China Family Panel Studies

CFPS 中国家庭动态跟踪调查

技术报告系列：CFPS-7

系列编辑：谢宇　责任编辑：胡婧炜

中国家庭动态跟踪调查
2010 年家庭关系数据库清理 [1]

许琪　张春泥　孙玉环　胡婧炜　吕萍

2012.12.20

[1] 特别感谢於嘉、黄国英、骆为祥、王骁、王佳、项军、武玲蔚、旎莎为家庭关系数据库清理所做的贡献，亦感谢谢宇教授、徐宏伟博士对本技术报告所提的建议。

1. 核查背景

1.1 T 表设计基本思路[2]

2010 年中国家庭动态跟踪调查（以下简称“CFPS”）对家庭成员关系和家庭成员基本信息的采集使用了 T 表设计。T 表包括 T1、T2、T3 三张表，这三张表位于家庭成员问卷的起始部分，其目的是识别家庭成员的身份、掌握家庭关系的全貌，并收集每一位家庭成员的基本社会人口信息（包括：性别、出生年/年龄、教育程度、职业、是否有行政职位、户口、居住地，等等）。其中，T1 表（同住家庭成员表）采集了每一位同住家庭成员的基本特征，“同住”以同灶吃饭来界定。该表中的家庭成员含两类：一类是三位个人编码以 1 开头的（如 101、102）与家庭有婚姻/血缘/领养关系的成员，另一类是三位个人编码以 3 开头的（如 301、302）与家庭同灶吃饭但没有婚姻/血缘/领养关系的成员（如保姆、司机、勤杂人员等）。T3 表（家庭成员不同住直系亲属列表）采集了 T1 表中 1 开头的同住家庭成员的不同住直系亲属的基本特征。T3 表中的直系亲属仅包括以婚姻/血缘/领养为基础的配偶、父母和子女三类关系，他们的三位个人编码均以 2 开头（如 201、202）。T2 表（家庭成员直系亲属关系表）将 T1 表中以 1 开头的家庭成员和 T3 表中的所有成员以血缘/婚姻/领养关系关联起来，以 T1 表中每一位 1 开头的家庭成员为核心，询问其父、母、配偶和最多 10 名子女的姓名。通过 T1 表，研究者可以知道受访家庭中每一个成员的基本特征；通过 T2 表，研究者可以知道家庭成员之间的关系和他们每一个人以血缘/婚姻/领养关系为纽带扩展出的这个家庭之外的直系亲属关系（包括父母、子女、配偶）；通过 T3 表，对即使不同住的直系亲属（包括父母、子女、配偶），研究者也可以了解其基本特征。T 表不是由每一位家庭成员和亲属亲自填答，而是访员入户后，在受访家户中选择一名家庭成员作为 T 表的代答人。

1.2 T 表设计的创新之处

在 T 表使用以前，问卷调查在收集父母、子女、配偶信息时常见的做法是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。这样的做法有四个缺陷：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，只提问与该受访者相对应的亲属关系，如提问这名受访者的父亲的情况、母亲的情况等。这种做法假定了家庭关系中只有一个中心，

[2] 关于 T 表的设计理念与操作方案还可以参考谢宇，2010，《中国家庭动态跟踪调查（2010）用户手册》。

即受访者本人，而对其他成员均只记录他们与受访者的关系。但在调查中应选择哪一位成员作为家庭关系的中心呢？又如何保证选中的受访者即是家庭关系的中心呢？这是以往的调查没有解决的问题。事实上，家庭中的每一个人都可以作为家庭关系的中心，家庭关系应该是由多个中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那些家庭成员/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“父亲”、“母亲”、“配偶”、“子女”的情况，不询问这些亲属的姓名，也没有专门的编号来识别这些亲属。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。T 表的产生解决了这些问题：T2 表采用“轮流坐庄”的方式，通过将每一个人轮流视为家庭关系的核心（或“庄主”），收集其所有家庭成员及亲属的姓名，并予以编码，因而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、跨代的关系均可以关联。此外，T 表采用代答的方式，不需要所有家庭成员都在场即可以收集到他们的信息。

1.3 T 表信息采集与使用的难度

尽管 T 表有更高的研究价值，但在执行操作上却对访员有较高的要求。理想的 T 表填写应以血缘/领养/婚姻纽带来判断父代与子代的关系以及配偶关系，应遵循父母-子女互认、夫妻互认的填答对称原则。每一个人的父母不应该为缺失，即使不知道父母的名字，也应以“XXX 的父亲”、“XXX 的母亲”等命名方式替代；如果一方填写了配偶，则其配偶在相应的配偶位置上也应该填写此人。但在实际操作中，有的访员以血缘关系界定子女的父母，有的访员则以婚姻关系界定子女的父母，后一种情况有可能导致继父母与其继子女的年龄差小于 16 岁或父母-子女不互认等情况。例如，在一个有继父-继子关系的家庭中，当继子“坐庄”时，其父亲位置上填写的是他的亲生父亲；而当继父“坐庄”时，其孩子位置上又填写的是该继子，由此产生了父母-子女不互认的问题。此外，或由于访员操作的失误，或由于部分家庭的成员关系过于复杂，或由于代答者不熟悉家庭情况（如女婿作为代答人时，有可能不知道岳父的父母的姓名、年龄等）等原因，会导致家庭成员姓名缺失、家

庭关系匹配错误、家庭成员信息填答错误等情况出现，从而也会产生逻辑上的一些问题。

总的来说，与 T 表有关的错误主要有两类：其一是匹配错误，即由于 T2 表家庭关系匹配不正确或匹配不上而表现出来的逻辑矛盾，比如一男性成员的儿子被误填在了其配偶的位置上，则会出现本人与配偶性别相同的错误；其二是信息错误，即由于受访者信息填报出错或访员录入错误而表现出来的逻辑矛盾，如某一男性成员的性别被填报为“女”，也会出现该成员与其配偶性别相同的错误。

由于 T 表本身具有一定的复杂性，加上上述操作过程中出现的一些问题，使得 T 表在使用上有一定的难度。首先，家庭成员的信息不能直接在个人问卷的数据库中使用，而需要从家庭关系数据库中调用 T2 表，并编写程序将相应的关系匹配上后才能使用。[3] 其次，在规模较大、关系较复杂的家庭中，由于受访者填报或访员录入出错导致的 T2 表中家庭关系填写不准确，使得研究者在使用数据时会发现由于关系匹配错误或匹配不上而造成的逻辑矛盾。如，某男性成员的父亲被误填在了其孩子的位置，则出现孩子年龄比父母年龄大的错误。再次，由于同时存在 T 表代答信息和个人问卷的自答信息，有可能出现同一信息的不同来源填答不一致或相互矛盾的错误。比如，某成员在 T 表中的婚姻状况被代答人填作“未婚”，但该成员在其个人问卷中自己填报的婚姻状况是“在婚”，这样，研究者就不得不在两个来源的填答中做出取舍。

为了提高 T 表的准确性，我们对已收集的 T 表信息进行了逻辑核查，并对已发现的错误进行了更正。本报告第二部分将介绍核查标准，第三部分将介绍人工更正方案。此外，报告在第四部分还介绍了数据清理后的回访计划，并在第五部分对 T 表操作提出了改进建议。

2. 核查标准

本次核查依据的是分解后的个人层面数据库，该数据库是以每一位 T1 表成员作为一条观测，以该成员在 T2 表中的亲属姓名、代码及该亲属在 T1 或 T3 表中的基本特征作为变量形成的家庭关系数据库。[4] 核查的具体过程是在家庭关系数据库中通过频数统计、交互表、条件筛选、本人与父

[3] 为了给研究者提供方便，我们已经将每一位个人问卷受访者在 T 表中的全部家庭成员信息匹配到其个人问卷数据库中。具体可参见孙玉环、谢宇、胡婧炜、张春泥、许琪、黄国英，2012，《中国家庭动态跟踪调查 2010 年家庭关系原始数据库的分解与匹配（CFPS-6）》。

4 CFPS 家庭关系数据库形成的具体方法详见孙玉环、谢宇、胡婧炜、张春泥、许琪、黄国英，2012，《中国家庭动态跟踪调查 2010 年家庭关系原始数据库的分解与匹配（CFPS-6）》。

母及子女年龄相减等方法查找含有异常值的个案，并将这些个案的个人代码及家庭代码生成清单。

本次核查所针对的是 T 表中的家庭关系，涉及的核查标准共 28 类，如表 1 所示。这里之所以用核查标准而不用错误类型，意在表明表中列出的 28 类并不都是错误，如“本人年龄与父母年龄相差小于 15 岁或者大于 50 岁”这一条，对于早育或晚育的个别人来说可能并没有错。我们之所以将这些情况也列进去是因为这些案例不符合常识，有出错的可能性。

表 1. 数据清理核查标准

编号	核查标准	涉及户数	级别
1	父亲性别为女	354	1
2	母亲性别为男	257	1
3	本人的性别与配偶的性别相同	232	1
4	未婚离婚或丧偶者有配偶信息	0	1
5	配偶健在但是婚姻状况为丧偶	0	1
6	配偶不健在但是婚姻状况为已婚或同居	64	1
7	家庭关系代码重复	79	1
8	由于配偶编码重复导致不同家庭成员的配偶相同	80	1
9	初婚配偶（双方均为初婚）各自填报的孩子总数不一致	246	1
10	将本人与其配偶匹配后，本人代码与配偶的配偶代码不一致	279	1
11	未婚但子女信息有效	90	2
12	一户之中有重名	457	2
13	本人的父亲不认本人是他的子女	379	2
14	本人的母亲不认本人是她的子女	461	2
15	孩 1 年龄与父母年龄相差小于 15 岁或者大于 50 岁	233	3
16	孩 2 年龄与父母年龄相差小于 15 岁或者大于 50 岁	130	3
17	孩 3 年龄与父母年龄相差小于 15 岁或者大于 50 岁	56	3
18	孩 4 年龄与父母年龄相差小于 15 岁或者大于 50 岁	25	3
19	孩 5 年龄与父母年龄相差小于 15 岁或者大于 50 岁	18	3
20	孩 6 年龄与父母年龄相差小于 15 岁或者大于 50 岁	7	3
21	孩 7 年龄与父母年龄相差小于 15 岁或者大于 50 岁	2	3
22	孩 8 年龄与父母年龄相差小于 15 岁或者大于 50 岁	10	3
23	孩 9 年龄与父母年龄相差小于 15 岁或者大于 50 岁	3	3
24	孩 10 年龄与父母年龄相差小于 15 岁或者大于 50 岁	0	3
25	本人年龄与父母年龄相差小于 15 岁或者大于 50 岁	426	3
26	本人年龄在 20 岁以下，但是配偶信息有效	77	3
27	本人年龄在 20 岁以下，但是子女信息有效	43	3
28	本人年龄在 20 岁以下，婚姻状况为已婚、同居、离婚或丧偶	86	3

表中级别为 1 的是一定要修改的类型(如“父亲性别为女”)，共有 10 项。这些项目有明显的逻辑错误，所以一定要修改。级别为 2 的是基本确定要修改的类型，共有 4 项。以“未婚但子女信息有效”为例，未婚生育在中国非常罕见，如果数据中出现则需要仔细推敲是否出错，但是也不排除其真实存在的可能性，所以视为“基本确定要修改”。剩下的 14 项级别为 3 的是必须在证据确凿的情况下才修改的类型，这 14 项都与年龄有关，如“本人年龄与父母年龄相差小于 15 岁或者大于 50 岁”等。

在 1 级核查标准涉及的类型当中，性别错误所占比重最高。“父亲性别为女”、“母亲性别为男”、“本人的性别与配偶的性别相同”涉及的家户数量分别为 354、257 和 232；夫妻、父母-子女匹配错误也是较常见的错误类型，其中，“本人代码与配偶的配偶代码不一致”涉及的家户共有 279 户，“初婚配偶的子女数量不一致”涉及的家户共有 246 户。在 2 级核查标准当中，父母与子女不互认是最常见的问题，其中“本人的父亲不认本人是他的子女”涉及 379 户家庭，“本人的母亲不认本人是他的子女”涉及 461 户家庭；另外，“一户之中有重名”的家户也很多，共 457 户。在 3 级核查标准中，“本人年龄和父母年龄相差小于 15 岁或者大于 50 岁”涉及的家户数量最多，共 426 户。

粗略汇总（即简单加总）来看，1 级核查标准涉及的问题家户总共有 1,591 户，2 级核查标准涉及的问题家户总共有 1,387 户，3 级核查标准涉及的问题家户总共有 1,116 户，三者合计总共涉及的问题家户总数为 4,094 户。

但是，表 1 中所列的 28 项核查标准并不是互斥的，也就是说，同一个家户可能同时出现几类错误，因此以上数字有可能被重复统计。如果对那些同时存在多类错误的家户只统计一次，这样汇总得到的问题家户总数为 2,511 户，其中第 1 级和第 2 级核查标准涉及的问题家户总数为 1,767 户。2010 年初访总共访问到的户数为 14,960 户，所有问题家户占总家户的比例（错误率 1）为 16.8%。若只计算第 1 级和第 2 级核查标准涉及的问题家户，则问题家户所占的比例（错误率 2）为 11.8%。

分省统计的错误率 1 和错误率 2 结果见表 2。从表中可以看出，北京、山东、吉林、天津和辽宁的错误率相对较低，表明这几个省的 T 表数据质量较高。而福建、湖北、重庆、四川和贵州的错误率相对较高，表明这几个省的 T 表数据质量较差。

分问卷版本统计的错误率 1 和错误率 2 结果见表 3。总体来看，无论是错误率 1 还是错误率 2 都呈现出随版本号时间的往后推移而逐渐降低的趋势，这表明时间越靠后的版本数据质量越好。相

比其他几个版本的数据来看，4月1日版本的数据质量最差，其错误率1为25.25%，错误率2为20.26%，明显高于其他几个版本。

表 2. 分省统计的错误率（%）

省份	错误率 1	错误率 2	省份	错误率 1	错误率 2
北京	8.57	5.71	广东	15.18	11.38
山东	8.74	4.89	江西	16.48	10.26
吉林	9.29	4.49	陕西	19.19	15.82
天津	10.75	8.6	甘肃	19.86	11.99
辽宁	11.48	7.12	河南	21.42	16.35
黑龙江	13.05	8.82	广西	21.65	16.49
河北	13.61	8.89	云南	21.71	14.99
湖南	13.7	10.43	福建	23.31	18.4
浙江	14.01	10.51	湖北	24.74	19.59
江苏	14.24	8.33	重庆	25.14	20.11
山西	14.35	8.55	四川	25.32	18.35
上海	14.42	11.77	贵州	27.55	18.22
安徽	14.81	12.12			
总计	**16.78**	**11.81**			

表 3. 分版本号统计的错误率（%）

版本号	错误率 1	错误率 2
4 月 1 日	25.25	20.26
4 月 16 日	15.49	12.68
5 月 4 日	15.97	11.33
6 月 8 日	14.40	9.25
合计	16.78	11.81

3. 错误判断与更正方法[5]

本次数据更正采用的是逐户人工判断和手工更正的方法，即在家庭层面查找上述核查标准所涉及的问题家户中可能存在的所有错误，然后按照逻辑关系人工判断应如何更正错误，并将更正方案写成 SAS 程序命令进行更正。下面逐一说明每一种错误的可能原因及修改方案。

[5] 此部分示例中的家庭成员姓名均为化名。

3.1 性别错误

性别错误包括三类：父亲性别为女、母亲性别为男、本人性别与配偶性别相同。这三类错误在逻辑上不应该存在，所以一定要修改。在修改过程中，我们发现性别错误可能由两种情况导致：第一种情况是匹配错误，其中将父亲填在母亲处、将母亲填在父亲处、将子女填在配偶处是最常见的三种匹配错误；第二种情况是填写错误，即由于受访者填报错误或访员录入错误而导致父亲、母亲或配偶的性别出错。下面分别举例说明：

3.1.1 匹配错误

表 4[6] 中 410211 这一户在家庭关系库中显示的 103、104 和 105 的父亲都是女，而母亲都是男。仔细看该户的家庭关系可以发现 101 和 102 是夫妻，生了三个孩子分别是 103、104 和 105。这一户的三个子女都姓王，按照中国人的一般命名原则——子女随父亲姓，判断王君（102）是父亲，任安（101）是母亲。而从数据中看，103、104 和 105 的父母位置恰好填反了，从而导致这三个子女的父母性别均出现错误，因而需要将父母的位置对调过来。

表 4. 父母匹配错误

hhno	410211	410211	410211	410211	410211
name_a_p	任安	王君	王礼明	王艳	王宏
name_a_f	任选	王永献	***任安***	***任安***	***任安***
name_a_m	杨凤	张兰花	***王君***	***王君***	***王君***
name_a_s	王君	任安	-8	-8	-8
name_a_c1	王礼明	王礼明	-8	-8	-8
name_a_c2	王艳	王艳	-8	-8	-8
name_a_c3	王宏	王宏	-8	-8	-8
code_a_p	101	102	103	104	105
code_a_f	201	203	***101***	***101***	***101***
code_a_m	202	204	***102***	***102***	***102***
code_a_s	102	101	-8	-8	-8
code_a_c1	103	103	-8	-8	-8
code_a_c2	104	104	-8	-8	-8
code_a_c3	105	105	-8	-8	-8

注：此表及以下各表中的名字均为化名。

[6] 表 4 中 name 代表个人名字，code 代表个人编码。其中，p 代表本人，m 代表母亲，s 代表配偶，c1-c10 分别代表孩 1-孩 10。下同。

另一种常见的匹配错误是将子女填在了父母的位置上。表 5 中 410324 一户的问题是 103 的性别和配偶的性别相同，都是男性。仔细看家庭关系可以发现，该户有 7 口人，101 和 102 是一对老夫妻；103 是其儿子，107 是其儿媳；104、105 和 106 是其孙子女。该户的问题在于 103 的配偶应该是 107，而不是 104，104 事实上是 103 的儿子。我们做出这样的判断的依据是：104、105 和 106 都承认 103 是其父亲，107 是其母亲，可见 103 和 107 是一对夫妻；且从 107 处也可以看出其丈夫为 103，子女为 104、105 和 106；进一步从年龄也可以发现 103 和 107 年龄相仿，而与 104 却相差约一代人，这更证实了我们的判断。综上可以得出结论：103 错将其儿子填到了配偶处，导致本人性别和配偶性别相同。修改方案是将 103 的配偶改成 107，并将 104 作为其子女补充到孩 3 的位置。

表 5. 子女填写到了配偶处

hhno	410324	410324	410324	410324	410324	410324	410324
name_a_p	李萌	王铮	王淮栋	王晟	王伊	王琪	王璇沁
name_a_f	李韧	王钟云	王铮	***王淮栋***	***王淮栋***	***王淮栋***	王嘉威
name_a_m	姚莞旬	邓	李萌	***王璇沁***	***王璇沁***	***王璇沁***	张可
name_a_s	王铮	李萌	***王晟***	-8	-8	-8	***王淮栋***
name_a_c1	王淮栋	王淮栋	***王伊***	-8	-8	-8	***王晟***
name_a_c2	王连栋	王连栋	***王琪***	-8	-8	-8	***王伊***
name_a_c3	-8	-8	***-8***	-8	-8	-8	***王琪***
code_a_p	101	102	103	104	105	106	107
code_a_f	201	204	102	***103***	***103***	***103***	206
code_a_m	202	205	101	***107***	***107***	***107***	207
code_a_s	102	101	***104***	-8	-8	-8	***103***
code_a_c1	103	103	***105***	-8	-8	-8	***104***
code_a_c2	203	203	***106***	-8	-8	-8	***105***
code_a_c3	-8	-8	***-8***	-8	-8	-8	***106***

除了这两种最常见的匹配错误，还有一些其他类型的匹配错误，比如将其他亲属填写到父母或配偶的位置上，此处不再一一举例。匹配错误的判断依据首先是家庭关系，如父母与子女要互认、夫妻双方要互认、子女填写的父母要互认为夫妻等；其次是年龄，如夫妻双方年龄应相差不大、父母与子女年龄应相差一代人左右等；此外还有姓名，如子女的姓氏一般与父亲相同、取名是否有明显的性别特征等。匹配错误的修改方案是，根据多方面的信息作出判断以后调整家庭关系，使性别错误不再出现。

3.1.2 性别填写错误

表 6. 非同住家庭成员性别填错

hhno	440842	440842	440842	440842
name_a_p	郑历	刘蓉	郑雪	郑文
name_a_f	郑率隆	***刘瑜***	郑历	郑历
name_a_m	姚娴	-8	刘蓉	刘蓉
name_a_s	刘蓉	郑历	-8	-8
name_a_c1	郑雪	郑雪	-8	-8
name_a_c2	郑文	郑文	-8	-8
code_a_p	101	102	103	104
code_a_f	201	***203***	101	101
code_a_m	202	-8	102	102
code_a_s	102	101	-8	-8
code_a_c1	103	103	-8	-8
code_a_c2	104	104	-8	-8

表 7. 同住家庭成员的性别填错

code_hhno	440942	440942	440942	440942
name_a_p	吴华	***冯苏森***	冯艳	冯钢
name_a_f	吴唯武	冯发	冯苏森	冯苏森
name_a_m	吕萱	郭梅	吴华	吴华
name_a_s	冯苏森	吴华	李祥	-8
name_a_c1	冯艳	冯艳	-8	-8
name_a_c2	冯钢	冯钢	-8	-8
code_a_p	101	***102***	103	104
code_a_f	201	203	102	102
code_a_m	202	204	101	101
code_a_s	102	101	205	-8
code_a_c1	103	103	-8	-8
code_a_c2	104	104	-8	-8

如果发现家庭关系没有错误，那么性别错误就只可能是性别信息本身的填答错误造成。这种情况在父母或配偶非同住的情况下较为常见。

表 6 中 440842 一户的错误是 102 的不同住的父亲 203 性别为女。检查该户的家庭关系发现，101 和 102 是一对夫妻，育有两个子女 103 和 104，匹配关系没有问题；从年龄看，102 和 203 的年

龄差没有异常；从姓名看，102 和 103 的姓氏相同，符合常识。所以，我们基本可以判断这个错误出现的原因是 203 的性别填写错了。修改方案是将 203 的性别改为女。

但在有些情况下性别错误则是由于同住家庭成员的性别填错导致的。表 7 中 440942 一户的问题是 101 和 102 的性别都与配偶性别相同，103 和 104 的父亲性别都为女。检查家庭关系发现，该户有 4 口人，101 和 102 是夫妻，103 和 104 是其子女，家庭关系匹配没有问题。而错误的来源是 102 的性别填写错误——数据中 102 是女，若将其改为男，则所有问题均迎刃而解。所以修改方案是将 102 的性别改为男。

3.2 父母与子女年龄差距异常

表 1 中的第 15 至第 25 类核查标准均属于这个范畴。一般来说，父母生育子女的年龄应在 15-50 岁之间，若超过这个范围则很有可能存在问题。

父母与子女的年龄差距异常可能由四种情况导致：第一，匹配错误；第二，年龄填错；第三，属于再婚重组家庭的继父母和继子女关系；第四，确实早育或晚育。

3.2.1 匹配错误

表 8. 匹配错误导致父母与子女年龄差距异常

hhno	130627	130627
name_a_p	孟子龙	孟飞
name_a_f	孟备	孟子龙
name_a_m	孙月英	***孙月英***
name_a_s	周甄	-8
name_a_c1	孟飞	-8
name_a_c2	孟乔乔	-8
code_a_p	101	102
code_a_f	201	101
code_a_m	202	***202***
code_a_s	203	-8
code_a_c1	102	-8
code_a_c2	204	-8

表 8 所示的 130627 一户中，102 同其母亲的年龄相差大于 50 岁。从家庭关系可以看出， 102 是 101 的子女，101 的配偶是 203，所以 102 的母亲应该是 203，但数据中却是 202。实际上 202 是 101 的母亲，即 102 的奶奶。所以 102 与母亲年龄相差大于 50 岁是因为将奶奶 202 填到了母亲处，而其母亲应为 203。修改方案是将 102 的母亲改为 203。

3.2.2 年龄信息错误

如果从家庭关系看没有匹配错误，即父母与子女互认、夫妻双方互认、子女填写的父母也互认为夫妻，那么这时进行修改就只能从年龄本身入手。此处又分为几种情况：

⑴ 不同来源的年龄信息不一致

年龄除了可以从家庭关系库中获得，还可以从个人库中得到。个人库中的年龄由受访者亲自填答，而家庭关系库中是代答，所以一般来说，个人库中的年龄更准确。基于此，若根据家庭关系库中的年龄信息判断父母和子女年龄之差有异常，而根据个人库中的年龄来推算的年龄差却是正常的，那么我们就以个人库中的年龄为准，修改家庭关系库中的年龄。

表 9. 年龄填错导致的父母与子女年龄差距异常

hhno	410640	410640	410640	410640	410640
name_a_p	董宝琴	赵敬铭	赵正	***任凤***	赵巧巧
name_a_f	董隐	赵毅一	赵敬铭	-8	赵正
name_a_m	杨莲	丁香	董宝琴	向敏	任凤
name_a_s	赵敬铭	董宝琴	任凤	赵正	-8
name_a_c1	赵雯	赵雯	赵巧巧	赵巧巧	-8
name_a_c2	赵正	赵正	-8	-8	-8
code_a_p	101	102	103	***104***	105
code_a_f	201	204	102	-8	103
code_a_m	202	205	101	206	104
code_a_s	102	101	104	103	-8
code_a_c1	203	203	105	105	-8
code_a_c2	103	103	-8	-8	-8

在表 9 所示的 410640 这一家户中，104 与其母亲的年龄差小于 15 岁。通过分析其家庭结构，

我们并没有发现匹配问题。在家庭关系库中，104 生于 1970 年，调查时 40 岁；而在个人库中，104 生于 1975 年，调查时为 35 岁。一般来说，个人库中的年龄为受访者本人填答，应更准确一些；而且若将其家庭关系库中的年龄改为个人库中的 35 岁，则发现与母亲年龄差扩大为 17 岁，不再显示出异常。所以我们判断，该处年龄异常是由于家庭关系库中 104 的年龄填写错误导致的。修改方案是将家庭关系库中的年龄改为个人库中的年龄，并同时修改出生年。

⑵ 只有家庭关系库中的年龄信息有效

这种情况很普遍，且可能由多种原因导致。比如受访者的父母或子女没有与其同住；受访者的同住父母或子女虽然与其同住，但没有填写个人问卷；受访者本人没有填写个人问卷；等等。

若个人库中没有找到相关的年龄信息，则一般保留家庭关系库中的原始状况，不做任何修改。但是如果父母与子女年龄相差太大（10 岁以下以及 60 岁以上），我们会将不同住的父母或子女一方的年龄改为缺失。

表 10. 年龄填错导致的父母与子女年龄差距异常

hhno	410633	410633	410633
name_a_p	董明浩	沈文珊	董琼
name_a_f	董村	董氏	董明浩
name_a_m	文婵	***李薇***	沈文珊
name_a_s	沈文珊	董明浩	-8
name_a_c1	董琼	董琼	-8
code_a_p	101	102	103
code_a_f	201	203	101
code_a_m	202	***204***	102
code_a_s	102	101	-8
code_a_c1	103	103	-8

表 10 中 410633 这一户无匹配错误，但是 102 在家庭关系库中的年龄是 42 岁（在个人库中的年龄也是 42 岁），而其母亲 204 在家庭关系库中的年龄为 46 岁，二者相差仅 4 岁，非常不合理。由于其母亲并不同住，没有个人问卷，无法从个人库中获得其母亲的年龄，所以无法进行准确的修改。但是鉴于年龄差异过于严重，我们将其母亲的年龄改为了缺失值“-9”。

3.2.3 继父母-继子女关系造成的年龄差距不合理

表 11 中的 411429 这一户家庭存在的问题是 102 与孩 1 年龄差距不合理。102 是 1974 年出生，没有个人库信息，其孩 1（103）是 1984 年出生，两人年龄差距过小。但我们在个人库中查到 102 的配偶 101 是再婚，其初婚是在 1985 年，与 102 再婚是在 1997 年，因此，103 应是其父 101 与前妻所生，后带入重组家庭中。

针对由于继父母-继子女关系造成的年龄差距不合理的情况，我们一般的做法是按照血亲关系进行调整。[7] 此例中应把 103 从 102 的孩子位置上去掉，并将 102 从 103 的母亲位置上去掉。

表 11. 继父母-继子女关系造成的年龄差距不合理

hhno	510012	510012	510012
name_a_p	仲夏	***刘慈***	仲若
name_a_f	仲致远	-8	仲夏
name_a_m	张芙	-8	刘慈
name_a_s	刘慈	仲夏	-8
name_a_c1	仲若	***仲若***	-8
code_a_p	101	102	103
code_a_f	201	-8	101
code_a_m	202	-8	102
code_a_s	102	101	-8
code_a_c1	103	***103***	-8

对继父母-继子女关系造成的父母-子女年龄差距不合理的问题进行更正时一定要慎重，须是在父母婚姻状况已知、父母初婚及再婚年份均已知、孩子出生年份已知、孩子的出生年份明显接近于亲生父/母的初婚年份且明显早于父母的再婚年份的情况下才做更正。

3.2.4 早育或晚育

我们对早婚早育的判断通常采用的标准是：在没有匹配错误、不存在重组家庭问题、在已填写的年龄信息确认一致或合理的情况下，如果父母生育孩子的年龄稍晚于 15 岁，可判断为早育，如果父亲一方生育孩子的年龄稍大于 50 岁，可判断为晚育，对这些情况不作修改。

[7] 我们仅对表 1 中 28 类核查标准所涉及的家户进行了这一调整。

3.3 20 岁以下结婚或生育

这里包括三类核查标准：本人的年龄 20 岁以下，但是配偶信息有效；本人的年龄 20 岁以下，但是子女信息有效；本人的年龄 20 岁以下，婚姻状况为已婚、同居、离婚或丧偶。我国 1950 年《婚姻法》规定的法定结婚年龄为男 20 岁、女 18 岁，1981 年《婚姻法》规定男 22 周岁以上、女 20 周岁以上准许登记结婚，所以若数据中个体年龄在 20 岁以下却已经结婚生育，可视为异常。但是异常并不代表错误，因为在一些少数民族地区以及农村地区，早婚早育现象依然存在。

有四种情况会导致 20 岁以下结婚或生育这一问题出现：第一，匹配错误；第二，年龄填错；第三，婚姻状况填错；第四，确实早婚。只有前三种情况才需要修改，最后一种不修改。

3.3.1 匹配错误

表 12. 匹配错误导致的早婚早育

hhno	620751	620751	620751	620751	620751	620751	620751
name_a_p	王冬梅	张家靖	张盈盈	张青青	张晖云	张宪学	张国
name_a_f	王鸿筹	张国	张家靖	张家靖	张家靖	张家靖	张
name_a_m	马氏	许如兰	王冬梅	王冬梅	王冬梅	王冬梅	李
name_a_s	张家靖	王冬梅	-8	-8	-8	-8	-8
name_a_c1	张盈盈	张盈盈	-8	-8	-8	***张盈盈***	张家妮
name_a_c2	张青青	张青青	-8	-8	-8	-8	张家靖
name_a_c3	张晖云	张晖云	-8	-8	-8	-8	张平川
name_a_c4	张宪学	张宪学	-8	-8	-8	-8	张佩佩
name_a_c5	-8	-8	-8	-8	-8	-8	张俏妮
name_a_c6	-8	-8	-8	-8	-8	-8	张家虎
code_a_p	101	102	103	104	105	106	107
code_a_f	201	107	102	102	102	102	204
code_a_m	202	203	101	101	101	101	205
code_a_s	102	101	-8	-8	-8	-8	-8
code_a_c1	103	103	-8	-8	-8	***103***	206
code_a_c2	104	104	-8	-8	-8	-8	102
code_a_c3	105	105	-8	-8	-8	-8	207
code_a_c4	106	106	-8	-8	-8	-8	208
code_a_c5	-8	-8	-8	-8	-8	-8	209
code_a_c6	-8	-8	-8	-8	-8	-8	210

表 12 所示的 620751 这一户中，106 的年龄为 18 岁，但是却有一个孩子 103。仔细比较家庭关系可以发现，103 的父母是 102 和 101，而且 102 和 101 也认 106 做小孩，所以 103 不是 106 的小孩，而是其兄弟姐妹。修改方案是删去 106 的小孩 103。

3.3.2 年龄填错

同样，我们可以比较个人库和家庭关系库中的年龄是否一致，若不一致，则可以根据个人库中的年龄校正家庭关系库中的年龄。

表 13 中所示的 510664 一户没有匹配错误（注：101 结过两次婚，其子女 201 和 202 可能是其与前夫所生，不算匹配错误），但是 102 的年龄是 0 岁，显然不合理。而在个人库中，102 的年龄是 58 岁（其配偶 101 是 61 岁，二者相差不大，比较可信）。所以修改方案是将家庭关系库中的年龄替换为个人库中的年龄。

表 13. 年龄填错导致的早婚早育

hhno	510664	510664	510664	510664
name_a_p	温婷	***肖诺***	肖晓	肖林
name_a_f	-8	-8	肖文钧	肖文钧
name_a_m	-8	-8	-8	-8
name_a_s	肖诺	温婷	-8	-8
name_a_c1	刘鹏	肖文钧	-8	-8
name_a_c2	刘娟	-8	-8	-8
name_a_c3	肖文钧	-8	-8	-8
code_a_p	101	102	103	104
code_a_f	-8	-8	203	203
code_a_m	-8	-8	-8	-8
code_a_s	102	***101***	-8	-8
code_a_c1	201	203	-8	-8
code_a_c2	202	-8	-8	-8
code_a_c3	203	-8	-8	-8

3.3.3 婚姻状况填错

与年龄类似，婚姻状况也同样有家庭关系库和个人库两个来源，而且个人库中的婚姻状况为受

访者本人填答，因而更为准确。所以若个人库中的婚姻状况与家庭关系库不一致，可以根据个人库中的婚姻状况更正家庭关系库中的婚姻状况。

表 14 所示的 370526 一户无匹配错误。在家庭关系库中，103 为 16 岁，但婚姻状况为已婚。而在个人库中，103 的婚姻状况为未婚，一般来说个人库中的信息可信度更高；而且从 103 的年龄来看，其已婚的可能性不大，所以我们判断 103 在家庭关系库中婚姻状况填错，将其修改为未婚。

表 14. 婚姻状况填错导致的早婚早育

hhno	370526	370526	370526
name_a_p	李小凡	吕成峰	***吕莫***
name_a_f	李泽	吕藩	吕成峰
name_a_m	王迪	-8	李小凡
name_a_s	吕成峰	李小凡	-8
name_a_c1	吕莫	吕莫	-8
code_a_p	101	102	103
code_a_f	201	203	102
code_a_m	202	-8	101
code_a_s	102	101	-8
code_a_c1	103	103	-8

3.4 未婚、离婚或丧偶者有配偶信息

根据 2010 年问卷的设计，未婚、离婚或丧偶者不填写配偶信息。为了检查这一原则的落实情况，故设计了这条核查标准，在检查过程中没有发现问题。

3.5 未婚但子女信息有效

在中国未婚生育的可能性很小，所以若数据中出现，则可视为异常。导致这种异常的可能性有三种：第一，子女匹配错误；第二，婚姻状况错误；第三，未婚生育或领养子女。只有前两种情况才修改，最后一种情况不修改。

3.5.1 匹配错误

表 15 中的 210750 一户有两处匹配错误：第一，103 的配偶应是 105，但是在数据中其子女 106 被填在了配偶处。将子女填在配偶处是一种很常见的匹配错误，前面已经举例，此处不再做重点说明。第二，数据中显示 104 婚姻状况为未婚，但却有子女 106。仔细观察家庭关系可以发现，106 的父母是 103 和 105，而 104 是其叔叔，所以 106 并不是 104 所生，应将其从 104 的子女位置去掉。

表 15. 匹配错误导致未婚有子女

hhno	210750	210750	210750	210750	210750	210750
name_a_p	王裕程	于璧	王复	王竹	高灵	王玉梓
name_a_f	王昌喜	于顺民	王裕程	王裕程	高一鸣	王复
name_a_m	刘秀秀	岳娥	于璧	于璧	顾真真	高灵
name_a_s	于璧	王裕程	***王玉梓***	-8	王复	-8
name_a_c1	王复	王复	***-8***	***王玉梓***	王玉梓	-8
name_a_c2	王竹	王竹	-8	-8	-8	-8
code_a_p	101	102	103	104	105	106
code_a_f	201	203	101	101	205	103
code_a_m	202	204	102	102	206	105
code_a_s	102	101	***106***	-8	103	-8
code_a_c1	103	103	***-8***	***106***	106	-8
code_a_c2	104	104	-8	-8	-8	-8

3.5.2 婚姻状况错误

婚姻状况填错是导致未婚有子女的主要原因，前面已经提到，婚姻状况既可以从家庭关系库获得，也可以从个人库获得，所以若个人库中的婚姻状况与家庭关系库中不一致，可以根据个人库中的婚姻状况校正家庭关系库的婚姻状况。

表 16 所示的 210414 一户中，103 的婚姻状况为未婚，但是却有一个子女 104。从家庭关系可以看出，104 的父母是 102 和 103，子女和父母是互认的，没有匹配问题，而且个人库中 103 的婚姻状况为在婚，由此可以推断，问题出现的原因是家庭关系库中 103 的婚姻状况填错了，应将其改为在婚，另外，修改婚姻状况以后还要为其补上配偶 102。

表 16. 婚姻状况填错导致未婚有子女

hhno	210414	210414	210414	210414	210414
name_a_p	刘佑	刘元法	王黛	刘启明	张普
name_a_f	刘长舜	刘佑	王兴旺	刘元法	张耀
name_a_m	徐素贞	罗丽	沈恬田	王黛	-8
name_a_s	-8	王黛	***-8***	张普	刘启明
name_a_c1	刘元通	刘启明	***刘启明***	-8	-8
name_a_c2	刘燕莎	-8	-8	-8	-8
name_a_c3	刘燕兰	-8	-8	-8	-8
name_a_c4	刘元法	-8	-8	-8	-8
name_a_c5	刘元玉	-8	-8	-8	-8
name_a_c6	刘燕娇	-8	-8	-8	-8
code_a_p	101	102	103	104	105
code_a_f	201	101	209	102	211
code_a_m	202	208	210	103	-8
code_a_s	-8	103	***-8***	105	104
code_a_c1	203	104	***104***	-8	-8
code_a_c2	204	-8	-8	-8	-8
code_a_c3	205	-8	-8	-8	-8
code_a_c4	102	-8	-8	-8	-8
code_a_c5	206	-8	-8	-8	-8
code_a_c6	207	-8	-8	-8	-8

3.6 配偶是否健在与婚姻状况矛盾

这包括两种类型：第一，配偶健在但是婚姻状况为丧偶；第二，配偶不健在但是婚姻状况为已婚或同居。这两种类型在逻辑上都不成立，若出现，则一定要修改。

导致这种错误的原因有两种：第一，婚姻状况填错；第二，配偶是否健在填错。

3.6.1 婚姻状况填错

婚姻状况错误是导致这类错误的主要原因。与之前相同，我们判断婚姻状况错误的主要依据还是个人库中的婚姻状况。

表 17 所示的 210838 一户没有匹配错误。102 的婚姻状况为在婚，其配偶为 201，但是 201 已经不健在了，出现矛盾。从个人问卷得知，102 的婚姻状况实际上为丧偶，由此得知出错的原因在

于 102 的婚姻状况填错了。当把 102 的婚姻状况改为丧偶时，其配偶也应删除，因为根据本次调查的原则，未婚、离婚和丧偶者不填写配偶信息。

表 17. 婚姻状况错误导致矛盾

hhno	210838	210838	210838	210838
name_a_p	迟顺松	***迟献***	唐小易	迟绍枫
name_a_f	迟献	迟常言	唐俊朗	迟顺松
name_a_m	马红芝	?	穆影	唐小易
name_a_s	唐小易	***马红芝***	迟顺松	-8
name_a_c1	迟绍枫	迟顺天	迟绍枫	-8
name_a_c2	-8	迟顺尹	-8	-8
name_a_c3	-8	迟顺敏	-8	-8
name_a_c4	-8	迟香玉	-8	-8
name_a_c5	-8	迟同缨	-8	-8
name_a_c6	-8	迟顺松	-8	-8
name_a_c7	-8	迟向军	-8	-8
code_a_p	101	***102***	103	104
code_a_f	102	202	210	101
code_a_m	201	203	211	103
code_a_s	103	***201***	101	-8
code_a_c1	104	204	104	-8
code_a_c2	-8	205	-8	-8
code_a_c3	-8	206	-8	-8
code_a_c4	-8	207	-8	-8
code_a_c5	-8	208	-8	-8
code_a_c6	-8	101	-8	-8
code_a_c7	-8	209	-8	-8

3.6.2 配偶是否健在填错

这种情况比较罕见，下面举例说明。

表 18 所示的 140583 一户没有匹配错误。103 的婚姻状况为在婚，其配偶为 205，但是 205 已经不健在，出现矛盾。从个人问卷来看，103 的婚姻状况同样是在婚，进一步检查 103 的年龄发现，103 仅 27 岁，其丧偶的可能性比较低。综合起来考虑 103 的婚姻状况为在婚的可能性比较大，而错

误的来源是 205 是否健在填错了，不过需要确认。[8]

表 18. 配偶是否健在填错导致矛盾

hhno	140583	140583	140583
name_a_p	王平	韩闰	***韩兆林***
name_a_f	王六	韩国强	韩闰
name_a_m	陈花	章玉	王平
name_a_s	韩闰	王平	***刘晶***
name_a_c1	韩兆林	韩兆林	-8
name_a_c2	韩倩倩	韩倩倩	-8
code_a_p	101	102	***103***
code_a_f	201	203	102
code_a_m	202	204	101
code_a_s	102	101	***205***
code_a_c1	103	103	-8
code_a_c2	104	104	-8

3.7 家庭关系代码重复

在一个家庭中，A 不可能既是 B 的父亲，又是 B 的儿子，也就是说，A 相对于 B 的家庭关系是确定且唯一的。从数据来看，也就是说每一列中的家庭关系代码不应该出现重复，若出现重复就一定要修改。家庭关系代码重复完全是因为匹配错误造成的。

在表 19 所示的 330365 一户中，101 的母亲和配偶都是 102，母亲和配偶的家庭关系代码出现重复，这是不符合逻辑的，所以一定要修改。仔细观察家庭关系以后发现，101 和 102 是夫妻，他们生了小孩 105 和 106，并且 105 和 106 也都承认 101 和 102 是其父母，由此可见 101 和 102 的夫妻关系是比较可靠的，那么必然是 101 的母亲填错了。进一步观察发现 103 和 104 是一对夫妻，他们生了 101，所以，101 的母亲应该是 104，而不是 102。修改方案是将 101 的母亲改成 102。

[8] 对于这种把握并不充分的修改，项目组计划在 2012 进行回访确认，并根据确认后的结果对 2010 年数据进行更新。关于回访的具体情况见下文。

表 19. 家庭关系代码重复

hhno	330365	330365	330365	330365	330365	330365
name_a_p	陈庚	丁岚岚	陈业继	贾月菲	陈雨	陈雪
name_a_f	陈业继	-8	-8	-8	陈庚	陈庚
name_a_m	***丁岚岚***	-8	-8	-8	丁岚岚	丁岚岚
name_a_s	***丁岚岚***	陈庚	贾月菲	陈业继	-8	-8
name_a_c1	陈雨	陈雨	陈庚	陈庚	-8	-8
name_a_c2	陈雪	陈雪	陈金金	陈金金	-8	-8
code_a_p	101	102	103	104	105	106
code_a_f	103	-8	-8	-8	101	101
code_a_m	***102***	-8	-8	-8	102	102
code_a_s	***102***	101	104	103	-8	-8
code_a_c1	105	105	101	101	-8	-8
code_a_c2	106	106	201	201	-8	-8

3.8 一户之内有重名

同一户中不同的家庭成员的姓名不应该完全相同，如果出现则应该视为异常。出现这种问题的常见原因是访员对同一个人重复录入了多次，导致一个人有多个编码。不过我们在核查时也发现在有些家户，姓名完全相同的两个人，确实是不同的两个人。

3.8.1 重复录入

这又分为两种情况：第一种是同住家庭成员与不同住的家庭成员姓名相同，第二种是多个不同住的家庭成员姓名相同。下面举例进行说明。

在表 20 所示的 140886 一户中，105 和 204 都叫赵璐璐，102 和 209 都叫任明伟。从家庭关系可以看出，赵璐璐和任明伟是母子关系，任明伟是赵璐璐的儿子。而进一步的比较可以看出 105 和 204、102 和 209 除了编号不同以外，姓名、性别、年龄、婚姻状况和教育程度都完全相同，由此可见他们本身是同一个人。出现这一问题的主要原因是访员操作不规范导致了重复录入。修改方案是将 2 开头的编码统一修改成 1 开头的编码。

表 20. 同住与不同住家庭成员同名

hhno	140886	140886	140886	140886	140886
name_a_p	杜美玲	***任明伟***	任龙	任虎	***赵璐璐***
name_a_f	杜三	任熙	任明伟	任明伟	赵五
name_a_m	张攀	***赵璐璐***	杜美玲	杜美玲	不知道
name_a_s	任明伟	杜美玲	-8	-8	-8
name_a_c1	任龙	任龙	-8	-8	任凡伟
name_a_c2	任虎	任虎	-8	-8	任国伟
name_a_c3	-8	-8	-8	-8	***任明伟***
name_a_c4	-8	-8	-8	-8	任庆伟
name_a_c5	-8	-8	-8	-8	任艺贞
code_a_p	101	***102***	103	104	***105***
code_a_f	201	203	102	102	205
code_a_m	202	***204***	101	101	206
code_a_s	102	101	-8	-8	-8
code_a_c1	103	103	-8	-8	207
code_a_c2	104	104	-8	-8	208
code_a_c3	-8	-8	-8	-8	***209***
code_a_c4	-8	-8	-8	-8	210
code_a_c5	-8	-8	-8	-8	211

另一种情况是多个不同住的家庭成员姓名相同。如表 21 所示，110056 一户中 205 和 208 都叫李兮，这两个李兮除了编号以外，性别、年龄、婚姻状况、教育程度都完全相同，由此可见是同一个人。修改方案是将 102 的孩 3 改为 205。[9]

[9] 这种情况下通常保留编号较小者。

表 21. 多个不同住家庭成员同名

hhno	110056	110056
name_a_p	李白	赵兰
name_a_f	李东临	赵宝玉
name_a_m	沈惠	陈珂
name_a_s	赵兰	李白
name_a_c1	李和君	李和君
name_a_c2	李斌君	李斌君
name_a_c3	***李兮***	***李兮***
code_a_p	101	102
code_a_f	201	206
code_a_m	202	207
code_a_s	102	101
code_a_c1	203	203
code_a_c2	204	204
code_a_c3	***205***	***208***

3.8.2 姓名虽相同但人不同

在有些情况下，同一户中的人虽然姓名相同，但是性别、年龄等其他信息并不一致，因而是不同的人，此时不做修改。这种情况可能是因为一户之中确实有人同名，也有可能是因为访员在录入姓名的时候输入错误。

在表 22 所示的 510360 一户中，105 和 106 这两个人的姓名完全相同，都叫吉姆么右，但是编号不同。进一步的比较发现，这两个人的年龄也不一样，一个 1 岁，一个 3 岁，由此可见是两个不同的小孩。这两个小孩可能确实都叫吉姆么右，但也有可能是访员在录入的时候将这两个小孩的名字输成同样的了，但是这并不重要。重要的是我们可以判断出他们确实是两个不同的人，因而不做任何修改。

表 22. 姓名虽相同但人不同

hhno	510360	510360	510360	510360	510360	510360
name_a_p	吉姆小小	吉巴阿巴	吉姆尔雅	吉姆么大	***吉姆么右***	***吉姆么右***
name_a_f	吉姆么次	吉巴古汉	吉姆小小	吉姆小小	吉姆小小	吉姆小小
name_a_m	阿什么央	阿克卓玛	吉巴阿巴	吉巴阿巴	吉巴阿巴	吉巴阿巴
name_a_s	吉巴阿巴	吉姆小小	-8	-8	-8	-8
name_a_c1	吉姆尔雅	吉姆尔雅	-8	-8	-8	-8
name_a_c2	吉姆么大	吉姆么大	-8	-8	-8	-8
name_a_c3	吉姆么右	吉姆么右	-8	-8	-8	-8
name_a_c4	吉姆么右	吉姆么右	-8	-8	-8	-8
code_a_p	101	102	103	104	***105***	***106***
code_a_f	201	203	101	101	101	101
code_a_m	202	204	102	102	102	102
code_a_s	102	101	-8	-8	-8	-8
code_a_c1	103	103	-8	-8	-8	-8
code_a_c2	104	104	-8	-8	-8	-8
code_a_c3	105	105	-8	-8	-8	-8
code_a_c4	106	106	-8	-8	-8	-8

3.8.3 由于取名为“不知道”等造成的姓名重复

还有一些情况是由于访员不规范操作使得输入的姓名异常所致。这种情况无法从姓名进行判断，但从家庭关系看基本能判断是不同的人，所以不做修改。

表 23. 姓名异常

hhno	450009	450009	450009	450009	450009
name_a_p	代元春	代策	秦凤	沈贺	代建江
name_a_f	代策	代任丘	***不知道***	周千顺	沈贺
name_a_m	秦凤	钱淑华	***不知道***	***不知道***	代元春
name_a_s	沈贺	秦凤	代策	代元春	-8
name_a_c1	代建江	代元春	代元春	代建江	-8
name_a_c2	-8	代秋云	-8	-8	-8
code_a_p	101	102	103	104	105
code_a_f	102	201	***204***	206	104
code_a_m	103	202	***205***	***207***	101
code_a_s	104	103	102	101	-8
code_a_c1	105	101	101	105	-8
code_a_c2	-8	203	208	-8	-8

如表 23 所示，450009 这一户中 204、205 和 207 的名字都叫“不知道”，因而出现了姓名重复。“不知道”肯定不是人名，而是访员无法收集到该信息时填入的文字。从家庭关系判断，103 和 104 是一对夫妻，204 和 205 是 103 的父母，而 207 是 104 的母亲，所以他们应该是不同的人，不做修改。

3.9 配偶匹配错误

配偶匹配错误有两种类型：第一，一个人同时是家中多个人的配偶；第二，夫妻双方不互认。这两种错误都是匹配出错导致的，必须进行修改。

3.9.1 一个人同时是多个人的配偶

表 24 所示的 450355 这一户中，105 同时是 103 和 104 的配偶，这是不符合逻辑的。而仔细核查家庭关系以后可以看出，103 和 104 实际上是 105 的父母，103 跟 104 才是一对夫妻。出现这种错误的原因是 103 和 104 都将子女填到了配偶处，这种错误在之前已经多次提及。修改方案是将 103 的配偶改成 104，104 的配偶改成 103，同时为 103 和 104 补上孩子 105。

表 24. 一个人同时是家中多个人的配偶

hhno	450355	450355	450355	450355	450355
name_a_p	郑金	李泉友	郑粲	李钰	李嫦
name_a_f	李林英	李少阮	郑晨申	郑金	***郑粲***
name_a_m	郑金莲	李少兰	郑小冉	李泉友	***李钰***
name_a_s	李泉友	郑金	***李嫦***	***李嫦***	-8
name_a_c1	李钰	李钰	-8	-8	-8
code_a_p	101	102	103	104	105
code_a_f	201	203	205	101	***103***
code_a_m	202	204	206	102	***104***
code_a_s	102	101	***105***	***105***	-8
code_a_c1	104	104	-8	-8	-8

3.9.2 夫妻双方不互认

在表 25 所示的 440468 一户中，102 的配偶是 104，而 104 本人无配偶，105 的配偶是 102，也就是说出现了配偶双方不互认的情况。仔细查看该户的家庭关系可以发现，102、103 和 104 都是 101 的子女，也就是说 102 和 104 是兄弟姐妹的关系，而不是配偶，所以 102 的配偶填错了，应该是 105。修改方案是将 102 的配偶改成 105。

表 25. 夫妻双方不互认

hhno	440468	440468	440468	440468	440468
name_a_p	刘诚	谢高洪	谢高合	谢俪	邓娜
name_a_f	刘甫	谢久恒	谢久恒	谢久恒	邓长江
name_a_m	谷照	刘诚	刘诚	刘诚	葛思
name_a_s	-8	***谢俪***	-8	-8	***谢高洪***
name_a_c1	谢高洪	-8	-8	-8	-8
name_a_c2	谢高合	-8	-8	-8	-8
name_a_c3	谢俪	-8	-8	-8	-8
code_a_p	101	102	103	104	105
code_a_f	201	203	203	203	204
code_a_m	202	101	101	101	205
code_a_s	-8	***104***	-8	-8	***102***
code_a_c1	102	-8	-8	-8	-8
code_a_c2	103	-8	-8	-8	-8
code_a_c3	104	-8	-8	-8	-8

3.10 父母与子女不互认

这种错误是指本人填写的父母却不承认本人是其子女。

在两种情况下会出现这种错误：第一，匹配错误；第二，子女是继子女，不是亲生。第一种情况必须修改，而第二种情况如果判断条件不充足则暂不修改。

3.10.1 匹配错误

表 26 所示的 410187 这一户的问题在于：103 母亲是 102，但是 102 的子女里却不包含 103；102

认 104 为其子女，但是 104 的母亲却不是 102，而是 207。仔细查看这一户的家庭关系可以发现，102 的孩 1 应该是 103，而 104 实际上是 102 的儿媳，而非其子女。其判断依据在于：首先，104 并不认 102 做母亲；其次，101 和 102 是一对，其中丈夫 101 姓杨，其子女应该也都姓杨，而 104 姓张不姓杨；再次，103 不仅姓杨，而且 101 认 103 这个儿子，103 也认 101 为父亲、102 为母亲；最后，105 和 106 均认 103 和 104 为父母，证明 103 和 104 为配偶关系。综上可以判断，103 才是 102 的小孩，104 是其儿媳。修改方案是将 102 的孩 1 改成 103。

表 26. 匹配错误导致父母与子女不互认

hhno	410187	410187	410187	410187	410187	410187
name_a_p	杨大海	杨娣	***杨战国***	***张文杰***	杨链	杨桐
name_a_f	杨慎	杨威	***杨大海***	***张远丰***	杨战国	杨战国
name_a_m	刘氏	孙氏	***杨娣***	***尹金枝***	张文杰	张文杰
name_a_s	杨娣	杨大海	张文杰	杨战国	-8	-8
name_a_c1	***杨战国***	***张文杰***	杨链	杨链	-8	-8
name_a_c2	杨春秋	杨春秋	杨桐	杨桐	-8	-8
code_a_p	101	102	***103***	***104***	105	106
code_a_f	201	204	***101***	***206***	103	103
code_a_m	202	205	***102***	***207***	104	104
code_a_s	102	101	104	103	-8	-8
code_a_c1	***103***	***104***	105	105	-8	-8
code_a_c2	203	203	106	106	-8	-8

3.10.2 子女非亲生

如果子女并不是本人亲生，则有可能出现不互认的情况。如表 27 所示的 140025 这一户中，104 填写的父母为 103 和 101，但是 103 并不认 104 为其子女。查看家庭关系可以发现，101 和 103 虽为夫妻，但是子女都不相同；而从个人库的数据来看，101 和 103 都是再婚。由此可见，其填写的子女都是与前妻或前夫所生。所以这里虽然存在子女与父母不互认的情况，我们并不做修改。这个问题之所以产生是因为 CFPS 在设计时没有规定父母填写的子女必须为其亲生子女、子女填写的父母必须为其亲生父母。对于这种再婚有继养子女的情况，我们在无法做出确定性判断时保留受访者原始的回答，暂不做修改。我们会在 2012 年的回访调查中针对这一类问题进行信息确认，并根据确认后信息对相关数据进行更新。

表 27. 子女非亲生导致父母与子女不互认

hhno	140025	140025	140025	140025
name_a_p	高晓松	赵晓宛	王超军	宋佳佳
name_a_f	高福	1	王益勤	***王超军***
name_a_m	李秀	2	赵晓宛	***高晓松***
name_a_s	王超军	-8	高晓松	-8
name_a_c1	***宋甜甜***	王志军	***王璐***	-8
name_a_c2	***宋佳佳***	王秋雨	***王也***	-8
name_a_c3	-8	王超军	-8	-8
name_a_c4	-8	王从军	-8	-8
code_a_p	101	102	103	104
code_a_f	201	204	209	103
code_a_m	202	205	102	101
code_a_s	103	-8	101	-8
code_a_c1	***203***	206	***210***	-8
code_a_c2	***104***	207	***211***	-8
code_a_c3	-8	103	-8	-8
code_a_c4	-8	208	-8	-8

3.11 初婚配偶填写的子女数目不一致

表 28. 初婚配偶填写的子女数目不一致

hhno	140778	140778	140778	140778	140778	140778
name_a_p	武三思	武仲	史馨	孟美	武蕊	武航
name_a_f	武仲	武大宗	史雪	孟子怡	武三思	武三思
name_a_m	史馨	文音	贺姿	张想	孟美	孟美
name_a_s	孟美	史馨	武仲	武三思	-8	-8
name_a_c1	武蕊	***武思齐***	***武思齐***	武蕊	-8	-8
name_a_c2	武航	***武三思***	***武三思***	武航	-8	-8
name_a_c3	-8	***-8***	***武江秀***	-8	-8	-8
code_a_p	101	102	103	104	105	106
code_a_f	102	201	204	207	101	101
code_a_m	103	202	205	208	104	104
code_a_s	104	103	102	101	-8	-8
code_a_c1	105	***203***	***203***	105	-8	-8
code_a_c2	106	***101***	***101***	106	-8	-8
code_a_c3	-8	***-8***	***206***	-8	-8	-8

如果夫妻双方都是初婚，那么他们的子女数应该相等，若不相等必须修改成相等。导致这种错误的唯一原因是匹配错误。

在表 28 所示的 140778 这一户中，102 和 103 是一对初婚夫妇，其中 102 填写了 2 个孩子，而 103 填写了 3 个孩子。由于他们是初婚，其子女数量应该一样，所以将 206 补充到 102 的孩 3 位置。

4. 个别案例的回访工作

尽管多数问题家户都可以通过以上逻辑关系纠正错误，但仍有少部分家户由于信息不足无法做出更正。例如，在仅出现父母性别相同这一问题时，如果父母的姓氏相同、父母的名字无法辨认性别，也没有其他孩子的信息判断谁是其父谁是其母，那么我们就没有足够的证据判断这一户是父母性别填错，还是父母位置填反。又如单人家庭户中出现父亲（或母亲）年龄与孩子年龄差大于 50 岁，作为父亲（或母亲）的本人又将婚姻状况填为未婚，没有配偶信息，我们难以判断不同住的孩子到底是作为父亲（或母亲）的本人的亲生子女，还是领养，还是将孙代误填作子代；我们也难以判断作为父亲（或母亲）的本人是未婚生育，还是丧偶或离异？对这些难以判断的家户，我们不得不以回访的方式来确认和补全信息。数据清理后的汇总结果显示，由于信息不足而难以作出判断和更正的家户共有 248 户，涵盖的错误类别主要包括：性别错误、父母与孩子年龄差不合理、父母与孩子不互认、一户之内有重名、未婚或不够婚龄（20 岁以下）却有配偶或子女。根据这些家户的错误类型，我们针对每一个问题家户填写的具体信息为其设计了个性化问卷（case-by-case questionnaire，或名“个别问卷”）（示例见图 1、2）。个性化问卷的作用是确认和补充信息。区别于计算机化的问卷，个性化问卷采用纸版问卷形式，由访员在回访的时候手工填写。个性化问卷中的问题针对的是每一个问题家户中个别人的个别信息（如：XXX 的母亲是谁？XXX 的性别是什么？），这些信息均是后期判断所需的关键信息，因此对每户提问的问题不同。每一份个性化问卷后都附上该家户的 T2 表，以供访员参考、核对。个性化问卷的回访与 2012 年的调查同时进行，执行时间是在 2012 年计算机化的调查问卷填答结束之后。待个性化问卷回收之后，数据组会根据问卷中新收集和已确认的信息，重新编写、核对这些家户的更正命令。

家户编号：xxxxxx　　访员姓名_____　　访员编号_____　　受访人姓名_____

个别问卷（I）

问题描述：婚姻状态存疑

操作提示：

1. 请问，**黄某**在 2010 年的婚姻状况是【出示卡片】？（请在记录表相应空格里填写婚姻状况，如：未婚、在婚、同居、离异、丧偶。）

[访员注意：未婚仅指从来没结过婚。若回答是“未婚”，则进一步确认：**黄某**从来没结过婚吗？]

2. 请问，**黄某**在 2010 年的配偶谁？（请在记录表相应空格里填写名字，如确认原填写，则打 √）

记录表：婚姻状况填写（未婚/在婚/同居/离婚/丧偶）。

	本人	婚姻状况	2010 配偶
原填写	102 黄某	在婚	201 李某
确认/更正	102 黄某		

[访员注意：　2010 年初访时，**李某**被填写为“不健在”]

备注表（如有）：填写同住家庭成员的名字，操作同 T2 表，并在家户资料中圈出此人

本人	父	母	配偶	孩 1	孩 2	孩 3	孩 4	孩 5

图 1.　个性化问卷示例（一）

家户编号：xxxxxx　　访员姓名_____　　访员编号_____　　受访人姓名_____

个别问卷（II）

问题描述：再婚重组家庭，确认亲生/继养子女
操作提示：
1. 请问，**黄小某**与**黄某**有血缘关系吗？（回答的“有”，则在相应的空格处打√；若回答“没有”，则打 X。）
2. 请问，**黄小某**与**李某**有血缘关系吗？（回答的“有”，则在相应的空格处打√；若回答“没有”，则打 X。如果某个孩子名字下方的两个空格均填 X，则继续提问：请问，**黄某**和**李某**是**黄小某**的父母吗？如果确认是孩子的父母，则结束提问。如果孩子另有父母，则继续提问 3*）
3*. **黄小某**的血缘父母是谁？（请在备注表的“本人”一栏填入**黄小某**，并根据回答在相应的空格里填入孩子的血缘父母姓名）
［访员注意：如果孩子的父母是家户资料中已有成员，则需在家户资料中圈出名字及编号，并将名字及编号填入表格中］

记录表：有血缘关系=√，没有血缘关系=X

		孩 1	孩 2	孩 3	孩 4
		204 黄小某			
父	102 黄某				
母	201 李某				

备注表（如有）此表填写名字及编号（如有，参见家户资料）

本人				
父				
母				

图 2.　个性化问卷示例（二）

5.　对 T 表在问卷设计和调查操作中的建议

基于以上错误类型说明和清理方案，我们建议在采用 T 表设计来收集家庭关系和家庭成员基本信息的计算机化问卷调查中应留意如下事项：

⑴ 访员在 T1 表录入同住家庭成员以及在 T2 表录入非同住家庭成员时，要防止出现一个人重复录入从而导致一人多码的情况。我们建议通过 CAPI 系统进行自动检查，如果发现新录入的（同住或不同住）家庭成员的姓名与 T1 或 T3 表中已有的姓名完全相同，系统需要提示访员确认是否重复录入。

⑵ 当遇到代答人不知道某一家庭成员名字时，不要以“不知道”、“未知”、“无名氏”、“某氏”来取名，最好用该成员对应的庄主名字及其与庄主的关系来联合命名，如“XXX 的爸爸”，“XXX

的妈妈”。这一点主要通过访员培训时来实现。

⑶ 在填写 T2 表时，应明确要求遵循血缘关系（领养等同于血缘）的原则，向代答人说明父母-子女关系须是亲生或者领养。同样，这一点也需要通过访员培训来实现。

⑷ T2 表在设计中应设置一些逻辑检验，限定每一行中（填写每一个同住家庭成员的家庭关系时）的代码不能有重复。如，若 A 已经填写在 B 的父亲位置；那么 A 就不应该出现在 B 的配偶、子女等其他家庭成员的位置上，以防出现家庭关系代码重复的问题。

⑸ 如有可能，在 CAPI 系统里面增加有关性别、年龄、婚姻状况和是否健在的逻辑检验。如 A 在家庭关系中是 B 的父亲，那么 A 的性别必须为男，年龄一般比 B 大 15 岁以上、50 岁以下。可以根据本报告的表 1 设计逻辑检验条件，其中级别为 1 的需要 hard check，级别为 2 或 3 的需要 soft check。

⑹ T2 表填完后，需要生成一张总表，并要求访员严格检查这张总表，重点检查项目是：子女与父母是否互认、夫妻双方是否互认、子女填报的父母是否也互认为夫妻、夫妻各自填报的子女数目是否相等。如果发现上述四项有异常，访员应该当场进行确认并修改。

⑺ 对于访员发送回来的数据要进行实时检查，可在表 1 列出的 28 项基础上建立一套常规核查标准，发现问题立即通过督导与访员取得联系，力争将错误解决在数据收集阶段。