPS 的个人问卷受访者数据除了年轻人口存在抽样偏差外，没有其他的偏差来源。

CHFS 数据的婚姻状态分布也较为接近于普查数据，但其与普查数据的差异大于 CFPS T1 表成员数据的分布（$\chi^2(3)=157.7124$，$p=0.000$）。CGSS 数据因设计不同，受访者限定为 18 岁以上的家庭成员，且城市人口过度抽样，因此，其分布与其他数据显著不同。

分性别来看，CFPS T1 表家庭成员数据的婚姻分布同样与普查数据一致，男性中未婚人口比例略高于普查数据，但女性的婚姻状态与普查数据几乎完全一致。

表 49. 15 岁及以上人口婚姻状态分布（单位：%）

		CFPS 2010		2010 年人口普查	CHFS 2011[a]	CGSS 2010[b]
		T1 表成员	个人问卷受访者			
全国	未婚	20.8	14.6	21.6	18.2	8.1
	有配偶	72.2	78.5	71.3	76.5	82.8
	离异	1.2	1.2	1.4	1.3	2.1
	丧偶	5.8	5.8	5.7	4.0	7.0
	样本量	30642	22197	105542243	24693	10154
男性	未婚	24.1	17.0	21.6	21.1	10.1
	有配偶	71.2	78.2	71.3	75.3	83.6
	离异	1.4	1.4	1.4	1.2	2.1
	丧偶	3.3	3.4	5.7	2.3	4.1
	样本量	15454	10732	52943450	12352	4932

女性	未婚	17.3	12.3	18.5	15.2	6.3
	有配偶	73.2	78.8	72.3	77.7	82.0
	离异	1.0	0.9	1.2	1.4	2.0
	丧偶	8.5	8.0	8.0	5.7	9.7
	样本量	15188	11465	52598793	12341	5222

注：[a] CHFS 2011 数据为加权后结果。

[b] CGSS 2010 数据为 18 岁以上人口的婚姻状态分布。

9. 权数计算

9.1 基线权数

CFPS 2010 对全国完全样本和全国再抽样样本分别计算了家庭问卷、成人问卷和少儿问卷三个数据库的权数。其中，全国完全样本的权数为五个“大省”（上海市、河南省、甘肃省、辽宁省和广东省）和一个“小省”（25 省市中的其他省市）共 6 个子总体的全部样本的权数的合并；全国再抽样样本的权数则为五个“大省”经再抽样后的样本的权数与“小省”权数的合并。权数的计算包括抽样设计权数、无回答调整权数、事后分层调整权数的计算以及对权数的极值调整。

抽样设计权数为第一阶段抽样、第二阶段抽样、第三阶段抽样和第三阶段调整抽样[83]概率的乘积的倒数。对于再抽样样本，在计算过程中还考虑了从第一阶段所抽取的样本区县中再抽取再抽样样本区县的概率。

无回答调整权数的计算在家庭成员问卷层面和家庭/个人问卷层面进行。家庭成员问卷层面的无回答调整采用了加权组调整的方法，使用家庭成员问卷完成的数量占村居样本中所有家庭样本数量的比例作为加权组调整系数。在家庭/个人问卷层面，家庭问卷的无回答调整权数同样采用了加权组调整的方法，调整系数为完成家庭成员问卷的家户中家庭问卷完成作答的家户数与需作答的家户数的比例。个人问卷的无回答调整采用两阶段的基于 logistic 模型的应答倾向概率作为调整系数：第一阶段将个人样本分为联系样本和无联系样本，利用数据中的辅助信息建立 logistic 回归模型，得到个人问卷联系层次的应答倾向概率；第二阶段将联系上的样本分为拒访样本和非拒访样本，建立 logistic 回归模型，得到个人问卷联系上的样本中拒绝访问的应答倾向概率。

事后分层调整（post-stratification）主要针对由于抽样设计的复杂性、实地调查过程中问题的多样性以及样本无回答的存在而导致的样本结构性偏差，用以减小抽样误差、提高估计精度。CFPS 采用性别、年龄、城乡三个变量对成人和少儿样本数据进行完全事后分层调整。

对权数的极值调整首先需要将无回答调整、事后分层调整过程中每一次的调整系数都控制在一定的范围内，以控制无回答和事后分层调整所带来的权数的方差；其次还需要通过调

[83] 第三阶段调整抽样针对同一个抽样地址下有多个满足条件的家户的末端抽样框误差。CFPS 采用了随机抽取一户的方法对该误差进行调整。

整将经抽样设计、无应答、事后分层三阶段调整后得到的最终权数控制在一定的范围内，以保证估计效率。

最后，为了使最终的权数和等于总人口数，还需要进行再次调整。通过这些调整，最终得到 25 省市家庭、成人和少儿完全样本的权数和再抽样样本的权数[84]。其中 25 省市完全样本的权数为“大省”和“小省”的权数的合并。权数的使用与用户需要使用的总体以及数据库类型有关，不同数据库类型及样本所代表的总体的基本情况可参见表 27。表 50 是全部数据库加权变量的简单列表：

表 50. CFPS 权数变量名及变量标签

数据库	变量名	变量标签
CFPS 2010		
家庭问卷数据库	fswt_nat	家庭权重-全国完全样本
家庭问卷数据库	fswt_res	家庭权重-全国再抽样样本
成人/少儿问卷数据库	rswt_nat	个人权重-全国完全样本
成人/少儿问卷数据库	rswt_res	个人权重-全国再抽样样本
CFPS 2012		
家庭问卷数据库	fswt_natcs12	家庭横截面权数:全国总样本
家庭问卷数据库	fswt_rescs12	家庭横截面权数:全国再抽样样本
家庭问卷数据库	fswt_natpn1012	家庭面板权数:全国总样本
家庭问卷数据库	fswt_restpn1012	家庭面板权数:全国再抽样样本
成人/少儿问卷数据库	rswt_natcs12	个人横截面权数:全国总样本
成人/少儿问卷数据库	rswt_rescs12	个人横截面权数:全国再抽样样本
成人/少儿问卷数据库	rswt_natpn1012	个人面板权数:全国总样本
成人/少儿问卷数据库	rswt_restpn1012	个人面板权数:全国再抽样样本
CFPS 2014		
家庭问卷数据库	fswt_natcs14	家庭横截面权数:全国总样本
家庭问卷数据库	fswt_rescs14	家庭横截面权数:全国再抽样样本
家庭问卷数据库	fswt_natpn1014	家庭面板权数:全国总样本
家庭问卷数据库	fswt_restpn1014	家庭面板权数:全国再抽样样本
成人/少儿问卷数据库	rswt_natcs14	个人横截面权数:全国总样本
成人/少儿问卷数据库	rswt_rescs14	个人横截面权数:全国再抽样样本
成人/少儿问卷数据库	rswt_natpn1014	个人面板权数:全国总样本
成人/少儿问卷数据库	rswt_restpn1014	个人面板权数:全国再抽样样本

CFPS 2010 权数的具体计算方法及加权结果统计可参考技术报告《中国家庭追踪调查

[84] 2010 年成人和少儿的权数是分开计算的，2012 年开始将成人和少儿当成一个整体计算权数。

2010 年基线调查权数计算（CFPS-17）》。

9.2 追踪权数

我们在前文已经介绍过，CFPS 仅对基因成员及其所在家庭进行追踪。随着时间的变迁，新的基因成员出生，已有基因成员死亡；新家庭因为基因成员的婚姻、分家等原因而不断产生，旧家庭因为基因成员的死亡或另组新的家庭而不断分化和消失。这些变化导致样本框发生改变。同时，由于调查中不可避免的存在样本流失和样本无应答的情况，调查总体和样本总体也都发生了变化。基于这些变化，若要使样本对总体仍旧有较好的代表性，需要对追踪数据（或称面板权数）进行加权调整，以便进行有效的统计推断。

CFPS 目前的推断主要是基于基因成员，因此权数调整主要针对基因成员进行，包括追踪年的追踪权数和追踪年的截面权数两大部分。其中追踪年的追踪权数仅对 2010 年的初始基因成员进行加权调整，追踪年的截面权数包含 2010 年的基因成员和之后新进基因成员两部分。追踪权数和截面权数的权数调整均包含个人权数调整和家庭权数调整。CFPS 的总样本包括六个子总体，我们对各个样本框分别进行权数调整。下面我们将主要介绍对全国完全样本的个人和家庭的权数调整，其它样本框加权调整的方法与之基本类似。

9.2.1 个人追踪权数调整

CFPS 个人追踪权数调整包含 2010 年基因成员个人初始追踪权数、新进基因成员的追踪权数、无应答调整权数、事后分层调整权数和极值调整权数。

CFPS 追踪调查的追踪对象是 2010 年完成家庭成员问卷的基因成员以及新进的基因成员（即 2010 年基因成员之后新生/领养的血缘子女），因此 2010 年基因成员的个人追踪初始权数是 CFPS 2010 家庭成员层面的无回答权数。需要注意的是，2012 年由于特殊原因没有调查四川某区县，该县在 CFPS 2012 的权数调整中视为区县层面的无回答，在计算个人追踪权数时将其从 2010 年的数据中去除，然后对 CFPS 2010 的家庭成员层面的无回答权数进行四川省内的区县层面的无回答调整，以此作为 CFPS 2012 的个人追踪初始权数。

对于新进基因成员，我们用其父母的个人追踪权数的均值作为他们的追踪权数。若仅有母亲（或父亲）一方是基因成员，则新进基因成员的追踪权数为其母亲（或父亲）的个人追踪权数。

为了提高权数的精度，各追踪年个人追踪权数的无应答调整需要利用 2010 年家庭成员问卷和个人问卷，以及各追踪年的问卷的信息，采用基于 logistic 模型的倾向权数计算方法，

得到个人层次的无应答调整系数。具体来说，即将追踪年的个人样本分为个人追踪完成样本和个人追踪未完成样本，利用数据中的辅助信息（年龄、年龄的平方、性别、家中人口数、家中是否有老人、家中是否有小孩、城乡代码、代际码、房屋所有情况和历年样本完成情况等），建立 logistic 回归模型，得到个人问卷联系层次的倾向回答调整系数。在 CFPS 的追踪调查中，成人和少儿被视为一体进行个人追踪权数的调整。需要注意，由于 CFPS 包含 5 个“大省”样本框和 1 个包括其余 20 个省的“小省”样本框，因此需要分别在六个抽样框中建立 logistic 模型，进而得到各追踪年家庭成员的无应答调整权数。

CFPS 抽样设计和实地调查过程的复杂性以及样本的无应答和流失导致某些关键变量存在样本结构性偏差，影响估计量的准确性。为了调整这些结构性偏差，减小抽样误差，提高估计精度，需要对个人样本数据进行事后分层调整。在个人问卷层面，性别、年龄、城乡是非常重要的指标，因此，在六个抽样框以及全国完全样本的成人和少儿数据中，我们用城乡（分为城镇和农村）、性别（分为男和女）、年龄（分为 16-19 岁、20-29 岁、30-39 岁、40-49 岁、50-59 岁、60-69 岁、70-79 岁、80 岁以上，共 8 类）变量进行事后分层调整。事后分层调整使用的是最新可获得的官方人口数据，例如 2010 和 2012 年使用的是第六次全国人口普查数据，2014 年使用的是 2014 年人口抽样调查数据。对 CFPS 个人问卷存在的极少量年龄、性别的缺失，我们采用均值和中位数插补方法对其进行插补。由上，我们得到追踪年各个抽样框的各层内的事后分层调整的系数。

为防止个人数据库因权数过大或过小而导致的方差较大、估计效率降低的问题，我们采用极值权数处理的方法，用权数分布的 5%和 95%的分位数作为最小值和最大值的极值点，对上述个人层面的权数进行极值调整，据此得到该部分的调整极值系数，并根据极值权数调整的情况进行事后分层调整。

由此，上述涉及个人调整部分的权数的乘积即为最终的个人权数。其中个人追踪权数不需要考虑新进基因成员的个人追踪权数，截面权数需要考虑新进基因成员的个人追踪权数。

9.2.2 家庭追踪权数调整

CFPS 家庭追踪权数调整包含家庭初始追踪权数、无应答调整权数、事后分层调整权数和极值调整权数。

CFPS 追踪年各家庭（包含分裂家庭）的初始追踪权数是追踪年基因成员的无应答调整权数的均值。

由于并非所有的家庭都完成了追踪年的家庭问卷，对没有完成问卷的含基因成员的家庭同样需要进行无应答权数调整，我们利用家庭样本的完成情况，基于 AAPOR 的回答率 RR1，使用区县层面的加权组调整方法，得到家庭的无应答调整系数。

为防止因家庭层面权数过大或过小而导致的方差较大、估计效率降低的问题，我们采用极值权数处理的方法（trimming），用权数分布的 5%和 95%的分位数作为最小值和最大值的极值点，对上述家庭层面的权数进行极值调整，得到该部分的调整极值系数。

由于对调查数据的加权调整要求最终权数的和等于总体权数，而通过极值调整后的权数的和不再与总体相等，因此需要对上述权数进行再次调整。此处，我们简单的将各个总体视为均匀总体，进行校准调整，使调整后的权数和与总体的权数相同。

由此，将上述家庭层面的调整权数相乘，其乘积即为最终的家庭追踪权数。

10. 技术报告系列

除此用户手册外，我们还建立了一个技术报告系列，包含与 CFPS 项目各个主题相关的一些文章，可以帮助用户更深层次地了解 CFPS 调查与数据。目前该技术报告系列的内容如下，以后我们还将结合项目的进展情况继续增加新的内容。

CFPS-1：《中国家庭追踪调查抽样设计》，谢宇、邱泽奇、吕萍，2012

CFPS-2：《中国家庭追踪调查 2010 年基线调查末端抽样框制作》，丁华，2012

CFPS-3：《中国家庭追踪调查 2010 年基线调查执行报告》，丁华，2012

CFPS-4：《中国家庭追踪调查 2010 年基线调查质量督导报告》，严洁、孙翊、滕学亮、任莉颖、孙妍，2012

CFPS-5：《中国家庭追踪调查 2010 年基线调查样本联系情况》，孙妍，2012

CFPS-6：《中国家庭追踪调查 2010 年家庭关系原始数据库的分解与匹配》，孙玉环、谢宇、胡婧炜、张春泥、许琪、黄国英，2012

CFPS-7：《中国家庭追踪调查 2010 年家庭关系数据库清理》，许琪、张春泥、孙玉环、胡婧炜、吕萍，2012

CFPS-8：《中国家庭追踪调查 2010 年职业行业编码》，任莉颖、李力、马超，2012

CFPS-9：《中国家庭追踪调查职业期望编码》，谢宇、李汪洋、马超、黄国英、柳皑然，2012

CFPS-10：《中国家庭追踪调查 2010 年职业社会经济地位测量指标构建（CFPS-10）》，黄国英、谢宇，2012

CFPS-11：《中国家庭追踪调查 2010 年综合变量（1）：字词与数学测试》，徐宏伟、骆为祥，2012

CFPS-12：《中国家庭追踪调查 2010 年综合变量（2）：受教育水平&抑郁量表》，谢宇、许琪、张春泥、徐宏伟，2012

CFPS-13：《中国家庭追踪调查 2010 年综合变量（3）：年龄、婚姻最佳变量》，张春泥、许

琪、孙妍，2012

CFPS-14：《中国家庭追踪调查 2010 年农村家庭收入的调整办法》，谢宇、张春泥、黄国英、许琪、徐宏伟，2012

CFPS-15：《中国家庭追踪调查 2010 年收入、消费支出数据整理》，沈艳、雷晓燕，2012

CFPS-16：《CGSS、CHIP、CHFS 和 CFPS 收入比较》，许琪、张春泥、周翔、谢宇，2012

CFPS-17：《中国家庭追踪调查 2010 年基线调查权数计算》，吕萍、谢宇，2012

CFPS-18：《中国家庭追踪调查 2010 年基线调查样本维护》，吕萍，2012

CFPS-19：《CFPS、CGSS、CHIP、CHFS 贫困率比较》，张春泥、许琪、周翔、张晓波、谢宇，2012

CFPS-21：《中国家庭追踪调查 2010 年受教育程度变量收集、清理与评估》，待发布

CFPS-22：《中国家庭追踪调查 2010 年综合变量（4）：父母社会地位》，张春泥、叶华、戴利红、胡婧炜、谢宇，2013

CFPS-23：《中国家庭追踪调查区县数据库模糊方法》，崔雅红、吴琼、徐宏伟、王广州，2014

CFPS-24：《中国家庭追踪调查 2012 年综合变量：生育子女数量和子女具体信息》，穆峥、谢宇，2014

CFPS-25：《中国家庭追踪调查 2012 年数据库介绍及数据清理报告》，吴琼、戴利红、崔雅红、张文佳，2014

CFPS-26：《中国家庭追踪调查 2012 年心理健康量表》，骆为祥、武玲蔚，2014

CFPS-27：《中国家庭追踪调查 2012 年家庭收入的调整办法》，许琪、张春泥，2014

CFPS-28：《中国家庭追踪调查方言编码》，武玲蔚、张文佳，2014

CFPS-29：《中国家庭追踪调查 2012 年和 2010 年财产数据技术报告》，靳永爱、谢宇，2014

CFPS-30：《中国家庭追踪调查 2012 年当前主要职业综合变量的建构》，李汪洋、胡婧炜、谢宇、吴琼，2014

CFPS-31：《中国家庭追踪调查 2012 年数列测试题》，徐宏伟、谢宇，2015

CFPS-33：《中国家庭追踪调查 2012 年家庭成员库的分解与家庭关系库的构建》，戴利红、孙妍、许琪、吴琼，2015

CFPS-34：《中国家庭追踪调查 2014 年数据库介绍及数据清理报告》，吴琼、戴利红、张聪、王玉磊、张文佳，2016

CFPS-35：《中国家庭追踪调查2016年数据库介绍及数据清理报告》，吴琼、戴利红、甄祺、张靖申、张聪、赵方圆，2018

CFPS-36：《中国家庭追踪调查2010年教育程度相关变量清理与评估》，胡婧炜 、黄国英、 张婧申 、崔雅红 、李汪洋、 程成 、吴琼、 谢宇 ，2019

CFPS-37：《中国家庭追踪调查 家庭社会经济地位综合变量：父亲和母亲的 出生年、最高学历、政治面貌和主要职业》，张春泥、叶华、李汪洋、马超、戴利红、胡婧炜、王祎睿、谢宇，2020

CFPS-38:《中国家庭追踪调查事件历史日历记录法（EHC）设计方案》，孙妍，2020

11. 参考书目

MCKINLEY, T. and K. GRIFFIN. "The Distribution of Land in Rural China." *Journal of Peasant Studies* 21, no. 1 (1993): 71-84.

Xie, Yu. "Evidence-Based Research on China: A Historical Imperative." *Chinese Sociological Review* 44, no. 1 (2011): 14.

Xie, Yu and Jingwei Hu. "An Introduction to the China Family Panel Studies (CFPS)." *Chinese Sociological Review* 47, no. 1 (2014): 3-29.

Xie, Yu and Ping Lu. "The Sampling Design of the China Family Panel Studies (CFPS)." *Chinese Journal of Sociology* 1, no. 4 (2015): 471-484.

北京大学中国社会科学调查中心，2009，《中国报告·2009·民生》，北京：北京大学出版社。

北京大学中国社会科学调查中心，2010，《中国报告·民生·2010》，北京：北京大学出版社。

北京大学中国社会科学调查中心，2011，《中国报告·民生·2011》，北京：北京大学出版社。

任强、谢宇，2011，“对纵贯数据统计分析的认识”，《人口研究》第 6 期。

孙妍、严洁、丁华、顾佳峰、刘月、姚佳慧、邹艳辉，2011，《中国家庭动态跟踪调查（2010）访员培训手册》，北京：北京大学出版社。

谢宇，2010，“认识中国的不平等”，《社会》第 3 期。

谢宇、董慕达，2011，“天地之间：东汉官员的双重责任”，《社会》第 4 期。

谢宇，2012，《社会学方法与定量研究》（第二版），北京：社会科学文献出版社。

谢宇、胡婧炜、张春泥，2014，“中国家庭追踪调查：理念与实践”，《社会》第 2 期。

谢宇、张晓波、李建新、于学军、任强，2014，《中国民生发展报告 2014》，北京：北京大学出版社。

谢宇、张晓波、李建新、涂平、任强，2016，《中国民生发展报告 2016》，北京：北京大学出版社。