China Family Panel Studies
中国家庭追踪调查
技术报告系列：CFPS-36
系列编辑：谢宇　责任编辑：谷丽萍

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

2019.1.14

一.背景介绍

受教育水平是社会科学研究中的核心变量之一。CFPS2010 收集的受教育水平相关变量主要包括在学状态、上学阶段、已完成的最高学历。CFPS2012 重新确认了这些变量，并对已离校者追加提问了离校时的阶段。CFPS 中教育数据来源的多样化虽然有助于为研究者提供更为准确与全面的信息，但也带来了一些数据使用上的困难。一是灵活性、个性化的信息采集方式极大地增加了问卷设计和数据库结构的复杂程度，这在一定程度上给用户理解问卷及合并数据带来了困难。二是不同来源的数据的不一致影响了用户对信息的判断选择与使用。

CFPS 数据团队综合利用 CFPS2010 和 CFPS2012 教育模块的多个变量信息，对 2010 年的受教育水平变量进行了补充与修正，并生成了最高学历、离校阶段（针对已经离开学校的个人）或正在上学的阶段（针对还在上学的个体）、受教育年限这三个最佳变量（表 1）。由于这些综合变量的生成需要综合考虑家庭库和个人库的信息，我们只对 2010 年存在有效个人问卷的受访者生成了这三个最佳变量。出于方便用户使用的考虑，我们同时将最高学历最佳变量放入了家庭成员数据库中，对于没有个人问卷的家庭成员，该最佳变量的取值为家庭问卷代答中的原值。本报告简述了创建 2010 年个人问卷受访者受教育水平最佳变量的方法，并对相关变量进行了综合质量评估。

表 1.　CFPS2010 受教育水平最佳变量列表

变量名	变量标签
cfps2010edu_best	CFPS2010 个人问卷受访者已完成的最高学历
cfps2010sch_best	CFPS2010 个人问卷受访者离校/上学阶段
cfps2010eduy	CFPS2010 个人问卷受访者已完成的受教育年限

注：表中三个变量可以从 2018 年发布的 2010-2016 年核心变量库中查找。

二.变量清理步骤

1. 受教育水平数据收集

在 2010 年 CFPS 基线调查中，受访者的受教育水平主要有两个信息来源：一是 10 岁及

以上受访者本人以及16岁以下少儿的监护人在个人问卷中提供的关于受访者“已完成（毕业）的最高学历”或“正在上哪个阶段”，二是家庭回答人在家庭成员问卷（T1表）中为每位家庭成员提供的“已完成的最高学历”。

为确保数据库中重要社会人口信息的完整性与准确性，CFPS在2012年追踪调查中再次向受访者确认了2010年的年龄、性别、婚姻状况、受教育水平等信息的准确性。其中，重新确认的与教育相关变量有：最近一次调查时受访者的在学状态、上学阶段、已完成的最高学历。这些信息可以为完善和修订2010年的相关教育变量提供重要的参考依据。

与此同时，自2012年追踪调查起，CFPS对受访者受教育水平信息的采集方式进行了调整。在2010年基线调查中，CFPS对离校受访者通过提问其“已完成（毕业）的最高学历”来采集受教育水平信息。然而，在后期的数据分析中，我们发现该提问方式存在两方面的问题。第一，该提问方式在测量上忽视了中途离校/辍学的受访者最后一个阶段的学校教育，使得这一类受访者的实际受教育年限被低估，并导致其最后一个阶段的教育信息缺失。第二，部分受访者由于未能准确理解问题，从而将其未完成的最后一个阶段误报为其已完成的最高学历，导致该部分受访者高报其最高学历。为解决上述问题，CFPS自2012年起在个人自答问卷中将对离校者受教育水平的提问方式改为了“您从哪个阶段离开学校”。相比2010年的问题，该提问方式不仅可以了解到受访者实际的受教育水平以及中途离校/辍学者最后一个阶段的教育信息，而且表达方式更为明确，可避免受访者对问题的错误理解。结合入学年份、离校年份、是否毕业等信息，我们可以进一步推断出受访者的最高学历。为弥补2010年的缺陷，CFPS在2012年追踪调查中对全部离校者重新采集了最后一个阶段的学校教育信息，可用来对2010年离校者错误或缺失的最后一个阶段的受教育信息进行更正和插补。

CFPS借助计算机辅助调查技术，通过加载单个受访者先前的调查数据和编写相应的问卷程序，为每一位受访者打造了适应其具体情况的问卷以进行有针对性的信息确认与信息补充，既保证了访问的灵活性与高效性，又提高了受教育信息采集的完善度与准确度。

2. 2010年受教育水平数据清理

CFPS2010最高学历、离校/上学阶段、受教育年限这三个最佳变量综合了CFPS2010和CFPS2012的教育模块的多个变量信息。本节将详细介绍数据清理与构建最佳变量的具体方法。

2.1 数据信息来源构建：代答值、自答值、逆推值

CFPS2010 教育信息最佳变量综合了三大类数据来源。第一类是“代答值”， 直接来自于 2010 年家庭成员关系问卷，涉及到的教育信息只有最高学历变量。第二类是“自答值”，综合了 2010 年成人库和少儿库个人问卷的多个教育变量，涉及到的教育信息有最高学历和离校/上学阶段。最高学历直接来自教育史模块的最高学历自答值，离校/上学阶段直接来自个人问卷中的上学模块或家长代答模块，但成人问卷未对当前不在上学的人提问离校阶段。第三类是“逆推值”，根据 2012 年成人库和少儿库个人问卷的多个教育变量逆推 2010 年的最高学历和离校/上学阶段。表 2 展示了逆推值的具体计算方法。附图 1 展示了构建代答值、自答值、逆推值中涉及到的具体变量。

表 2. 2012 年个人问卷教育信息逆推 2010 年最高学历和离校/上学阶段方法

	逆推方法	
	2010 年最高学历逆推为	2010 年离校/上学阶段逆推为
2010 年、2012 年均在上学		
2010 年调查时间在 2012 年入学年份的七月之前	2012 年最高学历降一级（或两级）[a]	2012 年上学阶段降一级（或两级）[a]
2010 年调查时间在 2012 年入学年份的七、八月升学季	2012 年最高学历[a]	2012 年上学阶段降一级（或两级）[a]
2010 年调查时间在 2012 年入学年份的八月之后	2012 年最高学历	2012 年上学阶段
2010 年在上学，2012 年不在上学		
2012 年离校时状态为毕业	2012 年最高学历降一级（或两级）[a]	2012 年离校阶段
2012 年离校时状态为辍学	2012 年最高学历	2012 年离校阶段
2010 年不在上学，2012 年在上学		
2010 年调查时间在 2012 年入学时间之前	2012 年最高学历	2012 年离校阶段降一级（或两级）[a]
2010 年调查时间在 2012 年入学时间之后	2012 年最高学历	2012 年离校阶段

2010 年、2012 年均不在上学	2012 年最高学历	2012 年离校阶段

注：a.对 2012 年最高学历或离校/上学阶段是大学本科、逆推 2010 年高学历或离校/上学阶段需要降级的情况，需要判断降一级或两级，即由本科降到高中或是大专。若上过大专，则降一级，否则降两级。

2.2 最高学历与离校最佳变量初步赋值

表 3 和表 4 分别展示了最高学历信息来源（代答值、自答值、逆推值）及离校/上学阶段信息来源（自答值、逆推值）的比对结果及最佳变量取值方案。最佳变量的取值原则如下：

（1） 2010 年在学的受访者，我们认为自答值的可信度最高，其次是逆推值及代答值；

（2） 对于 2010 年不在学的受访者，我们认为逆推值的可信度高于自答值及代答值。

（3） 若自答值和逆推值缺失，则用代答值补充。

表 3. 最高学历最佳变量信息来源（自答值、代答值、逆推值）与取值

信息来源对比	取值	N	%
自答值、代答值、逆推值均不缺失，且相等	相等值	18,710	43.93
自答值、代答值、逆推值三者不全相等			
自答值等于逆推值	相等值	10,437	24.51
自答值不等于逆推值，且 2010 年在上学状态	自答值	921	2.16
自答值不等于逆推值，且 2010 年不在上学状态	逆推值	2,888	6.78
自答值有值、逆推值缺失	自答值	9,463	22.22
逆推值有值、自答值缺失	逆推值	72	0.17
代答值有值，且逆推值、自答值均缺失	代答值	4	0.01
自答值、代答值、逆推值均缺失	缺失	95	0.22
总计		42,590	100.00

表 4. 离校/上学阶段最佳变量信息来源（自答值、逆推值）与取值

信息来源对比	取值	N	%
自答值等于逆推值	相等值	9,155	21.50
自答值不等于逆推值，且 2010 年在上学状态	自答值	569	1.34
自答值不等于逆推值，且 2010 年不在上学状态	逆推值	385	0.90

自答值有值、逆推值缺失	自答值	2,678	6.29
逆推值有值、自答值缺失	逆推值	22,176	52.07
自答值与逆推值均缺失	缺失	7,627	17.91
总计		42,590	100.00

2.3 最高学历与离校/上学阶段最佳变量初步赋值后的一致性分析

表 5 展示了初步赋值的最高学历和离校/上学阶最佳变量的比对结果。表 5 中对角线上方皆为零，说明离校/上学阶段不小于最高学历。对角线和对角线下方粗体字部分是离校/上学阶段等于最高学历或高于最高学历一级（或两级，大学阶段高于高中阶段、博士阶段高于硕士阶段）的情况。我们认为此类情况理论上合理，这些合理值占总观测的 99.9%。

表 5. 2010 年最高学历与离校/上学阶段最佳变量初步赋值结果比对

cfps2010sch_best	cfps2010edu_best 最高学历									
离校/上学阶段	缺失	从未上学	小学	初中	高中	大专	本科	硕士	博士	合计
缺失	95	2,179	1,439	2,132	1,071	397	295	16	3	7,627
从未上学	0	**10,340**	0	0	0	0	0	0	0	10,340
小学	0	**5,057**	**3,848**	0	0	0	0	0	0	8,905
初中	0	*4*	**3,208**	**6,255**	0	0	0	0	0	9,467
高中	0	*2*	*1*	**1,004**	**3,177**	0	0	0	0	4,184
大专	0	0	0	*23*	**216**	**951**	0	0	0	1,190
本科	0	0	0	*1*	**206**	**63**	**548**	0	0	818
硕士	0	0	0	0	0	*1*	**17**	**36**	0	54
博士	0	*1*	0	0	0	0	0	**3**	**1**	5
合计	95	17,583	8,496	9,415	4,670	1,412	860	55	4	42,590

表 5 中对角线下方斜体字部分是离校/上学阶段高于最高学历两级或以上的情况（共 33

条）。这些观测均为 2010 年在学，初步赋值为自答值。我们对这些观测做进一步处理，这 32 条疑似不合理样本中，1 条离校阶段设为缺失值，13 条用逆推值或代答值修正，18 条假定存在跳级情况并保留自答值。表 6 展示了具体处理方法：

表 6.表 5 中比对结果疑似不合理观测的进一步取值方案

最高学历对比上学阶段	频数	取值方案
逆推值缺失		
最高学历为小学以下、上学阶段为博士	1	设为缺失值
最高学历自答值等于代答值或代答值缺失	12	假定跳级，保留自答值
最高学历自答值低于代答值	6	取代答值
最高学历自答值低于代答值，代答值也属于跳级	1	假定跳级，取自答值
逆推值不缺失		
2010 年处于毕业年级、自答已完成最高学历低于当年毕业学历，2012 年逆推最高学历为 2010 年毕业学历	1	取逆推值
最高学历逆推值等于代答值且不低于上学阶段最佳变量两级或以上	6	取逆推值
无足够信息	5	假定跳级，保留自答值
总计	32	

2.4 受教育年限最佳变量的构建

在最高学历和离校/上学阶段最佳变量的基础上，我们按以下程序生成了受教育年限的最佳变量：

（1）当离校/上学阶段等于最高学历时，我们依照表 7 中的对应关系将最高学历转换为其相对应的受教育年限。

表 7. 基于最高学历的受教育年限转换表

编码	受教育程度	受教育年限（年）
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

（2）当离校/上学阶段不等于最高学历时，受教育年限变量按如下规则计算：

首先，对所有观测（无论是否跳级、是否读过大专）生成变量 1，取值方法为：变量 1=**已完成**的最高学历所对应的受教育年限，例如，如果受访者读完高中，然后读了 2 年大学，该变量取值为高中对应的 12 年；如果受访者读完大专，然后读了 2 年大学离校，该变量取值为大专对应的 15 年。

然后，对所有观测生成为变量 2，取值方法为：变量 2=与未完成阶段相邻的上一个阶段的年数+未完成阶段的已读年数。例如，如果受访者读完小学，然后读高中 2 年，那么取值为初中（而非小学）对应的 9 年+高中 2 年。在这个计算中，我们特别规定大学的上一个阶段为高中而非大专，所以不管受访者是读完高中后读大学 2 年，还是读完大专后读大学 2 年，这个变量的取值都是高中对应的 12 年+大学的 2 年。

最后，比对变量 1 与变量 2，取较大者作为最终受教育年限最佳变量的取值(cfps2010eduy),该变量在 2018 年发布的 2010-2016 跨年核心变量库中。

若受访者未完成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 cfps2010eduy 缺失。我们在生成变量 2 的过程中，使用 hot deck 方法对该变量的取值进行插补。具体来说，我们按照排序后的个人 ID（pid）找到上一位与该受访者上学或者离校阶段相同且同样没有完成该阶段的、但是该阶段已读年数不缺失的受访者，取其最后一阶段已读年数补充之，[1] 另生成插补版的教育年限（cfps2010eduy_im），对 cfps2010eduy 中有缺失的变量进行了填补。该变量在 2018 年发布的 2010-2016 跨年核心变量库中。

[1]如果上一位受访者的信息按此方法进行过插补，跳过此受访者。

三.数据评估

本节我们将基于教育变量的原始数据信息以及补充修正后的数据信息对教育数据进行综合的质量评估。

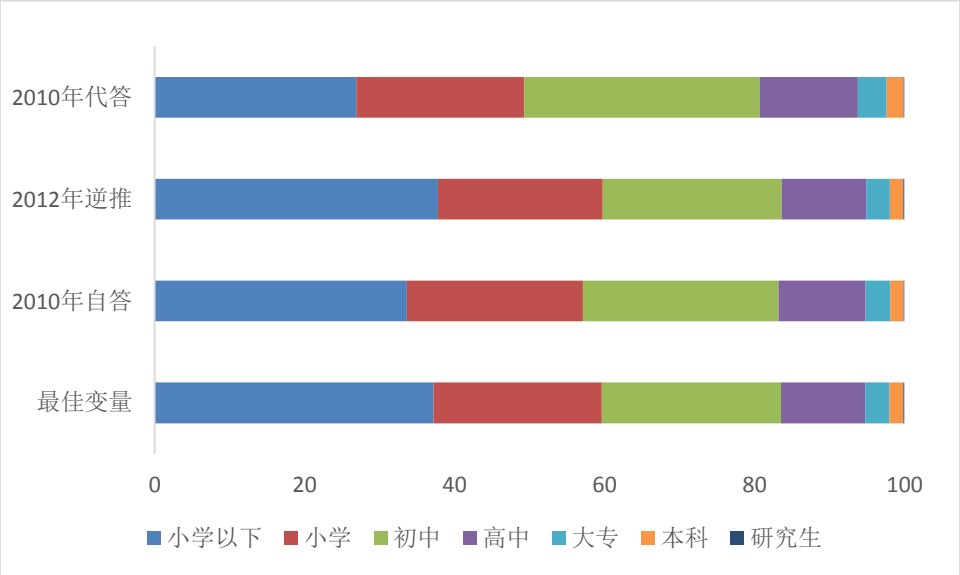
1.最高学历、离校/上学阶段测量

表 8 展示了不同信息来源的最高学历及其最佳变量的分布情况。代答值和逆推值的缺失比例较大，均在 20%以上。经过整理之后的最高学历最佳变量在小学以下阶段的分布进一步扩大，达到了 40%以上。由于只有 2012 年完成了自答问卷的受访者才有逆推值，为了比较不同信息来源之间的差异，我们关注那些同时完成了 2010 年和 2012 年自答问卷的受访者（见图 1）。由图 1 可知，在 6 岁及以上受访者中，2010 年自答的最高学历明显低于 2010 年代答的结果，高于 2012 年逆推的结果。

表 8. 2010 年个人问卷受访者最高学历最佳变量、自答值、逆推值、代答值分布

最高学历	最佳变量		自答值		逆推值		代答值	
	N	%	N	%	N	%	N	%
小学以下	17,577	41.27	16,468	38.67	14,248	33.45	8,801	20.66
小学	8,499	19.96	8,783	20.62	6,643	15.60	7,199	16.90
初中	9,412	22.10	10,062	23.63	7,204	16.91	10,422	24.47
高中	4,675	10.98	4,780	11.22	3,403	7.99	4,677	10.98
大专	1,412	3.32	1,430	3.36	947	2.22	1,429	3.36
本科	860	2.02	839	1.97	546	1.28	902	2.12
研究生	59	0.14	57	0.13	37	0.09	53	0.13
无缺失样本量	42,494	99.79	42,419	99.60	33,028	77.55	33,483	78.62
缺失样本量	96	0.23	171	0.40	9,562	22.45	9,107	21.38
总样本量	42,590		42,590		42,590		42,590	

图 1. 6 岁及以上受访者 2010 年已完成的最高学历



*注：1. 样本为 2010 年与 2012 年均接受自答问卷的受访者

2. 最佳变量、2010 年自答、2012 年逆推 N=30,142，2010 年代答 N=25,973

3. 研究生包含硕士和博士

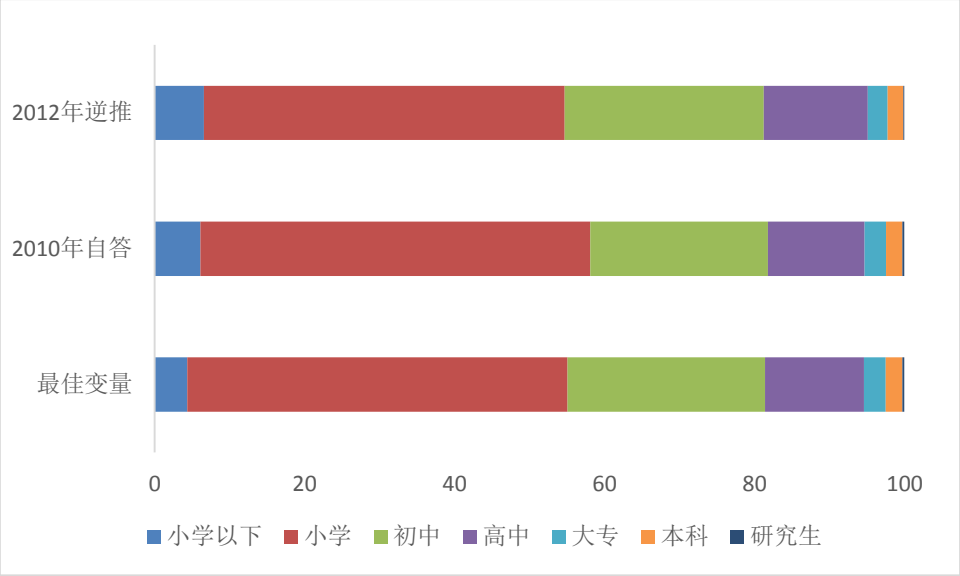
表 9 展示了不同信息来源的离校/上学阶段以及最佳变量的分布情况。图 2 展示了在基于同一样本的基础上，不同信息来源的离校/上学阶段的分布比较。相比原始数据，最佳变量确实起到了一定的修正作用。2012 年离校阶段存在大量缺失（后文将继续讨论），我们只能用最高学历的取值进行替代，对真实的离校/上学阶段存在一定程度的低估，因此，这一变量与真实的上学/离校阶段依然存在差距。

表 9. 2010 年个人问卷受访者离校/上学阶段最佳变量、自答值、逆推值、代答值分布

最高学历	最佳变量		自答值		逆推值	
	N	%	N	%	N	%
小学以下	10,340	24.28	4,139	9.72	9,619	22.59
小学	8,905	20.91	4,471	10.50	8,041	18.88
初中	9,466	22.23	2,243	5.27	8,972	21.07
高中	4,187	9.83	1,278	3.00	3,894	9.14
大专	1,186	2.78	332	0.78	1,047	2.46
本科	819	1.92	296	0.69	669	1.57
研究生	58	0.14	28	0.07	43	0.10

无缺失样本量	34,961	82.09	12,787	30.02	32,285	75.81
缺失样本量	7,629	17.91	29,803	69.98	10,305	24.20
总样本量	42,590		42,590		42,590	

图 2. 6 岁及以上受访者 2010 年离校/上学阶段



*注：1. 样本为 2010 年与 2012 年均接受自答问卷的受访者

2. 最佳变量、2010 年自答、2012 年逆推 N=7,295

3. 研究生包含硕士和博士

2. 受教育程度的测量

我们接下来分析一个设想：即上学/离校阶段相比已完成的最高学历对于受教育程度的测量更为有效。我们对比了 **2010 年离校受访者**的离校阶段和已完成的最高学历的最佳取值，发现二者在分布上存在显著的差异（见表 10），中途离校为未毕业的受访者。

表 10. 2010 年 6 岁及以上离校受访者最高学历与离校阶段分布(个人问卷，N=31,618)

	已完成最高学历（%）	离校/上学阶段（%）	中途离校 （%）
小学以下	36.02	28.00	
小学	21.41	21.34	37.58
初中	24.96	31.58	25.62

高中	11.67	13.04	11.32
大专	3.68	3.77	2.23
本科	2.12	2.14	0.93
硕士	0.14	0.14	0.00
博士	0.00	0.00	0.00

对 2010 年受访者最高学历最佳变量和离校阶段最佳变量分布的卡方检验结果为 $x^2(7) = 373.53$，$p = 0.000$。统计数据显示，在 6 岁及以上受访者中，有 31.01%的受访者的最高学历最佳变量低于离校阶段，这意味着如果仅采集最高学历的信息，我们将低估这一类受访者的真实受教育程度。这一点同样可以从分别基于最高学历与离校阶段最佳变量计算的平均受教育年限反映出来。基于最高学历的平均受教育年限是 5.82 年，基于离校阶段的平均受教育年限是 6.99 年，后者更接近 2010 年第六次人口普查的结果(8.80 年)。而且，对不同教育阶段的受访者来说，其受教育程度被低估的比重是不同的。表 10 第 4 列显示，中途离校的人群主要集中在初中及以下的低学历人群，具体而言，各有 37.58%、25.62%的受访者在小学或初中阶段中途离校，未完成学业；仅有 2%左右的受访者在大专及以上教育阶段中途离校，且受教育经历越高，中途离校的比例越低。

3.数据偏差

本部分我们将进一步探讨除测量方法以外的其他可能导致数据结果产生偏差的原因。

3.1 数据修正与受访者特征

测量误差（measurement error）是数据采集中常见的误差之一，该误差既有可能由测量工具的选择与使用所导致，也有可能来自于受访者在理解问题、回忆和处理相关信息，以及填报答案时产生的偏差。表 11 给出了我们创建的 2010 年受教育程度最佳变量的修正情况。在 42,590 名个人问卷受访者中，我们对 6.99%的样本在已完成的最高学历最佳变量上做过修正，对 2.24%的样本在离校/上学阶段最佳变量上做过修正。如果假定最佳变量是经过综合判断后得来的最准确的取值，什么样的人的答案更容易不准确而引起修正呢？我们进一步分析了这些受教育程度信息被修正的受访者的社会人口特征。

表 11. 2010 年受访者教育变量修正比例

教育变量	修正比例（%）	修正样本量	总样本量
最高学历	6.99	2,977	42,590
离校/上学阶段	2.24	954	42,590

由表 12 可知，在已完成的最高学历最佳变量中，相对于未修正过信息的受访者，教育信息被修正过的受访者在年龄、性别和婚姻状态上存在显著的差别。这一类受访者的平均年龄为 45.04 岁，比信息未修正的受访者大了 8.09 岁；女性占比高出 5.54%；在婚的比例高出近 8 个百分点。此外，对于离校/上学阶段最佳变量，信息未修正过和修正过的受访者同样在年龄、性别和婚姻状态上也存在显著的差别。

表 12. 数据修正与受访者特征

	未修正受访者	修正受访者	差值
最高学历			
年龄（年）	36.95	45.04	-8.09***
性别（%）			
男性	50.99	45.45	5.54***
女性	49.01	54.55	-5.54***
在婚比例（%）	78.95	86.65	-7.70***
外出比例（%）	0.34	0.60	-0.26
离校/上学阶段			
年龄（年）	26.39	47.40	-21.01***
性别（%）			
男性	48.12	52.80	-4.68***
女性	51.88	47.20	4.68***
在婚比例（%）	65.33	86.66	-21.33***
外出比例（%）	0.46	0.27	0.19**

注：* p<0.05, **p<0.01, ***p<0.001。

3.2 代答的样本选择性

除测量上的误差外，由无应答导致的样本选择性误差也会对数据结果产生影响。这种选择性可以部分解释 CFPS 代答值与自答值在整体分布上存在的差异。表 13 第 1 列给出了 T1 表成员最高学历最佳变量的分布，第 2 列给出了 T1 表成员中完成了个人问卷的受访者最高学历最佳变量的分布。卡方检验结果（p = 0.000）表明，个人问卷受访者与 T1 表代答的最高学历分布存在显著差异。二者之间的差异表现为，与 T1 表成员相比，CFPS 个人问卷受访者中文盲/半文盲的比例大约高出 16 个百分点，初中及以上人群所占的比例略低。以“小学”为参照组，以 T1 表数据为暴露期（exposure）的对数比率模型（log-rate model）分析结果显示，显著的差异来源于文盲/半文盲、初中至大学本科这几个组。从受教育年限来看，个人问卷受访者的平均受教育年限比 T1 表成员略低，前者比后者低 1.76 年。

表 13.　T1 表代答和最高学历最佳变量的分布

	T1 表全体成员	个人问卷受访者	个人问卷缺失者
最高学历（%）			
文盲/半文盲	24.46	41.33	19.78
小学	20.96	19.99	19.57
初中	32.02	22.16	34.32
高中	14.50	11.02	15.88
大学专科	4.71	3.33	5.83
大学本科	3.15	2.03	4.31
研究生	0.18	0.13	0.26
博士	0.02	0.01	0.05
平均受教育年限（年）	7.13	5.37	7.79
样本量	46,522	42,419	13,044

CFPS 的 T1 表代答数据让我们有机会一定程度上了解到缺失者的信息。表 13 最后一列给出了个人问卷缺失者最高学历最佳变量的分布。与个人问卷受访者相比，两者的学历分布存在显著差异（p = 0.000）。在个人问卷受访者中，小学以下人群的比例比个人问卷缺失者高出 21 个百分点，而初中及以上人群的比例大大低于后者。在平均受教育年限上，个人问

卷缺失者比个人问卷受访者的平均受教育年限高 2.42 年。这一差值同样具有统计上的显著性（p=0.000）。个人问卷的受访者与缺失者在受教育程度上存在的显著差异正体现了样本的选择性。为了验证这一点，我们进一步考察了个人问卷受访者与缺失者在其他代答变量上的差异。

表 14 给出了 T1 表成员中个人问卷受访者与缺失者在由家庭成员代答的年龄、性别、婚姻状态和外出情况上的分布差异。对样本均值的 t 检验结果表明，个人问卷受访者和缺失者在性别、年龄、婚姻状态和外出情况上都存在显著差异。总的来说，个人问卷受访者的年龄更高，比缺失者平均年长 0.67 岁；个人问卷受访者中女性比例略低于男性，而个人问卷缺失者中女性所占比重更高；在个人问卷受访者中，超过六成处于在婚状态，而个人问卷缺失者中，处于在婚状态的家庭成员比重低于个人问卷受访者；在外出情况上，几乎所有个人问卷受访者都在家中，而个人问卷缺失者中，近一半的人属于外出人员。

表 14. 个人问卷受访者与缺失者的描述性统计结果

	个人问卷受访者	个人问卷缺失者	差值
年龄（年）	37.60	36.93	0.67**
性别（%）			
男性	50.59	44.08	6.51***
女性	49.41	55.92	-6.51***
在婚比例（%）	63.08	57.39	5.69***
外出比例（%）	0.36	48.11	-47.75***

注：* p<0.05, **p<0.01, ***p<0.001。

上述数据结果表明，CFPS 2010 年个人问卷未能访到外出工作的、年轻的、非在婚状态的、高学历的家庭成员。CFPS2010 个人问卷受访者与 T1 表成员在受教育程度上的差异，很大程度上可以归因于个人问卷受访者的样本选择性问题。

3.3 缺失值

缺失值是影响数据质量的另一个重要因素。表 15 给出了 2010 年自答最高学历和 2012 年自答离校/上学阶段变量的缺失情况。个人问卷受访者共计 42,590 名，在自答最高学历变

量上，有 11 个个案缺失，仅占 0.03%；而在 T1 表全体成员（N=57,155）中，代答最高学历缺失的比例是 18.60%。

如前文所述，2012 年的自答离校/上学阶段变量，对于准确测量受访者的受教育程度具有非常重要的作用。然而，在 2012 年追踪调查中，有 9539 名受访者流失，流失的受访者的离校阶段相应缺失。此外，为了获得尽可能多的信息，2012 年追踪调查对个体受访者根据外出情况区分了长短问卷两种提问方式。其中，长问卷是向受访者本人提问，包含完整的访问信息，短问卷访问的是无法追访到的外出个体，由家庭成员代答。出于对代答问题的简洁性和可代答性的考虑，代答短问卷仅提问了个人已完成的最高学历，离校阶段和教育史信息则没有采集，因而外出家庭成员的离校阶段信息缺失比例较大。总的来说，在 2010 年个体受访者中，有 9,539 人的离校阶段缺失，占全部受访者的 22.40%。对于数据缺失的这一部分受访者，我们在对最佳变量赋值时使用 2010 年自答的信息。鉴于这一缺失比例，研究者在使用离校阶段这一变量时，需要慎重考虑缺失值的情况。

表 15.　2010 受访者教育变量缺失值分布（N= 57,155）

	缺失比例（%）
2010 代答最高学历	18.60
2010 自答最高学历	25.78
2010 自答离校/上学阶段	77.63
2012 自答离校/上学阶段	43.51

表 16 给出了这一类受访者的社会人口特征。相对于自答离校/上学阶段信息完整的受访者，信息缺失的受访者在年龄、性别分布、婚姻状态上存在显著的差别，但外出比例都很小。总的来讲，信息缺失的这一类受访者的年龄更小，男性比例略高，在婚比例更低。

表 16.2012 年自答离校/上学阶段缺失值的社会人口特征

	有效值	缺失值	差值
年龄（岁）	37.81	35.25	2.57***
性别（%）			

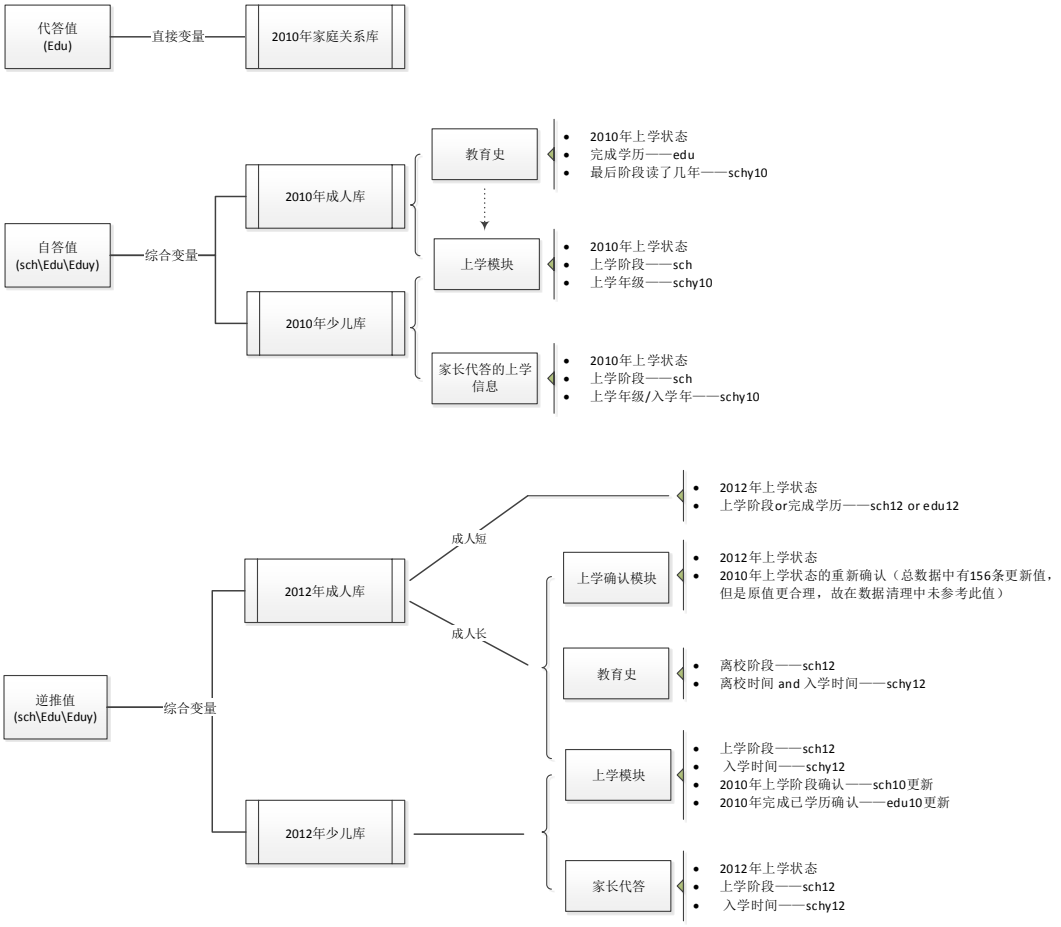
男性	48.81	53.85	-5.04***
女性	51.19	46.15	5.04***
在婚比例（%）	65.35	53.46	11.53***
外出比例（%）	0.32	27.74	-27.72***

注：* p<0.05, **p<0.01, ***p<0.001。

CFPS 数据采集过程中在不同问卷针对不同访问对象采集了教育相关信息，本报告介绍了针对基线个人问卷受访者（n=42,590）最高学历、上学/离校阶段、受教育年限这三个最佳变量的创建过程，并对创建变量的分布以及一些分布特点可能的原因进行了分析。最佳变量是 CFPS 数据管理人员基于对问卷设计及产生过程的理解所创建的综合了各种信息之后的一个较为理想的选择，我们推荐用户在一般的分析任务中使用这些最佳变量。

四．附图

三类数据来源（代答值、自答值、逆推值）及最高学历和离校/上学阶段构建结构图[2]



注：schy10 表示 2010 年调查时，受访者回答的其 2010 年离校/上学阶段的已读年数；schy12 表示 2012 年调查时，受访者回答的其 2012 年离校/上学阶段的已读年数。

[2]schy10 表示 2010 年调查时，受访者回答的其 2010 年离校/上学阶段的已读年数；schy12 表示 2012 年调查时，受访者回答的其 2012 年离校/上学阶段的已读年数。