China Family Panel Studies

CFPS 中国家庭追踪调查

技术报告系列：CFPS-39

系列编辑：谢宇　责任编辑：赵逸文

中国家庭追踪调查
2018 年数据库介绍及数据清理报告

吴琼 戴利红 甄祺 谷丽萍 王祎睿 吕萍

2020.12

（2021.8 更新）

中国家庭追踪调查 2018 年数据库介绍及数据清理报告

一．CFPS2018 总体介绍

1.1 执行总体情况

第五轮全国调查（CFPS2018）于 2018 年 6 月 5 日开始，2019 年 3 月结束。CFPS2018 的电访数据采集工作于 2018 年 6 月 5 日启动，持续至 2019 年 5 月。这一轮的全国追踪调查以 2010-2016 年全国调查所界定出来的家庭为基础，发放样本不仅包括 2016 年完访的所有家庭，还包括 2010-2014 年任一轮次完访、但 2016 年并未成功追踪的家庭。CFPS2018 调查全部结束时，我们共完成约 15,000 个家庭的访问，采集个人问卷约 44,000 份。[1]在调查执行过程中，近 400 名实地和电话访员共同完成了数据采集任务，电访完成的问卷数占所有问卷的 22%。

由于人口在不同地域间的流动，CFPS 样本分散程度进一步扩大。CFPS 基线调查时样本集中分布于 161 个区县和 649 个村居。到了 2018 年成功联系上的样本已分布在 900 多个区县的近 3000 多个村居，这使得联系上受访者变得越来越困难。此外，随着调查轮次的增加，部分受访者配合度下降，退出率有所上升。尽管面临着诸多挑战，CFPS 的独特设计使得我们能维持较为理想的追踪率。一是 CFPS 对于所有从原家庭流动出去的家庭成员，只要他们满足追踪的条件，我们会在全国范围内进行数据采集。二是我们从 2012 年就开始逐步引入电话访问和家庭成员代答的模式，在受访者自身面访不可实施的情况下，采用电话访问的方式进行补救。如果电话也无法触及受访者本人，我们会基于从其家庭的其他成员处采集一些基本信息（代答问卷）进行补充。三是针对那些以往调查没有追踪成功的受访者，我们会在后续的追踪调查中再次尝试。完访数据显示，不少上一轮没有成功的受访者样本在本轮调查中完成访问了。以完成家庭成员问卷为家庭层面完访标准，CFPS2018 家庭层面截面完访率为 69.3%，跨轮追踪率为 86.6%。个人样本的截面应答率为 67.4%，跨轮追踪率为 80.8%。若仅关注基线基因成员，2018 年的完访率为 64.5%。与同期执行的英国家庭追踪调查（UKHLS）相比，CFPS2018 第五轮追踪的应答率仍具备国际水平。

[1] 问卷统计数目来自 CFPS2018 第一轮正式版数据，原始采集的总问卷数目多于最终发布的问卷数。对于后期产生有效自答问卷的样本，我们发布版中不包含其代答问卷。

1.2 问卷设计调整

从问卷结构上看，2018 年 CFPS 进行了一些调整。一是将以前相互独立的成人问卷（针对 15 岁以上的个人）和少儿问卷（针对 0-15 岁个人）重新整合，形成一套整合型的个人自答问卷（针对 10 岁及以上个人）和少儿家长代答问卷（针对 0-15 岁个人）。个人自答问卷通过年龄和其他相关的筛选变量进行跳转控制。二是 2018 年首次在问卷设计中引入了家庭代答的概念，它针对所有离家单元，无论其经济上是否与原家庭独立，CFPS 都会要求原家庭成员对于离家单元的每个个体提供一些基础问题的代答，我们将这类由于“离家”产生的代答称之为“家庭代答”。而在往期调查中，这类代答只针对经济上有联系（也即与原家庭同处一个家户）的外出个体。在 2018 年调查中，除了离家单元的代答，我们还会由经济上同处一家的回答人对于那些由于身体原因无能力回答问卷的个体提供代答，我们将其称之为“个人代答”。“家庭代答”和“个人代答”的问卷产生原因不同，但问卷内容一样，主要采集个人的基本信息，与自答问卷相比内容有大幅度消减。

从问卷内容上看，CFPS2018 秉持着针对相同的测量目标与之前设计保持一致的基本原则，添加了一些新的测量。新的模块包括第一份工作（GD）模块，采集了第一份工作的单位、工作内容、求职渠道以及所需的教育程度。新的量表包括大五人格问卷、以及针对青少年的偏差行为的测量，这些量表的基本结构以及推荐用法可以参照第三部分的“个人库”介绍。除此之外，2018 年问卷设计时还在个人问卷上添加了个人的捐款行为、个人金融素养评估，在成员问卷上针对家庭成员基本信息（出生年、性别）中存疑的部分进行了重新采集。其他一些设计上的细微变动用户可以通过项目网站上“调查问卷”页面的 2010-2018 历年问卷汇总表查看。

1.3 数据结构变化

与问卷设计调整相应， CFPS2018 的数据结构也发生了一些变化。一是取消之前成人和少儿库的区分，个人层面包含两个数据库：10 岁及以上个人问卷库(person 库)，0-15 岁少儿的家长代答问卷库(childproxy 库)。其中 10 岁及以上的个人问卷包括个人自答以及代答数据，代答问卷为自答问卷的精简版本。这种数据结构上的变化是为了与前面介绍的问卷设计的调整相呼应。如果用户需要分析所有 10 岁及以上样本的自答数据，CFPS2018 的数据结构会比以往轮次的数据更加便利，省去用户链接成人和少儿库的步骤，而且变量也进行了统一。但如果用户需要同时分析 10-15 岁少儿的自答和家长代答数据，CFPS2018 的数据结构

需要用户在进行分析前先跨库链接个人问卷库和少儿家长代答库。这两个库可以直接通过pid 进行链接，然后保留年龄上符合条件的样本。

CFPS2018 在数据结构上第二个调整是 CFPS2018crossyearid 的发布将取代之前所有轮次的 crossyear 数据集。Crossyear 数据库从创建初始就包含了截至到当期调查的所有个人样本，无论这个样本在当期调查时是否被成功访问。由于更新的 crossyearid 库包括之前年份 crossyearid 库的所有样本，本期 crossyearid 发布之后将撤销所有之前轮次的 crossyearid 数据集的下载权限，用户可以直接使用 CFPS2018 的 crossyearid 数据集, 通过 inrosterXX 变量（样本是否在 XX 年份的关系库中）以及 entrayear（样本初次进入 CFPS 年份）来定义所需子样本。譬如用户只感兴趣截止到 2014 年的样本，则可以取 inroster10, inroster12 和 inroster14 中任意变量为 1 的样本。有关跨年核心变量库的详情，请参见稍后发布的《个人层面跨年核心变量库技术报告》。

二、2018 年问卷数据清理步骤

2.1 中断样本的确认

2018 年电访样本比例在 2016 年的基础上进一步提升，而随着电访比例的提高，中断样本的数目也有所增加。为了给用户提供更全面的数据，我们将中断样本中数据完整度达到一定标准的观测纳入发布数据集。根据各库特点，我们在选择纳入标准时有一定的差别。对于家庭成员库，纳入标准为至少完成家庭成员问卷 A、B、C 三个模块，因为根据这些信息我们可以基本确认家庭成员及其家庭关系信息。根据这个标准，最终达到纳入家庭成员库的中断家户样本为 85 条，总共涉及到 320 名家庭成员。家庭经济库的纳入标准为至少完成了经济问卷中“家户收入“整个大模块的数据采集，最终有 56 条经济库样本达到标准被纳入发布库。对于个人库和少儿家长代答库，纳入标准为问卷完成度大于 50%并且受访者至少回答了相关的主要模块：自答问卷中，符合要求的受访者需要完成“上学确认”、“教育史”、“婚姻”模块才能入选发布库；少儿家长代答问卷的入库标准为至少回答到了“入学情况”模块。最终，个人库纳入中中断样本 418 条（包括个人自答问卷 413 条，个人代答问卷 2 条，家庭代答问卷 3 条），少儿家长代答库纳入中断样本 16 条。中断样本在各问卷中均以 interrupt 变量来指征，中断样本的 interrupt 变量值取 1。

2.2 各库样本编码清理

CFPS2018 的各库样本编码清理工作包含以下环节。1) 以家庭关系库为出发点，确定有效家庭样本和个人样本，删除中间过程中产生的无效样本或重复样本。在家庭层面主要涉及两部分，一是结合原家庭和离家单元双重判断来界定离家单元是否为独立的分裂家户，删除非独立离家单元产生的新家庭编码，将有效的问卷信息放回到原家庭；二是结合跨轮次相关家庭的成员列表，保持一个成员只能所属一个家庭的原则，如果出现家庭 A 的成员全部被包含在家庭 B 中的情况，则删除家庭 A 的样本。在个人层面，结合跨轮次相关家庭的个人样本列表来判断个人样本有效性。一方面我们需要核查新加入成员是否与已有个人样本重复，重复的话保留已有样本编码；另一方面我们需要查看新进成员是否第一次同时出现在多个关联家庭中，是的话则需要结合完访个人问卷和所属家庭情况进行编码的取舍。

2）根据家庭关系库界定出的有效家庭和个人样本，决定家庭经济库、个人库、少儿家长代答库的发布样本，基本原则是所有的家庭济库样本、个人库和少儿家长代答库样本都需要从关系库中找到相应的家庭或者个人，但并非所有在关系库中出现的家庭或个人都会在经济库或个人库及家长代答库中出现，因为存在家庭关系库完成后，户内样本没有完全完访的情况。

3）整理问卷内部和跨问卷的重复样本编码。原始问卷内部的重复样本主要通过调查过程中的结果代码、关系库清理结果以及问卷数据的完整度等多类信息综合来取舍。个人库的发布样本来自个人自答、个人代答和家庭代答，在整合最终发布的 person 库中，我们以个人自答样本优先，删除相应样本在代答库中的观测。当样本重复出现在少儿家长代答和其他代答库时，我们以少儿家长代答样本为优先，删除相应样本的个人或家庭代答。总体来说，我们对于个人层面重复样本的处理原则如下：个人自答样本最多保留一条，代答样本也最多保留一条，10-15 岁之间个人自答和家长代答样本可以重复。

2.3 基于实时数据清理的数据更新

调查过程中的实时数据清理有助于我们及时的搜集访员或受访者的反馈，也顺便起到了提醒访员规范访问行为的作用。CFPS2018 实时数据清理包含了以下几方面内容：1）访员备注（F2）信息的整理；2）重点变量极大或极小值的筛查和访员反馈；3）重点变量奇异值的录音核查；4）文本信息中不规范输入的访员反馈。基于这些信息，我们完成了约 80

个问卷变量的更新，共涉及 3047 条观测。改动较多的变量有以下几类。1）职业编码信息：涉及到职业相关变量（如 KGD4、QG303、QEA203），一般为原文本记录不规范，在实时核查过程中通过访员反馈进行了信息补充，共涉及反馈条目 1273 条；2）疾病编码信息：涉及到疾病相关变量（如 QP402A、QP402B、WC5_B_1、WC5_2010），同样为文本记录不规范，涉及反馈条目 1196 条；3）经济库中“万元“和”元“单位混淆，涉及变量包括 FM401、FQ5、FQ6、FR2、FT301、FT302，我们根据访员反馈更新条目 194 条。

2.4 问卷逻辑核查

在调查的测试环节、实施过程中以及调查结束后，我们分批对 CFPS2018 年成员问卷、经济问卷、个人自答问卷的所有模块间的跳转以及重点模块的内部跳转进行了逻辑核查。虽然电子化的问卷系统避免了访员由于个人操作失误导致的跳转错误，但程序一旦失误则有可能带来一整批样本的损失。在确保所有大模块跳转无误的基础上，我们对于教育、婚姻、工作以及迁移等重点模块的跳转路径进行了一一排查。

2.5 重点变量的清理

我们对于部分变量进行了重点清理。以经济库为例，家庭的收入和资产相关变量，它们较易由于单位的混淆或时间区间范围混淆而导致数值过大或过小。对于这些变量，我们进行了如下步骤的清理。

FM401（私营企业或个体经营总资产，单位为万元）：对于所有原始数值超过 1000 的观测，根据家庭收支信息及 FM4（经营净利润）的数值进行判断。当收支或经营利润明显与经营资产不处于同一数量级时，我们向访员发送反馈需求并依据反馈进行相应修正。如果实时数据核查未获得访员方面的有效反馈，数据管理员在后期清理时将结合往期数据综合考虑“万元”和“元”单位混淆的可能性，并酌情进行修正。

FQ5、FQ6、FR2（房屋购建成本、房屋当前市价、其他房产市价，单位为万元）：对于所有原始数值超过 1000 的观测，数据管理员首先根据家庭所在地址信息以及同村居其他家户的相关信息进行判断。若样本所处同村居的其他观测对应的房产相关数值与该家户呈现出明显的数量级差异，则向访员发送反馈需求并依据反馈进行相应修正。如果实时数据核查未获得访员方面的有效反馈，数据管理员考虑访员将题目单位的“万元”错当做“元”的可能性并酌情进行修正。

收入相关变量 FM4、FN301、FR501（经营净利润、离退休金/养老金总额、房租总收入，单位为元）：使用 fid 与其他年份数据的对应变量分别进行链接，若相邻年份呈现出明显的数量级差异，则考虑访员将题目单位的“元”错当做“万元”的可能性，酌情进行数值修正。

资产相关变量 FS6V、FS7V、FT1、FT101、FT201（耐用消费品、农用机械总值、现金及存款、定期存款、金融产品总值，单位为元）：使用 fid 与其他年份数据的对应变量分别进行链接，若相邻年份呈现出明显的数量级差异，则考虑访员将题目单位的“元”错当做“万元”的可能性，酌情进行数值修正。

负债相关变量 FT501、FT601、FT901（待偿贷款额、亲友借款待偿额、尚未归还借款总额，单位为元）：使用 fid 与其他年份数据的对应变量分别进行链接，若相邻年份呈现出明显的数量级差异，则考虑“访员将题目单位的元错当做万元”的可能性，酌情进行数值修正。

需要提醒用户的是，项目组在修改相关数值时的决定非常谨慎。一方面需要数据本身呈现存疑状态，另一方面项目组综合其他信息能基本确定出错原因（譬如单位错误，或者是时间限定错误，或是访员确认），只有在两个条件均满足的情况下，项目组才会对数据进行修正。用户在使用数据的过程中，可能还会发现其他一些存疑情况，但项目组并未对其进行修正。用户可以根据自己的具体研究需要，酌情进行数据处理。

2.6 不同类型问卷的合并

CFPS 的家庭问卷存在面访问卷和电访问卷两种类型，这两种问卷的主体内容没有任何差异，只在访员观察题部分有细微差别，去除了少量不适合电访访员的试题。家庭层面面访和电访的整合可以直接根据样本唯一性来进行操作。

CFPS 在个人层面的样本不仅存在面访和电访模式的差别，还存在自答和代答的区分。面访和电访个人问卷的最主要差别在于认知模块和访员观察题，这些题目只包含在面访问卷中，因此面访和电访数据的合并较为直接，数据结构上的差别就是电访问卷在认知模块和访员观察题上的相关变量是缺失值。但个人问卷自答和代答样本的整合相对复杂，代答问卷一方面在问题的数量上有了大幅度的消减；另外一方面在具体问卷上的问法可能也有所差别。表 1 展示了 10 岁以上个人层面不同问卷类型之间的差别。

在自答和代答样本合并时，在样本层面，如果自答和代答样本同时存在，我们保留自答样本，使得数据更加完整；在变量层面，大部分代答问卷中的问题都与自答问卷中相应问题

相同，我们将其变量名称进行了统一（譬如自答问卷中的 QG12 和 PG12 统一成自答问卷变量 QG12）；对于少量不完全可比的问题（譬如 PG02）维持了代答问卷自己的变量名。

表 1. 个人层面不同问卷类型之间的区别

	面访	电访
自答	**问卷内容**: 最完整; **如何识别**：iwmode=1, selfrpt=1	**问卷内容**：与自答面访相比，少了认知模块和部分访员观察题; **如何识别**：iwmode=2, selfrpt=1
代答	**问卷内容**：跟自答面访相比大为精简; **如何识别**：iwmode=1, proxyrpt=1	**问卷内容**：与代答面访相同； **如何识别**：iwmode=2, proxyrpt=1

2.7 综合变量的添加

除了从问卷中直接生成的变量之外，CFPS 发布数据中还包括了项目组基于问卷原始变量后期生成的综合变量。所有综合变量的基本算法，都可以从项目网站上 “数据文档“页面的”综合变量查询表“进行查询。

这些综合变量的基本情况如下：

2.7.1 家庭收入（家庭经济库）

家庭收入综合变量包括总的家庭收入、人均家庭收入和具体分项收入（家庭工资性收入、经营性收入、转移性收入、财产性收入和其他收入）。其中工资性收入（fwage_1）是指家庭成员从事农业或非农受雇工作的税后工资、奖金和实物形式的福利。工资性收入除了家庭经济库的数据之外，个人问卷中也采集了个体的工资信息，我们在综合变量生成过程中引入了个人自答中的工资信息。我们将经济问卷所采集的工资性总收入与所有完成个人问卷的家庭成员的工资性收入的总和进行比较，最终的家庭总工资收入的取值为二者中的高值。由于 2018 年代答问卷只涵盖了“目前从事的最主要一份工作”或“受访者最近结束的一份工作”

的情况，无法较为准确估计过去一年的工资，我们在 CFPS 2018 工资性收入的计算中未引入个人代答数据，用户如果需要可以自行计算。

经营性收入（foperate_1）是指家庭从事农林牧副渔业生产经营扣除成本后的净收入（包括自产自销部分），以及从事个体经营和开办私营企业获得的净利润。具体计算方法为农业相关收入和私营企业、个体经营收入之和减去农业经营成本得到的结果，当收入低于成本时则此值置为 0，也即经营性收入不为负值。如果用户需要了解经营亏损的情况，可以依据问卷数据中的相关变量自行计算。

转移性收入（ftransfer_1）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（fproperty_1）是指家庭通过投资、出租土地、房屋、生产资料等获得的收入。其他收入（felse_1）是指通过亲友的经济支持和赠予获取的收入。

CFPS2018 在经营性收入、转移性收入、财产性收入的设计与 CFPS2016 基本相同。在工资性收入的设计上的主要区别在于个人层面的代答问卷。2018 年的代答问卷只提问了最近结束的一份工作的相关情况，而在 2016 年则包括了主要工作、实习工作以及过去 12 个月总工资的概括性问题。在其他收入方面的设计上 CFPS2018 与 CFPS2016 也有些细微不同。2018 年的问卷设计将来自不同住的亲戚的赠予收入进一步细分为“来自不同住子女的赠予收入”和来自“其他亲戚的赠予收入”，生成“其他收入”的综合变量时使用的公式与先前稍有差异。具体来说，CFPS2016 中，“其他收入”由“不同住亲戚给的钱”（包括不同住的子女）和“其他人给的钱”两个变量组成，CFPS2018 中这个综合变量则是由“不同住的子女给的钱”、“不同住的其他亲戚给的钱”、“其他人给的钱”三个变量组成。 “其他收入”这个综合变量在两轮调查中的实质组成没有变化，但 CFPS2018 年的设计更加细分。

在生成家庭总收入 fincome1 变量时，我们采用了如下步骤：1）原始数据进行清理后，分别生成工资性收入、经营性收入、财产性收入、转移性收入、其他收入这五项收入。2）将前一步生成的五个分项收入进行加总，将加总的数值与经济问卷受访者所回答的家庭年总收入进行比较，最终的家庭总收入综合变量取值为二者中的高值。人均收入为总收入除以家庭人口数，注意此处的家庭人口数为经济问卷访问过程中出现的，由成员问卷加载过来的 fml_count。这个变量定义了经济问卷访问过程中的家庭规模，与后期清理过后生成的家庭规模变量 familysize18 可能稍有不同。

为了方便用户与 CFPS2010 基线数据的比较，我们还同时生成一版与基线可比的系列变量系列，用尾缀_2 标识，以和之前提及的_1 系列变量进行区分。有关 2010 年可比变量的详细信息，用户可以参考项目网站上第三版用户手册中的相关内容以及“综合变量查询表“的相关信息。

2.7.2 家庭支出（家庭经济库）

家庭支出综合变量包括家庭总支出以及分类别的四大类支出，它们分别是居民消费性支出 PCE（包含食品 FOOD、衣着 DRESS、居住 HOUSE、家庭设备及日用品 DAILY、交通通讯 TRCO、文教娱乐 EEC、医疗保健 MED、其他消费性支出 OTHER），转移性支出 EPTRAN（包括家庭对非同住亲友的经济支持、社会捐助以及重大事件中人情礼），保障性支出 EPWELF（包括家庭购买各类商业保险），建房购房贷款支出 MORTAGE。CFPS2018 家庭支出方面的设计与 CFPS2016 相同。各支出变量的算法均可参照“综合变量查询表“。

家庭总支出的算法类似于家庭总收入。一方面我们可以通过分项加总的方式得出家庭总支出（消费性支出、转移性支出、保障性支出、建房购房贷款支出），另一方面受访者在经济问卷的最后需要给出家庭总支出。最终的家庭总支出综合变量以加总所得的支出为主，当分项信息中受访者无法给出具体数值或者分项加总的数值小于 100 且自报总支出大于 100 时，家庭总支出综合变量的取值来自于自报总支出。

2.7.3 家庭资产（家庭经济库）

家庭资产包括家庭总体净资产和分类别的各项资产，其中包括住房净资产（houseasset_net）、土地资产（land_asset）、生产性固定资产（fixed_asset）、金融资产（finance_asset）、耐用消费品（durables_asset）、房贷外金融负债（nonhousing_debts）。具体而言，住房净资产为现住房价值（resivalue）和其他房产的总价值（otherhousevalue）减去房贷（house_debts）所得，土地资产为农业经营收入的估算产物（具体估算方式与往年相同，可参考项目网站上“数据文档”页面的“综合变量查询表”），生产性固定资产为经营资产（company）和农用器械价值（agrimachine）之和，金融资产为存款（savings）、金融产品（financial_product）、他人欠自家款项（debit_other）之和。房贷外金融负债则包含了来自银行、亲友及其他机构的借贷总额。最终家庭的净资产（total_asset）等于各项资产加总减去各项负债加总。

2.7.4 个人收入 income（个人库）

个人库中的 income 变量的基础是系统自动生成的，基本算法是 incomeA（一般工作的总工资性收入）+incomeB（主要工作的工资性收入），也即 income 变量主要反映个体过去 12 个月（持续到访问最近一年的工作）的工资性收入。需要注意的是，问卷只针对受雇类型的工作进行“工资性收入”的相关提问，如果一般工作或主要工作并非受雇（譬如从事个体/私营经济/其它自雇），则不对工资性收入进行提问。用户可以通过 lastyjob（工作是否延续到最近一年指示变量）以及 jobclass（工作类型）等系列变量对相关问题的跳转进行确认。

我们针对系统自动生成的 income 变量进行了如下修正：当原始变量为拒绝回答或不知道（-1 或-2）时，问卷设计进行了展开式提问，我们生成了相应的收入区间变量，并利用区间的两个端点（最大值、最小值）取均值的方法来获得估计值。如果最终估计的区间在两端，也即比最小值还小，或是比最大值还大，我们则采用最小值的二分之一或最大值的两倍取估计值。

关于问卷中工作总收入的校验（QG1202），incomeA，incomeB 二者如果只有其中一项有值，可以不用进行题目 QG1202 工作总收入的校验。因此，我们对原始的 QG1202 数值进行了修正，如果只有一项有值并且还提问了 QG1202，那么 QG1202 更新为-8。同时我们对问卷也进行了更新，对“QG1202 工作总收入校验”这道题目前的跳转条件“若 Income>0”更新为了“若 IncomeA 和 IncomeB 同时有值”的情况下，再继续提问 QG1202。

2.7.5 认知水平（个人库）

CFPS2018 的认知测试从设计上沿袭了 CFPS2014 的问卷，包括识字测试和数学测试两部分。识字测试和数学测试的原始设计是按照受访者的教育水平采用不同的起点来进行测试，以提高测试效率，但在基线调查时，我们发现不少受访者在起点问题上就选择“不知道”或者打错。为了更准确地估计受访者的认知水平，我们从 CFPS2014 开始允许受访者在较高起点的首道题答错后，降到低一级起点再尝试。

我们按照每答对一道题记一分的方式计算受访者在识字和数学方面的总分，由变量 wordtest18_sc2 和 mathtest18_sc2 表示。同时，我们另外生成了一套假设起点固定时受访者有可能得到的分数，以确保其与 CFPS2010 认知分数的可比性，由变量 wordtest18 和 mathtest18 表示。两种算法的得分相关度非常高，其中识字测试的两套得分相关系

数接近 1，数学测试的相关系数也超过.97。如果用户使用 2014 年或/和 2018 年认知测试数据时，我们推荐使用_sc2 系列变量，但如果分析中包含有 2010 年数据时，我们推荐使用与 2010 年可比变量，也即 wordtest18 和 mathtest18。认知测试只在个人自答的面访模式中采集，接受电话访问的个人没有相关数据。

2.7.6 CESD 抑郁变量

CFPS2018 中采用 Center for Epidemiologic Studies Depression Scale (CES-D)量表来测试个人的抑郁水平。在 CFPS2012 中，我们使用的是包含 20 道题的 CESD20。根据现场访员的反馈，该题受访者接受度不高，于是在 CFPS2016 中，我们选择了一部分随机样本使用了精简的 8 道题版本，另一部分样本依然使用 CES-D20，准备逐步过渡到精简版本。

在 CFPS2018 调查中，我们全部转换到精简的 8 道题版本的 CES-D8。但为了能有效比对不同轮次间的抑郁分数，我们利用 2016 年随机化的样本分配将 CES-D8 和 CES-D20 两套题目的分数进行了对等的操作，使用的方法是百分位数等化方法(equipercentile equating)。在 2018 年数据集中，我们依据 2016 年数据建立的对应关系同样生成了可比的分数 CESD20sc（构建的 CESD20 总分）。这个分数保持了 CES-D20 的打分区间，与 CFPS2012 和 CFPS2016 中的 CES-D20 量表得分也是可比的。数据集中保留了原始的单题分数，用户也可以自行生成自己认为更合适的对等分数。我们建议用户如果只需要使用 CFPS2016 或者/和 CFPS2018 年的数据时，采用 8 道题版本的总分，但是如果用户需要同时使用 CFPS2012 的数据进行分析时，采用 CESD20sc。

2.7.7 教育

教育模块的跳转非常复杂，我们综合加载数据以及不同模块的信息，生成了四个教育系列综合变量：受访者已完成的最高学历 CFPS2018EDU、受访者离校/上学阶段 CFPS2018SCH、受访者已完成的受教育年限 CFPS2018EDUY 和经插补之后的受访者已完成的受教育年限 CFPS2018EDUY_IM。

在算最高学历时，针对个人自答，我们会结合教育史模块的 W01 和 WEDU 以及 EDU_LAST 算受访者已完成的最高学历；针对少儿家长代答和个人代答，由于这两个板块没有教育史，我们会根据 R1_LAST 和 QC3 算受访者已完成的最高学历。在算受访者离校/上学阶段时，我们会根据 R1_LAST 和 QC3 算出离校阶段。初步整理完最高学历和离校/上学阶段之后，我们会对这两个变量之间的相互关系进行一轮逻辑校验和修正，原则上来说，离校/上学阶段应该等于或高于最高学历。

在以上两个变量清理的基础上，我们再计算受访者已完成的受教育年限，具体步骤如下。（1）当离校/上学阶段等于最高学历时，我们将最高学历转换为其相对应的受教育年限。（2）当离校/上学阶段不等于最高学历时，受教育年限变量则会依据最高学历、离校/上学阶段这两个变量为基础进行两种算法的估计，然后在两种算法中取高值，生成受教育年限综合变量。当以最高学历为基础时，受教育年限等于最高学历转化成的受教育年限；当以离校/上学阶段为依据时，受教育年限等于离校/上学阶段的前一个阶段所对应的受教育年限加上后续未完结教育阶段的已完成年限。

若受访者未完结阶段的已读年数缺失，但最高学历和离校上学阶段不缺失，按照上述方式生成的受教育年限可能缺失。我们使用 hot deck 方法对该变量的取值进行插补，并生成插补版的教育年限（CFPS2018EDUY_IM）。

有关教育变量清理的更详细内容请参考技术报告《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》。由于 CFPS2018 已经是第五轮的追踪调查，我们没有仅仅依据 CFPS-36 的方法计算 CFPS2018 当年的值，我们还会结合受访者历年的教育综合变量对 2018 的数据进行填补和修正，做到了历年可比。CFPS2018 教育变量主要有 3 个数据源，分别是个人自答、父母代答和个人代答，同一个受访者根据不同的来源可能会算出不同的教育教育变量值，所以我们也对在三个来源里出现了两次及以上的受访者的教育变量选择了更合理的教育值。除此之外，家庭成员关系库中还有关于最高学历的相关内容，用户也可以考虑根据关系库的学历内容进行适当填充。

用户如果希望自己通过原始变量来进行整理或者核对，需要注意问卷中涉及教育阶段的原始值和项目组最终教育综合变量的取值有所不同。为了与历年的教育综合变量可比，我们在清理过程中进行了数值转换，譬如在问卷中小学教育程度是“3”，我们在算综合变量时会把其转换为“2”。

2.7.8 当前工作状态 employ

个人自答的 employ 变量是系统根据受访者在【GB 当前工作状态确认】模块回答的信息自动生成的，基本算法如下：（1）如果受访者①过去一周至少工作了 1 个小时、或②能够在确定的时间或者 6 个月以内回到原来的工作岗位、或③从事个体经营活动，但是目前处于生意淡季，等过一段时间还会继续经营、或④从事农业方面的工作但是目前处于农闲季节，则判定受访者有工作，employ=1；（2）如果受访者过去一个月找过工作，且如有工作机会能在两周内开始工作，则判定受访者失业，employ=0；（3）如果受访者过去一个月没有找过工作，

或如有工作机会不能在两周内开始工作，则判定退出劳动力市场，employ=3；(4)其他情况，employ=-8。

个人代答问卷并没有采用整套【GB 当前工作状态确认】模块，但是我们根据 GB1 补充了部分受访者的当前工作状态。如果受访者过去一周至少工作了 1 个小时，则判定受访者有工作，employ=1；其他情况，employ=-8。

2.8 各类编码工作

2.8.1 职业编码、行业编码和职业威望

CFPS2018 采集了受访者的详细工作信息，涵盖了自家农业生产活动、农业打工、受雇、非农自雇以及家庭帮工。我们对这些原始的信息进行编码后，生成不包含隐私信息且较方便分析的数据。为了方便用户使用，我们在个人数据库中生成行业编码系列变量，包括实习工作行业（QGA4CODE）、第一份工作行业（KGD3code）、主要工作行业（QG302code）；职业编码系列变量，包括实习工作职业（QGA401CODE）、第一份工作职业（KGD4code）、主要工作职业（QG303code）、配偶/同伴职业系列变量（QEA203code、EEB4022_A_1code）。关于职业和行业编码的更详细信息，用户可以参考即将发布的技术报告《中国家庭追踪调查中的文本编码》。

我们根据生成的职业编码，创建了职业威望系列变量：与实习工作职业编码 QGA401CODE 相关的三个职业威望变量（qga401code_isco、qga401code_isei、qga401code_siops）、与第一份工作职业编码 KGD4code 相关的三个变量（kgd4code_isco、kgd4code_isei、kgd4code_siops）、以及与主要工作职业编码 QG303code 相关的四个变量（qg303code_isco、qg303code_isei、qg303code_siops、qg303code_egp）。有关这些职业威望的具体计算方法，用户可以参考《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》。

2.8.2 疾病和死亡编码

CFPS2018 调查在问卷的两个地方采集死亡原因，一是在家庭成员问卷中由成员回答人提供家庭成员中是否有人去世以及去世原因；二是在个人问卷中，对于初次确认死亡状态的配偶进行死亡原因的采集。访员在现场对死亡原因进行编码。

CFPS2018 年在个人问卷中询问了关于慢性疾病的信息，并在后期处理过程中生成变量慢性疾病编码（QP402Acode 和 QP402Bcode)。在少儿家长代答问卷中，我们询问了儿童过去 12 个月最严重疾病，并在后期生成编码变量（WC5_B_1code）；对于之前年份未进行相关数据采集的受访者，我们还会询问其出生后最严重疾病，并形成后期编码（WC5_2010code）。

CFPS2018 将“过去两周身体不适”相关问题（QP302）进行了设计上的调整，由之前的选择题变成了自由文本题。为了与其他年提供尽量可比的数据，我们对此变量进行了编码（QP302code）。编码之后各个数值代表的具体含义可以从项目网站上“数据文档“页面的 codebook 中查找。

2.8.3 其他编码

除了上述的编码之外，CFPS2018 还对职业期望（个人库中变量为 QS801_B_2CODE，少儿家长代答库中变量为 WD101code）、行政管理职务（个人库中变量为 QG1401code）进行了编码。我们采集了正在上学的受访者的就读学校，涵盖了小学、初中一直到博士阶段的院校名。对于大专及其以上的学校，我们采用之前年份相同的编码方案形成了高等教育的学校类型编码（QS1_B_1CODE）。除此之外，我们依据教育部发布的《中等职业学校专业目录》（2010 年修订版）生成了专业编码。在个人自答库有针对在读的职业初中学生的所学专业（QS401code）、在读的职业高中学生的专业（QS501_B_1code）、初中阶段所学专业（KW1002_B_1code）、高中阶段所学专业（KW1002_B_2code）；在少儿家长代答库中，专业变量有初中在读学生所学专业（WS401code）、高中在读学生的专业（WS501code）。我们依据教育部发布的《学位授予和人才培养学科目录(2011 年)》生成了学科编码，包括学校基本情况模块的大专在读学生所学专业的学科 QS701_B_1 和本硕博在读学生所学专业的学科 QS9code，教育史模块的大专阶段主修专业的学科（KW1003_A_1code）、大学阶段主修专业的学科（KW1003_A_2code）、硕士阶段主修专业的学科（KW1003_A_3code）和博士阶段主修专业的学科（KW1003_A_4code）。

2.8.4 地址编码和城乡状态

与往年相同，CFPS2018 的地址信息给出了三级编码：省级（provcd）、区县级(countyid)和村居级(cid)，他们分别代表家庭或者个人在接受当期 CFPS 调查时的居住地址。CFPS 家庭层面的地址信息来自家庭问卷回答人，如果该信息存在缺失，我们也会尝试利用在家个人的个人问卷信息进行适量补充。CFPS 个人层面的地址信息主要来自个人地址模块、EHC 地

址模块，当该信息缺失时，我们也会根据家庭层面的地址（针对在家个人）以及离家单元地址（针对外出个人）进行适量补充。如果样本是在家个人，他们的地址取自家庭地址；如果样本是外出个人，他们的地址则来自于关系问卷中的外出单元模块。由于外出单元模块中的地址为家人代答，信息不完整的现象较为常见；同时由于外出单元地址中村居层面不是结构化信息，这些外出样本的地址在村居编码(cid18)以及城乡属性(urban18)上存在系统性的缺失。代答样本中由于离家比例较高，再加上没有自答问卷所提供的 EHC 地址，在地址上的缺失情况相对于自答样本更为突出。

发布数据集中的省码是对应到具体省份的国标码，区县和村居码均为 CFPS 项目编制的顺序码，不是标准代码，数值本身没有实质性含义，也无法与外部数据链接。需要使用更具体地址信息的用户请参考 CFPS 项目网站上“数据中心“下面的”限制数据“栏目。需要注意的是对于家庭代答的个体样本（proxytype=2），他们的个人地址信息来自于原家庭的外出地址模块。CFPS 的区县和村居编码基本实施与国标码一一对应的关系，也即当样本家庭所居住区域的国标码发生变动时，其对应的地址编码也会发生变动。

我们同时提供了按 2018 年国家统计局网站上定义的各样本所在村居的城乡性质(urban18)。对于部分村居样本无法准确进行编码的情况，我们通过地图对其村居性质进行了一定的补充。

2.9 权数

CFPS2018 权数依然分家庭和个人层面两大类。如往期不同，我们不再单独提供针对再抽样样本的权数，而只提供针对所有样本的权数。个人层面的权数针对 CFPS 定义的基因成员（包括基线界定的基因和后期追踪调查时新产生的基因），存在截面权数（rswt_natcs18n）和追踪权数（rswt_natpn1018n）两组。家庭层面权数针对所有家庭，只包含截面权数，而不再包含追踪权数，这是由于家庭内部基因成员的死亡、婚姻、迁移在不断变化，跨年之间缺乏类似于个人的可比性。

个人层面截面权数针对 2018 年完成个人访问的基因成员，它的基本算法分为两步：一是计算 2010 年个人基础权数（针对 2010 年家庭成员关系中所有满足条件的个人）；二是计算 2018 年个体的应答权数，个人截面权数是这两个权数的乘积。个人层面的面板权数针对 2010 年完访并且在 2018 年也完成了个人访问的基因成员，它的计算方法是先构建 2010 年

个人完访问卷的权数，在此基础上再考虑 2018 年的流失情况，形成 2010 和 2018 年的面板权数。

家庭层面的截面权数建立在个人权数的基础上。由于 CFPS 追踪的是基因成员[2]，家庭层面截面基础权数为2018年该家庭中所有基因成员的2018年个人权数的均值。在此基础上，我们考虑家庭层面的样本流失，形成家庭层面的流失权数。最终的家庭截面权数是 2018 年的家庭基础权数与家庭流失权数系数的乘积。

由于 CFPS 样本由基线所在的 25 个省市已经扩散到全国，之前的总量权数在追踪调查中缺乏特定的参照体系，我们将权数改为标准化形式，它由总量权数除以总量权数的均值得出。跟往年一样，如果用户进行截面数据分析，使用家庭或个人层面的截面权数；如果用户分析个人在跨年间的变化，则可以使用我们提供的个人面板权数。如果用户需要进行家户层面的跨年变化分析，需要在家户号跨年匹配的基础上额外关注家庭成员的结构在跨年间是否有所变化，并根据自己定义的跨年家庭进行流失概率分析，构建相应的流失权数；在没有自己构建的面板权数时，也可以直接借用家庭截面权数。

有关 2018 年权数创建和使用方法的更具体介绍，可以参考后续推出的技术报告《中国家庭追踪调查的权数调整》。

三、数据库简介

CFPS2018 访问问卷包括家庭成员问卷、家庭经济问卷、个人自答问卷、个人代答问卷以及少儿家长代答问卷。在家庭成员问卷访问过程中，我们会让家庭问卷回答人对于外出的个人（包括经济离家和物理离家）的个人提供一份家庭代答问卷。调查问卷发布在项目网站“文档中心“下面的”调查问卷“栏目。

在数据发布时，我们将个人自答问卷数据和个人层面的代答问卷数据进行了整合，形成了针对 10 岁及以上个体的个人库。其余各个问卷（成员问卷、家庭经济问卷、少儿家长代答问卷）分别对应一个单独的数据集。每个数据集相应的 codebook 发布在项目网站“文档中心”下面的“数据文档”栏目。CFPS2018 的数据库基本情况如表 2 所列。跨年个人样本综合变量库（crossyearid）将后续单独发布。

[2] 其他家庭成员依据其与基因成员的关系是否存续决定是否需要调查

表 2 CFPS2018 年各库基本状况[3]

数据库	样本量	变量数
家庭关系库	58504	296
家庭经济库	14218	321
个人库	37354	1371
少儿家长代答库	8735	289

3.1 家庭成员关系库(famconf)

家庭成员关系库以 CFPS 界定的每个家庭成员为一行，家庭成员以 pid 标识，包括 2010 年基因成员及之后调查年新增的家庭成员的配偶(_s 系列变量)、父亲（_f 系列变量）、母亲（_m 系列变量）及子女（_c1-_c10 系列变量）的基本信息。同处一个家庭的成员拥有同样的家户号 fid18。除了 fid18 之外，关系库还标识出了该家户在以往轮次中的家户号，分别以 fid16、fid14、fid12、fid10 来标识。对于同一个观测，如果不同年份的家户号数值不同，说明该家户曾经发生过由于部分成员经济独立所导致的家户分裂。2018 年家庭成员关系库中包括来自 15,051 个家庭的 58,504 条个人样本。在 CFPS2018 关系库中，我们将所有个人都只保留了一条观测，其中优先保留其在当前家庭的观测（co_a18_p=1），如果当前家庭在家庭层面未完访，则保留个体在上一级家庭的观测(co_a18_p=0)。

与往年家庭关系库数据相比，2018 年对已有系列变量（TB601_A18_*：离家原因）进行更新。TB601 基于家庭成员问卷中的问题 A3，其中对于“77”类别（77. 其他原因【请记录受访者原话】）所采集的开放性文本信息进行了归类编码，在原有分类基础上新增以下类别：出境、外出打工、外出上学、离婚、结婚、探亲、分家、迁移、外出就医。除此之外，在 CFPS2018 关系库中还新增了一些变量。为了便于用户使用关系库，总结了综合变量和新增变量的信息来源及生成算法，详见下表 3。

表 3　CFPS2018 家庭关系库新增变量

变量名	变量标签	家庭成员问卷问题编号	算法简要描述
往期成员库已有变量			
FID_PROVCD18	2018 年家庭省级国标码		根据地址模块清理省级国

[3] 关系库统计值基于发布版本为 1 的数据集，其它各库统计值基于发布版本为 2.1 的数据集。

			标码
FID_COUNTYID18	2018 年家庭区县顺序码		根据地址模块清理后区县国标码再重新编码
FID_CID18	2018 年家庭村居顺序码		根据地址模块清理后村居国标码再重新编码
FID_URBAN18	2018 年家庭城乡分类（基于国统局）		家庭所在村居根据国统局的城乡类型进行村居属性的分类
SUBSAMPLE	是否在全国再抽样样本中		基于基线原家庭（fid_base）对应的抽样信息
SUBPOPULATION	抽样子总体		基于基线原家庭（fid_base）对应的抽样信息
GENETYPE18	2018 年基因类型		根据 2018 年家庭成员类型表、是否是基因成员重新编码
fid1*	2010、2012、2014、2016 对应的家庭编码		历年的家庭归属
FAMILYSIZE18	家庭成员人数		汇总家庭中 co_a18_p=1 的值
TB2_A_P	个人性别	BC2、E1、D105	成员问卷新采集的、个人问卷中采集的、及往年已有的性别信息的综合。
TB1Y_A_P	个人出生年	BC3、E2、D104	成员问卷新采集的、个人问卷中采集的、及往年已有的出生年信息的综合。
TB1M_A_P	个人出生月	BC3、E2、D104	成员问卷新采集的、个人问卷中采集的、及往年已有的出生月信息的综合。
TB3_A18_P	个人婚姻	BC4、E3	成员问卷新采集的、个人问卷中采集的、及往年已有的婚姻信息的综合。
TB4_A18_P	个人最高学历	BC5、E4	成员问卷新采集的、个人问卷中采集的、及往年已有的最高学历信息的综合。
HUKOU_A18_P	个人户口	BC6、E5、D106	成员问卷新采集的、个人问卷中采集的、及往年已有的户口信息的综合。
TB6_A18_P	个人是否居住在家	A2、A201	
CO_A18_P	个人是否与该家庭经济上是一家人	F102、B1	优先以离家人自己主观判断是否经济独立为准，其次以原生家庭的主观判断来界定。
OUTPERS_R_WHERE18_P	离家人（个人）的居住区域	G1、H1	整合成员问卷中单人离家和多人离家的信息

TB602ACODE_A18_P	离家（个人）省国标码	G101 、H101	整合成员问卷中单人离家和多人离家省份信息
TB601_A18_P	离家（个人）原因	A103	值（1-6）保留问卷原选项值，将问卷选项 77 采集的文本信息进行归类，并重新编码为（10-18）
OUTUNIT18	外出单元序号	F1	将原生家庭把离家人员划分在不同的单元，依次顺序编码。
COREMEMBER18	是否是核心成员	“CFPS 家庭人员类型”表	表中的 4 以及往年的核心成员
CFPS2018_INTERV_P	个人本轮个人问卷是否完访		汇总各类个人问卷的完访情况
ALIVE_A18_P	个人是否健在	A3	成员问卷新采集的、及往年已有的死亡信息的综合。
TA4Y_A18_P	个人去世年份	A4	成员问卷新采集成员去世年份信息。
TA4M_A18_P	个人去世月份	A4	成员问卷新采集成员去世月份信息。
TA401_A18_P	个人去世原因	A401	成员问卷新采集成员去世原因信息。
pid_a_*	父亲、母亲、配偶、10 个孩子的样本编码	C2、C3、C4、C5	2018 年新采集信息与历年家庭关系信息的整合。

2018 年成员库新增变量

TB602CCODE_A18_P	离家（个人）区县顺序码	G102 、H102	整合成员问卷中单人离家和多人离家区县信息，并重新编码。
TB602CCODE_A18_*	离家单元区县顺序码	G102、H102	区县国标码再重新编码
C105_A18_P	个人进入该家庭的原因	C105	
RTYPE_END18	家庭成员问卷中 rtype 的含义		见家庭成员问卷“CFPS 家庭人员类型”中的说明
PSU	基线家庭初级抽样单位		Fid_base 在 2010 年基线采样时所对应的 PSU
C105_A18_P	个人进入该家庭的原因	C105	
ADS1_18	是否搬家	ADS1	
KZ103_18	访问使用的主要语言	Z103	
FID_BASE	基线家庭样本编码		样本所在 fid18 回溯到 2010 年基线调查时的源头家庭

3.2 家庭经济库(famecon)

家庭经济库以家庭为单位，fid18 为每个家庭的唯一标识符。如前所述，该家户在之前年份的家户号用 fid10-fid16 标识。对于跨年间家户未发生变动的家庭来说，其家户号维持不变；但如果跨年间家户发生分裂，被认定为依然在“原家庭”的那些家庭成员所在的家户号依然不变，只有被认定为是“新家庭”的家庭成员所在的家户号才会发生改变。因此我们不能完全依据家户号来判断一个家庭是否发生结构上的变化。准确判断一个家户是否与上一轮调查完全一致要依赖于家庭关系库中的家庭成员构成。如果需要了解关于家户分裂的相关信息，用户可以参考“用户手册”。

经济库中的样本包括往期调查所界定出来的原生家庭以及在 2018 年调查时发现由家庭因婚姻变化、子女经济独立等原因所派生出来的新组家庭。访问方式为面访或电访（以 iwmode 标识），面访和电访的问卷内容完全相同。家庭经济库中的所有观测都来自于家庭关系库，但并非所有存在于家庭关系库的家户都一定会在经济库中有观测。少量家庭在完成关系库之后并未完成家庭经济问卷的回答。

在 CFPS2018 家庭经济库中，有两个变量与家庭规模相关，它们分别是 familysize18 和 fml_count。其中 familysize18 是后期生成的，来自家庭关系库，由关系库中属于同一个家庭的成员数确认（fid18 相同，且 co_a18_p=1）。Fml_count 是调查实时的加载变量，由调查在进行过程中由家庭关系问卷中直接加载过来，代表了访问时调查系统界定的家庭成员数目，这个数目是受访者在回答家庭经济库时的参考。二者大部分时候相同，但在某些情况下不一致，不一致的主要原因在于如果原家庭存在离家单元，当离家单元对自己的经济关系与原家庭认定不一致时，我们将以离家单元自身的界定为准。譬如原家庭回答人可能认为离家单元与自己有经济联系依然是一家人，但离家单元自身界定自己已经经济独立，变成了独立的家户，这时 familysize18 和 fml_count 可能会出现差异。

3.3 个人库(person)

个人库包括所有 10 岁及以上个人的问卷数据，个人样本由 pid 唯一标识。Pid 在跨年跨

问卷的数据集中保持不变，个人层面的不同数据集都可以通过 pid 来进行链接。除了个人标识符之外，个人库还给出了个人所在的家户号 fid18，如果需要将个人问卷与家庭问卷进行链接，则可以通过 fid18 作为链接变量进行跨库链接。与往期的成人库相比，CFPS2018 的个人库多出了 10-15 岁样本的个人自答问卷。如果用户需要构建与往期在结构上相似的成人库，可以通过年龄的筛选条件保留 15 岁以上的个人样本即可。

CFPS2018 个人库包含个人自答、个人代答（针对因为身体原因无法完成自答）和家庭代答（针对离家单元中的个人，一般由成员问卷的回答人统一代答）的样本，访问模式为面访和电访（用 self_iwmode 和 proxy_iwmode 进行标识）。个人自答的面访和电访的区别主要在于认知模块，面访中包括字词和数学这两个认知测试，但电访中没有；除此之外，其他问题在面访和电访中已经统一。个人自答的样本使用 selfrpt 变量标识，当 selfrpt＝ 1 时，该样本为个人自答样本。

在个人库中，我们用变量 PROXYTYPE 标识出该样本是来自于家庭代答还是个人代答，注意家庭代答的样本可能来自上一级家户的家庭成员代答，而个人代答一般来自当前家户的家庭成员代答。从问卷内容上，家庭代答和个人代答的问卷内容完全相同；但代答问卷和自答问卷的差别较大，代答问卷从内容上到具体问题上均为自答问卷的简版。在整合自答和代答数据的过程中，我们对自答和代答问卷的问题进行了比对，如果提问方式上没有任何差别，我们则统一了变量名（以自答变量名为准），作为同一个变量进行处理；但如果提问方式上有差异，我们则保持了自答和代答各自的变量。

在个人问卷的实地访问阶段，我们发现部分样本进入个人自答问卷的第一份工作和主要工作模块的跳转存在错误，有 4000 多条样本未采集到相关的工作信息，我们及时采取了补访措施。对这部分样本重新采集第一份工作和主要工作模块的内容，成功补访了 3000 多条样本，并在数据中生成 GDGE、GDGEyear 和 GDGEmonth 三个标识变量，分别指示出对第一份工作和主要工作进行补访的样本，以及相应的补访时间。

对于数据中的多项选择题，为了方便用户使用，我们在原始变量的基础上又生成一套，针对每一个选项显示是否符合的0、1变量。譬如KGD2A这个系列添加了一组KGD2A_a_1-KGD2A_a_5，分别显示“求职渠道：是否自己直接与用人单位联系”等。涉及到的多项选择题有7组，分别是KGD2A、QG7、QG8、QG9、KG13A、QI301和QP605。

CFPS2018 自答个人问卷中新增了两套量表，分别是针对 16 岁及以上个体的“大五人格”量表，以及针对 10-15 岁青少年的问题行为量表。

3.3.1 大五人格(big five personality measure)

在CFPS2018个人问卷中，CFPS首次对15 岁以上的受访者从人格测量的五个维度进行了数据采集，这五个维度分别是：尽责性、外向性、亲和性、开放性和情绪不稳定性。我们使用的测量工具源自美国、德国以及英国的家户调查（PSID、GSOEP 和BHPS）中人格测试基本相同的版本，包含15道题，每个维度对应三道题（见表4）。用户在使用过程中，可以将各维度三道题综合形成五个综合变量，在生成综合变量的过程中注意各维度内部存在的反向提问。用户也可以考虑去掉相关反向提问的问题后创建综合变量，这种做法可在一定程度上提高各维度内部的一致性。[4]

表4 CFPS2018中简版人格量表15个条目内容及相关维度

变量名	问卷内容	相关维度
QM201	做事严谨认真	尽责性
QM202	爱说话	外向性
QM203	有时对别人粗鲁、不客气【反向提问】	亲和性
QM204	具有独创性，会产生新点子	开放性
QM205	经常会担心	情绪不稳定性
QM206	天性比较宽容	亲和性
QM207	往往很懒惰【反向提问】	尽责性
QM208	开朗、善社交	外向性
QM209	重视艺术和审美的体验	开放性
QM210	容易紧张	情绪不稳定性
QM211	做事有效率	尽责性
QM212	含蓄、保守【反向提问】	外向性
QM213	为他人着想、对几乎每一个人都和蔼	亲和性
QM214	想象力丰富	开放性
QM215	是放松的，能很好地应付压力【反向提问】	情绪不稳定性

3.3.2 青少年偏差行为(problem behavior)

在CFPS2018个人问卷中，CFPS首次对10岁至15岁的青少年受访者采集了他们的偏差行为情况，包括内化偏差行为(internalizing problem behavior)和外化偏差行为(externalizing problem behavior)。青少年内化或外化偏差行为的测度最早来源于James Peterson和Nicholas Zill设计的Achenbach偏差行为量表。美国收入动态调查（PSID）的扩展调查“儿童发展调查”（CDS）中引入了Behavior Problems Index，来测量内化和外化问题两个维度，共包含32道题。我们在CFPS2018年采用了源自美国Early Childhood Longitudinal Study中较为精简的

[4] 有关 CFPS2018 中大五量表的数据评估和用法推荐，可以参考 “吴琼,谷丽萍.简版人格量表在中国大型综合调查中的应用[J].调研世界,2020(05):53-58.”

版本，包含14道题，其中内化问题8道（变量名以Qint打头），外化问题6道（变量名以Qext打头），详情见表5。我们评估了这套量表的信度和效度。我们首先依据原始量表所设定的维度，利用验证性因子分析模型评估了内化偏差行为这个维度的8道题是否同属于一个维度，以及外化偏差行为这个维度的6道题是否同属于一个维度。初始结果表明，单维度的验证性因子分析模型与实证数据拟合度不佳。[5] 后续分析发现，在内外化维度中，分别有一对变量（内化偏差行为中的QINT005和QINT007；外化偏差行为中的QEXT004和QEXT006）显示出超越于与其他变量相关的共同因素。这一对变量的存在导致单一的维度无法很好地解释实证数据。当我们把其中的一个变量从该量表中删除后（譬如内化偏差行为中的Qint007, 以及外化偏差行为中的QEXT006），剩下的变量则可以较好的由一个共同因素来解释，也即与单维度的结构形成较好的拟合度。调整之后的量表在结构效度上呈现出单维度的特性。[6]内化问题维度内部一致性系数Cronbach's alpha是0.65，外化问题维度的Cronbach's alpha系数是0.64。

基于以上分析结果，我们对这套量表的推荐用法如下：在创建内化偏差行为综合变量时，采用QInt001、QInt003、QInt005、QInt009、QInt010、QInt011、QInt014这七个变量，形成总分（或是平均分）；在创建外化偏差行为综合变量时，采用QExt002、QExt004、QExt008、QExt012、QExt013这五个变量，形成总分（或是平均分）。我们在发布库中将这套量表的原始14个变量全部包含在其中，用户也可以选择自己认为更合适的方式进行数据分析。

表5 CFPS2018中的简版青少年问题行为量表14个条目内容及相关维度

变量名	问卷内容	相关维度
QInt001	学习上遇到困难我会生气	内化问题
QExt002	我经常和同龄人吵架	外化问题
QInt003	我害怕考试	内化问题
QExt004	我很难集中注意力	外化问题
QInt005	我经常感到寂寞	内化问题
QExt006	我很容易分心	外化问题
QInt007	我经常觉得悲伤难受	内化问题
QExt008	我很难完成学校的作业	外化问题
QInt009	我担心自己在学校表现得不够好	内化问题

[5] 对于内化偏差行为这个维度，利用单维度验证性因子分析模型分析原始的 8 道题，得出的拟合度指数如下：chi-square=338.44, p<.001, RMSEA=.081, CFI=.868, TLI=.816, SRMR=.044. 对于外化偏差行为，相应的拟合度指数如下：chi-square=357.14, p<.001, RMSEA=.126, CFI=.834, TLI=.723, SRMR=.057

[6] 对于内化偏差行为这个维度，利用单维度验证性因子分析模型分析去掉 QINT007 之后的 7 道题，得出的拟合度指数如下：
chi-square=104.653, p<.001, RMSEA=.051, CFI=.950, TLI=.926, SRMR=.027.对于外化偏差行为，基于去掉 QEXT006 后的 6 道题分析得出的拟合度指数如下：chi-square=71.091, p<.001, RMSEA=.073, CFI=.955, TLI=.909, SRMR=.030.

QInt010	我担心做不完作业	内化问题
QInt011	我担心在学校没有玩伴	内化问题
QExt012	我因为多嘴打扰到别人而惹麻烦	外化问题
QExt013	我因为和同龄人打架而惹麻烦	外化问题
QInt014	在学校犯错误的时候我感到羞愧	内化问题

3.4 少儿家长代答库（childproxy）

少儿家长代答库包含的是所有 0-15 岁少儿的家长代答数据。少儿家长代答库的观测单位为少儿，每个被访问到的孩子为一行，以 pid 标识，其代答的家长以 respc1pid 来标识。代答人一般为最熟悉孩子的家庭成员，可能是父母，还可能是其他家庭成员。与往期的少儿库相比，CFPS2018 的少儿家长代答库只有家长代答数据，没有 10-15 岁少儿的自答部分。如果用户需要创建跟以往年份结构上可比的少儿库，可以从 CFPS2018 的个人库中提取 10-15 岁的少儿样本，与少儿家长代答库用 pid 作为链接变量进行链接。

少儿家长代答库包括少儿家长代答样本，以及少量来自家庭代答的 0-15 岁少儿样本。数据库中同样用标识符变量 PROXYTYPE 对少儿家长代答（1）和家庭代答（2）进行区分，家长代答问卷的内容比家庭代答丰富。访问模式为面访(IWmode=1)和电访(IWmode=2)。

在 CFPS2018 少儿家长代答问卷访问过程中，有 36 条样本由于 iinterv 的调用失误，导致 WA302（现在的户口所在地）、WA102（孩子的胎龄）、WA106（孩子在哪出生）三个变量出现了缺失，我们在数据中用“-9”进行标识。