大型追踪调查样本流失规律及模式

——以中国家庭追踪调查为例

孙　妍　王　赫　潘修明　张春泥*

内容摘要：本文基于中国家庭追踪调查 2010—2022 年七轮数据，分析家庭样本流失模式及变化趋势。研究结果表明，以截面视角的流失率来衡量，家庭样本随着追踪期数的增加，流失规模逐渐扩大，但流失速度呈现前快后缓的特征。从追踪视角来看，最常见的流失类型是样本家庭在不同轮次间处于失访与完访状态间来回切换的摇摆流失，这意味着将曾经失访的样本保留在追踪库中有助于减少样本的损耗。社会外部环境的变化导致的访问模式转换，也会引发样本快速流失。失联和拒访是导致样本流失的主要因素，对于因不同原因失访的家庭，后续挽回的难度也有差异。对于失联流失的样本，挽回的难点在于获取有效的联系方式；而对于拒访流失的样本，挽回难度较大。家庭规模较大、居住流动性较低、农村和低收入家庭更倾向于全程参与调查，受访者对往期轮次调查的态度也对后续流失有一定影响。在实践中，综合运用往期追踪状态划分的流失组别能够较好地预测下一轮的完访率，能够为样本维护及实地执行提供有效参考信息。

关键词：追踪调查；样本流失；序列聚类分析；中国家庭追踪调查

中图分类号：C31　**文献标识码：**A　**文章编号：**1004-7794(2025)03-0063-14

DOI: 10.13778/j.cnki.11-3705/c.2025.03.006

一、引　言

随着现代社会调查技术的发展，具有全国代表性的固定样本追踪调查日益成为高质量社会科学研究的重要微观数据来源。2010 年以来，国内兴起了多个具有特色的大型固定样本追踪调查项目。例如，北京大学中国社会科学调查中心实施的中国家庭追踪调查（CFPS）、中国健康与养老追踪调查（CHARLS），西南财经大学开展的中国家庭金融调查（CHFS）等。

固定样本追踪调查关注事件的时间属性，通过对固定一批家庭或个人在不同时点上进行重复观测的追踪设计，充分捕捉个体随时间变化的特征和社会现象的动态趋势。与截面调查相比，追踪调查数据具有无可比拟的优势。首先，其数据采集的长期性有助于研究者了解总体及个体的发展规律和变化趋势，为社会现象研究提供更全面、深入的视角；其次，对同样个体的重复观测有利于研究者在因果推断时较好地控制个体异质性的干扰，并通过纳入时滞变量更准确地反映变量

* 孙妍，2009 年毕业于北京林业大学，获管理学博士学位，现为北京大学中国社会科学调查中心副研究员，研究方向为社会调查方法。王赫，北京大学社会学系硕士研究生，研究方向为社会分层与不平等、计算社会科学、文化与认知。潘修明，2024 年毕业于北京大学社会学系，获法学硕士学位，研究方向为青年研究、社会分层与流动研究。张春泥（通讯作者），2013 年毕业于香港中文大学，获社会学专业博士学位，现为北京大学社会学系长聘副教授，研究方向为社会分层与不平等研究、社会人口学，邮箱：chunnizhang@pku.edu.cn。本研究得到教育部人文社会科学研究一般项目“追踪调查样本流失模式及维护策略探索性研究”（18YJC840034）、教育部重大课题攻关项目“新中国成立以来社会调查发展进程、经验与展望研究”（20JZD032）的资助。

间的因果顺序；最后，追踪调查持续采集的数据也有助于研究者深入探究社会现象背后的机制和过程，如代际传递、社会流动等（任强和谢宇，2011），上述数据优势将随着追踪轮次的增加而逐渐体现出来。例如，美国密歇根大学的收入动态调查（Panel Study of Income Dynamics，PSID）从 1960 年持续至今，已成为世界上持续时间最久的追踪调查，也成为美国对主要社会问题进行学术和政策研究的重要数据来源（Smeeding，2018）。

然而，固定样本追踪调查的实施难度会随着时间推移和追踪轮次增加呈几何级数的增长，在项目实施过程中需要应对诸多挑战，其中最为突出的问题就是样本流失。以美国 PSID 为例，经过前三轮逐年追踪，原始样本流失比例可达 15%；至 1989 年第 20 轮追踪时，原始样本已有近半数流失（Fitzgerald et al.，1998）。单纯从社会调查与统计分析的角度而言，样本流失会产生诸多负面影响：首先，它直接导致样本量缩减，削弱统计的效度，降低研究结果的可靠性。其次，若存在选择性的样本流失，会导致应答样本总体代表性受损，基于剩余样本的研究结论可能仅适用于特定的子群体，难以具有普遍性。再者，如果流失样本的属性与研究所关注的变量有关，则流失也会成为估计偏误的来源。为了应对样本流失问题，国际上许多追踪调查项目会定期或不定期对追踪样本进行更新，以期与总体特征保持一致（王俊等，2024）。

鉴于样本流失问题对研究结果的影响，如何减少样本流失成为调查界持续关注和深入研究的重要课题。从 20 世纪 70 年代起，国外学者已针对发达国家追踪调查项目中受访对象流失的特点及模式进行了细致的研究，分析受访者的人口统计学特征、社会经济因素、家庭因素、调查设计因素等对受访对象是否持续参与追踪调查（或追踪应答）的影响（Groves et al.，2000；Watson，2003；Uhrig，2008）。为了厘清影响追踪应答的机制，Groves & Couper（1998）提出了递进式应答模型，即将追踪应答分解为接触样本和邀约访问两阶段递进式的行为，进而指出影响接触成功率和邀约成功率的因素会各有不同。既有文献发现，在以家庭为调查对象的追踪调查项目中，接触成功率主要受家庭特征的影响，例如，家庭规模、拥有孩子的情况、家庭流动性（通常用住房拥有情况、居住年限等进行度量）会影响受访家庭是否被成功联系（James，2023）。影响邀约成功率或个体参与访问意愿的因素则更多是受访对象个人特征，如年龄、性别、受教育程度等（Uhrig，2008；Lynn et al.，2012），但学历、收入等个人属性对样本流失的影响则较为复杂，在不同项目和地区间并未取得一致的结论（Fitzgerald et al.，1998；Lillard & Panis，1998；Behr et al.，2005；Frankel & Hillygus，2014；James，2023）。除个人特征外，个人行为模式也影响他们的应答意愿和行为。例如，社区活动参与度高的群体在调查活动中的参与度也更高，受到调查是否提供激励的影响较小（Groves et al.，2000）；前期调查参与表现较好的受访者后续持续参与访问的可能性更高（Watson，2003；Nicoletti & Peracchi，2005）。

追踪调查在中国起步较晚，但上述提及的三项大型追踪调查项目也实施了 5~8 个轮次，各个项目均面临不同程度的样本流失问题。然而，既有国内学界关注追踪调查方法研究成果相对有限，迄今为止仅有少数几篇论文探讨过国内追踪调查项目的样本流失问题。这些研究或将追踪状态简化为完访、失访进行二元化处理（梁玉成，2011；齐嘉楠，2016），或仅选择拒访一个维度进行分析（孙妍等，2011），对引发样本流失的具体原因及随着追踪轮次增加的演变规律识别仍有所不足。我国在过去十年中经历了城市化进程加速、人口流动水平上升、人户分离现象增多等变化，这些社会变化均可能影响国内追踪调查项目的实施难度。在样本流失的问题上，发达国家追踪调查项目经验虽然有借鉴价值，但由于国情和文化的差异，中国追踪调查样本流失的特点和模式仍有待专门研究，并从中寻找提高样本追踪率、降低样本损耗的有效手段。

鉴于此，本研究将依托中国家庭追踪调查（China Family Panel Studies，CFPS）项目，重点回答两个问题：首先，从截面及追踪两个维度上看，样本流失有哪些模式和特点？其次，区分不

同类型的样本流失，关注流失受哪些因素的影响？最后，基于这两个问题的回答，运用研究归纳出的样本流失规律对后续轮次的应答进行预测和检验。

二、数据与方法

（一）数据来源

本研究关注的 CFPS 是北京大学中国社会科学调查中心设计和实施的一项全国代表性、综合性的长期固定样本追踪调查。该调查于 2010 年启动全国首轮调查（基线调查），在全国的 25 个省（区、市）（不含香港、澳门、台湾、新疆、西藏、青海、内蒙古、宁夏、海南）成功访问了 14960 户家庭及其家庭成员。在基线调查之后，CFPS 保持每两年回访一次的追踪模式，本研究运用的是 CFPS 2010—2022 七轮调查数据。

（二）CFPS 追踪规则

CFPS 将国际家户追踪调查项目经验与自身特点相结合，在构建追踪数据库的同时，也尝试保证其截面样本具有总体代表性，据此设计了一套具备自我更新功能的追踪规则，包括访问对象的识别和追踪样本的纳入条件两个方面。

1．访问对象识别。

CFPS 以追踪基因成员的家庭为原则，每轮次动态识别需要访问的家庭及个人。“基因成员”是指 2010 年基线调查完访家庭内的所有家庭成员。在基线调查之后，基因成员新出生的血缘子女以及新领养的 10 岁以下子女，同样会继承“基因成员”的身份，以此保证基因成员样本的自我更新。CFPS 将基因成员界定为永久追踪对象，每轮调查中基因成员所在的家庭作为 CFPS 的目标访问家庭。在目标家庭中，CFPS 运用嵌套数据采集模式，在家庭层次采集受访家庭成员的构成和家庭经济状况，在个体层次采集该家庭中的基因成员及其非基因直系亲属（父母、子女及配偶）的个人信息。

这一规则意味着在 CFPS 项目中，虽然基因成员是相对固定的，但基因成员所属的家庭及其家庭成员的构成在各个轮次间均可能发生变化。当基因成员因婚姻、分家、流动等各类原因迁移至新的家庭时，其所在的新家庭也会被纳入为目标访问家庭；反之，当家庭中不再存在基因成员时，对该家庭的访问则会停止。从 2010 年到 2020 年，经过六轮访问后，CFPS 基线家庭中约有 30%发生了分裂，目标受访家庭数量每轮以约 7%的速率增长，从最初的 14960 户扩张至 19593 户。

2．追踪样本纳入条件。

为了最大限度降低样本损耗，CFPS 并未沿用国际上知名家庭追踪调查项目通常采用的“移除连续两轮次未应答样本”这一常规操作（Goebel et al.，2019；PSID，2023），而是设定了较为宽松的无条件放回方案。具体而言，CFPS 仅在样本自然消亡（全家去世）、不存在基因成员或受访家庭强烈要求退出这三种情形下才会移除受访家庭；而对于其他因各类原因在各轮次访问中无应答（失访）的家庭，无论其累计失访轮次有多少，均会被保留在样本库中，在后续轮次的追踪时仍作为继续尝试接触并劝访的对象。这种无条件放回追踪样本纳入规则不仅适用于家庭样本，也同样适用于户内的个人样本。因此，在 CFPS 数据库中，家庭或个人层次都存在追踪不连续的样本数据（孙妍等，2024）。

（三）样本流失的界定

本研究的因变量是样本流失，样本流失的界定取决于分析单元的界定和流失的界定两个方面。

1．家庭作为流失的分析单元。

现有文献中大多以永久追踪的个人作为分析单元（Lillard & Panis，1998；Watson，2003；Nicoletti & Peracchi，2005；Michaud et al.，2011；Friedel & Birkenbach，2020），这种做法的优势在于能得到一个相对恒定的研究总体，但弊端也较为明显。首先，如调查涉及家庭及个人两个递进层次的信息采集，个人样本的流失其实是家庭与个人两级决策叠加的结果，而只以个人作为分析单元，实际上忽视了家庭层次无应答的群体效应，即由于整个家庭失访引发的家庭成员集体失访。以 CFPS 为例，若以所有目标受访个人为观测总体，各轮所有个人失访中超过半数属于因整个家庭失访导致个人失访的情况，换言之，整个家庭流失是研究家庭追踪调查样本流失时不可忽视的类型。其次，家庭样本流失与户内个人样本流失形成的机制存在差异（Lipps，2009）。家庭失访通常是由于无法与目标受访家庭取得联系或受访家庭集体决策的结果，而户内个人失访则多由于个体家庭成员的参访意愿较低而导致。鉴于 CFPS 的追访是从家庭到个人的问卷生成顺序，且家庭样本流失是导致个人样本流失的最主要来源，本研究选择以家庭为分析单元来研究样本流失。

对于像 CFPS 这类不断纳入新分裂家庭的追踪调查而言，存在原家庭和新分裂家庭两类家庭，因此，需要明确家庭样本流失的基数构成。从以往其他追踪调查项目对追踪家庭的界定来看，存在地址识别和人员识别两种模式（Lipps，2009）。地址识别模式下，居住在原地址的家庭视为原家庭、继承原家庭编码，新分裂的家庭获得可识别家庭关联的新编码。人员识别模式下，是根据项目所关注核心人员所在家庭归属判别应该由哪个家庭继承原家庭编码。CFPS 原家庭识别是由实地接受访问的顺序来决定的，兼具地址识别及人员识别两种模式。面访为主的追踪模式下，访员会先返回上期受访家庭所在的地址寻找原家庭，原家庭成功访问后再对离家个人进行追踪。因此，面访追踪基本遵循的是地址识别模式。在电访模式下，若存在家庭分裂，则将先接受访问的个人所在的家庭定义为原家庭、继承原家庭编码。由于原家庭识别与访问顺序紧密相连，因此仅仅依靠家庭编码识别基线家庭会导致识别标准不一致。为降低原家庭识别标准差异对研究结论的影响，本研究对所有分裂家庭进行整合，统一放回基线原家庭，并以基线家庭为分析单元开展样本流失分析。

2．家庭流失的界定。

现有聚焦追踪调查样本流失议题的文献，界定流失的方法大体可以分为三类：单调分析（Uhrig，2008）、时点分析（Watson，2003；Lynn et al.，2012）及追踪分析（Lillard & Panis，1998；Michaud et al.，2011；James，2023）。在单调分析和时点分析视角下，受访对象在某轮次或某时点的失访即被定义为流失样本。这种视角的问题在于忽视了受访对象在后续被成功逆转的可能性和其在追踪过程中跨轮应答状况的动态变化，而且其研究结果也会受所选择的特定轮次和时点的影响。相比之下，追踪分析能综合考虑各轮次的完访状态，把握流失的动态性。

鉴于 CFPS 采取的是无条件放回的追踪策略，能够持续采集所有目标受访家庭样本在所有追踪轮次的访问状态，因此既可以采用截面视角，分析单一轮次或时点的失访或完访，也可以采用追踪视角，综合各轮次家庭完访的情况来界定流失模式。

此外，除了从时点角度关注完访状态外，本文还进一步纳入了受访家庭每轮次失访的原因，具体分为失联、拒访、其他原因未完访及全家死亡四类。需要说明的是，遵循前文提及的回归基线的家庭界定原则，若基线家庭产生分裂，所有关联家庭中有一个家庭完成访问则视为基线家庭完访；若所有关联家庭均失访，则提取获取信息最多的联系结果作为家庭失访的原因。例如，某基线家庭分裂出一个新家庭，在某轮追踪时，这两个关联家庭均未完访，失访的原因分别是失联和拒访，则将该基线家庭单元当轮次的追踪状态定义为“失访”、流失原因定义为“拒访”。

（四）影响流失的因素

本文的第二个研究问题关注样本流失的影响因素。受访对象是否在追访中流失由样本能否被成功接触以及受访对象是否愿意接受访问共同决定。影响样本可及性及受访意愿的因素来自家庭、决策者个人、社会环境三方面。本研究中，家庭层面纳入了反映样本联系困难程度的指标，包括家庭规模（人）、家庭成员构成（60 岁及以上老人的数量、15 岁及以下孩子的数量）、家庭类型、新分裂家庭数量。考虑到搬迁、流动最易引发样本流失，纳入了家庭成员中是否有人拥有本地户口、家庭居住房屋是否为自有住房这两个可预测家庭流动性的变量。由于本文研究的是家庭层面的流失，因此影响家庭参与调查兴趣的因素纳入了家庭收入水平及家庭主事人（户主）的个人属性，包括其性别、年龄和受教育水平。区域层面主要考虑城乡的差别，关注居住在城市及农村家庭的流失差异。与截面调查一次性访问不同，追踪调查受访对象前期的受访经历及表现可能会影响他们后续的参访决策。本研究中选取了访员在完成家庭成员个人问卷后对应答家庭成员所做的三项主观评价指标：访问的配合度、对调查感兴趣的程度、对调查的疑虑水平，每项评价在家庭层面取平均值。与此同时，CFPS 在一个家庭中会向所有符合访问资格的家庭成员发出完成个人问卷的访问邀约，将家庭中个人问卷的完访率视为是该家庭对项目访问配合程度的客观测量指标。由于本研究采用回归基线家庭的模式识别流失，故影响因素分析模型纳入的各项指标数据也同样是来自基线调查的数据。

（五）分析方法

本研究将采用序列聚类分析方法来分析追踪视角下家庭样本的流失类型。具体做法是：首先，根据家庭样本从 2010 年到 2020 年六轮调查的完访情况构造了每个家庭应答状态（分完访和失访两种状态或分不同失访原因）的序列；其次，参考 Potârcă et al.（2013）的做法对特征相近的序列进行聚类，得到不同的流失类型。

在探讨影响流失的因素上，本文以家庭样本在六轮调查中是否非全程追踪（1=是/存在流失，0=否/全程追踪）或是否为某种类型的流失（1=特定流失类型，0=全程追踪）作为因变量，运用二元 Logit 回归模型来探讨各因素的效应。

三、研究发现

（一）截面样本流失特点及演变规律

截面样本流失率是指在第 N 轮调查时，全体基线家庭单元中未完访的比例。根据流失样本是否存在于截面总体中，截面样本流失可划分为样本损耗（Sample Attrition）和总体损耗（Population Attrition）两类。总体损耗特指因自然原因引发的样本流失，典型的例子如受访家庭全家去世、移民。由于总体损耗的样本在当轮调查中入样概率为零，不会影响样本总体的截面代表性。因此，本文聚焦样本损耗，关注随时间推移样本流失类别的变化规律。

表 1 展示了 2012—2020 年五轮追踪历程中基线家庭样本的访问状态，可以看到，家庭样本的截面流失率随追踪期数增加呈现增长的趋势。因全家死亡导致的总体损耗始终位于一个较低水平，而失联和拒访是导致家庭截面无应答最主要的两类样本损耗类型。相较于失联率，拒访率随着时间推移的增长幅度更高。这表明随着调查轮次数的增加，受访家庭参与调查的疲惫感与日俱增，主动退出项目的群体规模扩大。较之 2018 年，2020 年的截面流失率出现了激增，这主要是由于外部环境的变化迫使访问模式由面访为主转换为以电话访问为主，由此加剧了样本流失，尤其是推高了样本失联率。可见，访问模式转变会对成功联系到受访户造成很大的负面影响。

表 1　　CFPS 家庭样本逐轮应答情况

	2012 年	2014 年	2016 年	2018 年	2020 年
完访	12724	12436	12103	11763	10046
失联	1026	1346	1430	1515	2544
拒访	744	868	1066	1257	1672
其他原因未完访	431	221	203	192	407
全家死亡	35	89	158	233	291
样本家庭总数	14960	14960	14960	14960	14960
截面流失率（%）	14.9	16.9	19.1	21.4	32.8
拒访率（%）	5.0	5.8	7.1	8.4	11.2
失联率（%）	6.9	9.0	9.6	10.1	17.0
其他流失（%）	2.9	1.5	1.4	1.3	2.7
总体损耗率（%）	0.2	0.6	1.1	1.6	1.9

（二）追踪样本流失特点及演变规律

1．按追踪状态分类的流失模式。

从追踪角度分析样本流失的优势在于能综合家庭样本在所有访问轮次的应答状态，对流失模式相似的家庭进行聚类分组。根据家庭样本从 2010 年到 2020 年六轮调查的应答状态（分完访和失访两种），本研究通过构造序列和对序列的聚类得到五种类型：全程追踪、摇摆流失、冲击流失、中途流失、首期流失。其中，全程追踪是完整参与六轮访问的家庭，首期流失、中途流失、冲击流失和摇摆流失则存在不同程度的样本流失。

图 1 展示了全程追踪组别外其余四个组别的家户数量和应答序列。首期流失和中途流失的特征是在当前分析时点上呈现单调样本损耗，即在第一轮或中间某轮追踪时停止参与访问且不再返

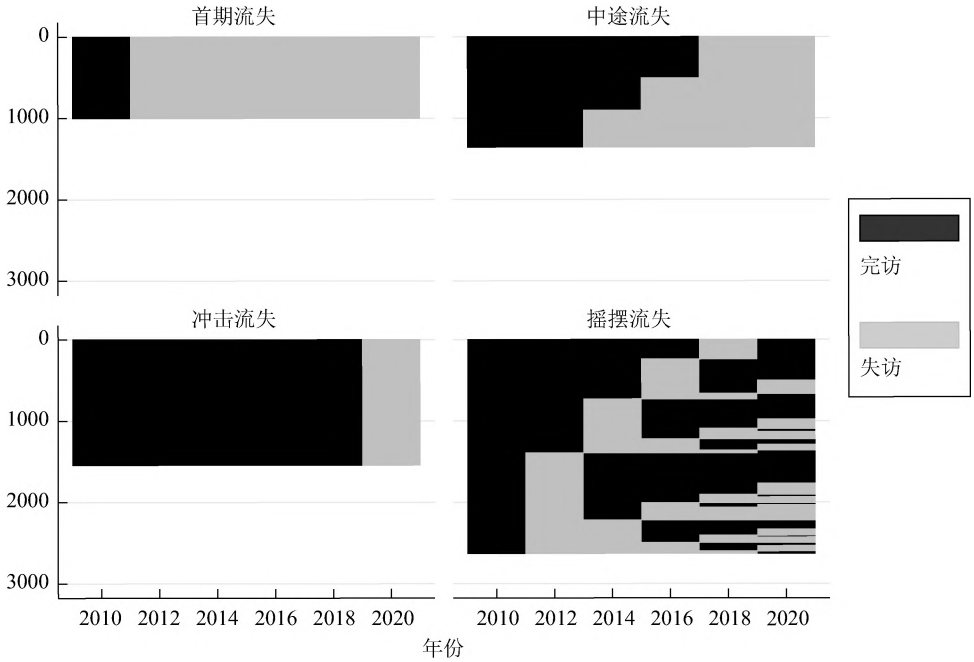

图 1　按追踪状态分类的四种流失模式的应答序列

回。冲击流失主要是在 2020 年轮次访问模式转换前持续接受访问，但切换为电访模式后未应答的家庭，这体现了特定外部环境下访问模式转化所造成的影响。摇摆流失汇聚了各轮次应答状态在完访、失访间来回切换的家庭，这些家庭可能在个别轮次未参与调查，但在随后的轮次中出现了不同程度的回归。

表 2 显示，在全部的 14960 个基线家庭分析单元中，全程追踪的样本占比约为 56.4%。在有过失访经历的家庭中，摇摆流失占比最高（约四成），在全部基线家庭分析单元中占比为 17.6%。排在第二位的是冲击流失，有 10.3%家庭是在访问模式由面访切换为电访后中断了持续访问，可见访问模式的切换对家庭持续应答造成了较大负面影响。两类单调损耗样本在所有家户占比约为 15.8%，其中，持续接受两期或更多期访问后退出的家庭占比为 9.1%；仅完成基线调查，从首期追踪（即 2012 年第二轮访问）起就不再应答的家庭在所有家庭中占比为 6.7%。

表 2　　按追踪状态分类的各类家庭分布

	家户数（户）	占比（%）	非全程追踪家庭中的占比（%）
全程追踪	8435	56.4	
摇摆流失	2629	17.6	40.3
冲击流失	1541	10.3	23.6
中途流失	1355	9.1	20.8
首期流失	1000	6.7	15.3
合计	14960	100.0	100.0

整体而言，以单调损耗为特征的首期或中途流失并不是主导的流失模式。由于访问模式变化引起的冲击流失是外部环境突然变化的结果，具有偶发性和不可预防性。而摇摆流失这类中断访问又返回的样本最值得关注。若按照以往国际惯例，连续两轮次未应答样本即从追踪对象中移除，项目将会永远失去这些可挽回的摇摆流失样本，而 CFPS 采取将往期失访样本保留在追踪样本库的做法则很大程度上有助于降低因样本流失所带来的样本规模损耗。

2．按失访原因分类的流失模式。

根据非全程追踪家庭样本每期失访原因，也可以构造六轮调查的失访原因序列，序列聚类的结果显示，非全程追踪的家庭样本可分为四种模式：死亡流失、失联流失、拒访流失、其他（受访家庭原因或执行原因）流失（见图 2）。

表 3 展示了按失访原因分类的家庭分布。首先，死亡流失对样本损耗造成的影响最小，经历了六轮调查，CFPS 样本总体损耗保持在一个相对低的水平。失联流失是导致家庭样本无法持续追踪的首要原因，在非全程追踪家庭中的占比约为 48.8%。其次是拒访，引发样本流失的比例为 35.7%。此外，11.0%的样本失访是由受访者身体不适等其他原因引起的。这一结果表明，未能与目标受访家庭取得联系（即失联）是导致样本家庭无法持续参与访问的最大障碍，若能获得受访家庭有效联系方式，将会极大改善追踪调查家庭样本流失状况。

表 4 显示，除死亡流失外，失联流失、拒访流失和其他流失中都有较高比例的摇摆流失，即在初次失访后回归访问。其中，因其他原因流失的家庭中属于摇摆流失类别的样本占比最高，达 68.2%；失联流失和拒访流失类别中，也均有接近四成属于摇摆流失。这意味着对于这些非死亡流失的家庭，如采取继续追踪的策略，后续均存在逆转的可能。

（三）样本流失影响因素分析

接下来，本文将以全程追踪家庭为参照组，探讨各类因素对是否发生流失（非全程追踪）和各种类型流失的影响。

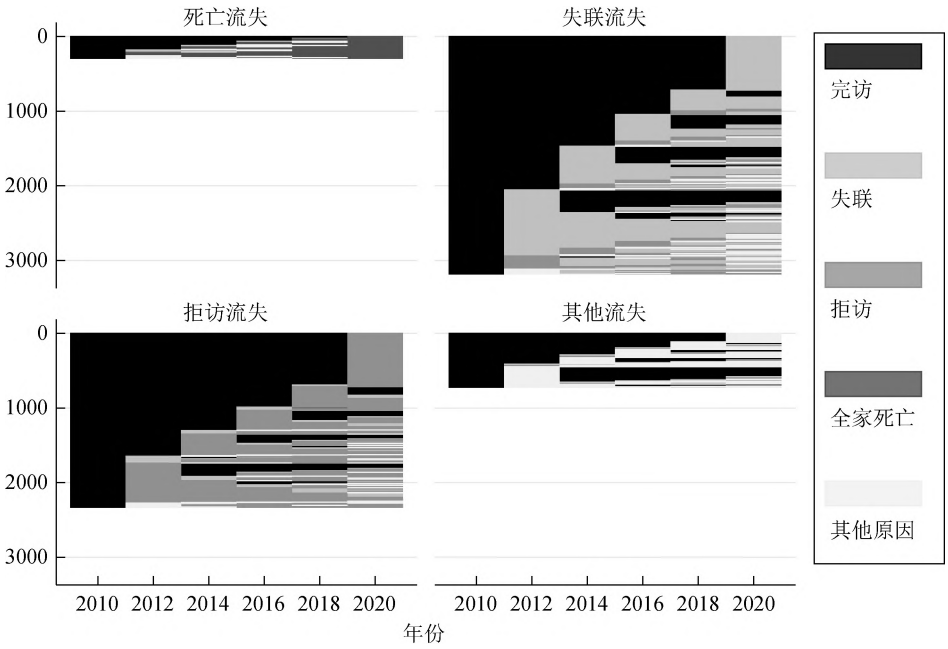

图 2　按流失原因分类的四种流失模式的应答序列

表 3　　按流失原因分类的各类家庭分布

类型	家户数（户）	占比（%）	非全程追踪家庭中的占比（%）
全程追踪	8435	56.4	
死亡流失	291	2.0	4.5
失联流失	3183	21.3	48.8
拒访流失	2330	15.6	35.7
其他流失	721	4.8	11.1
合计	14960	100.0	100.0

表 4　　按流失原因分类的各类家庭的追踪流失状态分布（%）

	首期流失	中途流失	冲击流失	摇摆流失	合计
失联流失	18.0	20.7	22.5	38.7	100.0
拒访流失	13.6	19.4	29.6	37.4	100.0
其他流失	3.3	13.2	15.3	68.2	100.0
死亡流失	29.6	50.9	8.2	11.3	100.0

1．家庭非全程追踪影响因素。

从表 5 列（1）关注家庭是否出现任意一种情况的流失即非全程追踪（1=是，0=否），结果显示，家庭结构、家庭经济状况、户主的社会经济特征、受访对象前期参访经历均会在不同程度上影响家庭是否能够持续参与访问。具体而言，家庭结构的紧密和完整对全程持续参与追踪有正向促进效果，与核心家庭相比，单亲家庭和单人家庭更难以全程追踪，而主干家庭则更不容易出现失访现象。居住流动性较低的家庭，如拥有自有住房或本地户口的家庭，全程追踪可能性显著较高，更不易发生流失。相较于城镇家庭，农村家庭对追踪调查的持续配合度较高。人均收入较低的家庭被全程追踪的可能性更高。男性户主家庭调查持续参与度会更高一些，户主年龄与是否持续参与访问呈现倒 U 型关系。在访问过程中表现出对调查感兴趣、户内个人问卷完访率高的家庭，后续持续接受访问的可能性更大；反之，对调查疑虑高的人员在后续追踪过程中更倾向于退出。

与国际其他追踪调查项目相比，运用 CFPS 数据甄别出的流失因素既有一致之处，亦存在不同之处。家庭流动性高导致流失加剧、受访家庭在往期调查中的态度和表现影响流失，这些发现均与已有研究较为吻合。但家庭及个人属性对全程追踪的影响则呈现出男性户主、低收入群体、农村地区样本更配合的特点，这是与其他国外调查受访家庭参与配合度不一致之处。 基于欧洲社区及家庭追踪调查（ECHP）的研究表明社区事务参与程度高的群体对调查的接受度高（Groves et al.，2000），PSID 受访者中高学历人群参与调查的比例更高（Lillard & Panis，1998），这些发现背后的逻辑是地位较高的群体对社会事务更感兴趣，也会对参与调查更热情，具有这些特征的受访者会将参与调查视为社会参与的行为。但在现阶段中国，参与调查更应理解为是一种遵从行为，社会经济地位较低的群体对调查的配合度更高，也相对更容易接纳陌生的访员。

2．家庭流失类别影响因素。

表 5 列（2）~（5）以全程追踪作为参照类，针对每种流失类别分别建立 Logit 回归模型。分析发现，每种流失类型的影响因素之间既有共性也有差异性。共性的方面体现在家庭居住的流动性显著提高了各类流失的可能性，城镇家庭比农村家庭更易流失，这些发现充分体现了在我国人口流动和城市化背景下持续追踪调查的难度。此时，基线调查中受访家庭中个人问卷的完访率可以很好地预测未来持续参与的可能性。

表 5　　影响追踪流失类型因素的 Logit 回归模型的平均边际效应估计

	（1）	（2）	（3）	（4）	（5）
	非全程追踪	首期流失	中途流失	冲击流失	摇摆流失
新分裂家庭数量	−0.215***		−0.183***	−0.101***	−0.135***
家庭规模	−0.005	−0.022***	−0.006	−0.006	0.000
儿童数量	−0.002	0.011*	−0.010	−0.000	0.002
老年人数量	−0.004	−0.005	−0.005	0.014*	−0.012
家庭类型（核心家庭=0）					
主干家庭	−0.023*	−0.003	−0.008	0.007	−0.023*
联合家庭	0.023	0.010	0.026	0.014	0.020
破损核心家庭	0.102***	0.025	0.063**	0.025	0.078**
单人家庭	0.168***	0.086***	0.112***	0.059*	0.101***
家庭是否有本地户口居民	−0.173***	−0.108***	−0.080***	−0.041	−0.100***
现住房是否自有住房	−0.089***	−0.050***	−0.043***	−0.030**	−0.070***
家庭城乡所在地（农村=0）	0.100***	0.082***	0.078***	0.021**	0.063***
家庭人均净收入的对数	0.010**	0.012***	0.009**	−0.004	0.007
户主性别（女=0）	−0.022*	−0.023***	−0.005	0.007	−0.020*
户主年龄	−0.018***	−0.008***	−0.012***	−0.004*	−0.012***
户主年龄平方项	0.000***	0.000***	0.000***	0.000**	0.000***
户主受教育年限	0.001	0.005***	0.002*	−0.002*	0.001
家庭成员平均配合度	−0.000	−0.007*	−0.006	0.012**	−0.003
家庭成员平均调查兴趣	−0.008*	−0.004	0.003	−0.006	−0.007
家庭成员平均调查疑虑	0.010***	0.008***	0.001	0.003	0.009***
家庭个人问卷完访率	−0.308***	−0.144***	−0.131***	−0.115***	−0.278***
样本量	14400	9133	9568	9764	10751

注：1．*p<0.05，**p<0.01，***p<0.001；2．由于首期流失的家庭没有新分裂家庭，列（2）不含“新分裂家庭数量”变量。下同。

具体来看，首期流失和中途流失这两类单调流失更可能发生在家庭人均收入较高、户主受教育年限较高、城镇地区的家庭，而家庭人均收入和户主受教育年限对另外两类流失类型则无此影响，家庭居住地的影响在另外两种流失类型家庭的效应也相对较小，可见，社会经济地位较高的家庭的流失更难以逆转。对于首期流失，拥有未成年子女的家庭和规模较小的家庭更可能在完成基线访问后便退出，对调查的配合度和疑虑对首期流失的影响也更突出，可见，取得受访家庭最初的信任和配合这些“第一印象”是能够持续追访的重要因素。前面提到，冲击流失是在社会外部环境变化迫使访问模式发生变化后导致的流失，这类流失有一定的特殊性。列（4）显示，家庭居住的流动性对冲击流失的影响较小、往期调查的疑虑也不会显著增加冲击流失的可能性，甚至过去配合度更高的家庭显著更可能发生冲击流失，这一定程度上印证冲击流失是外因所致。或者说，若访问模式不转变，冲击流失样本存在较大可能性能够持续接受访问。其中，家庭老人数量多、受教育年限低的家庭显著更可能发生冲击流失，这很可能是由于这类人群更适应于面对面访问，改为电话访问后遇到了联系不上和沟通的困难，他们受访问模式转变的影响更大。对于摇摆流失，最主要的影响因素是家庭居住的稳定性，拥有本地户口及自有住房的家庭在失访后仍有较高可能性回归。换言之，只要家庭不迁移后续始终存在数据采集的可能性。

3．家庭流失原因影响因素。

本文这部分的分析聚焦非全程追踪样本，仍以全程追踪家庭为参照组，将是否因某类原因导致的流失设为因变量进行二元 Logit 回归分析。表 6 报告了各模型的平均边际效应。

表 6　　影响按原因分类的流失类型的因素的 Logit 回归模型的平均边际效应估计

	（1）	（2）	（3）	（4）
	死亡流失	失联流失	拒访流失	其他流失
新分裂家庭数量		−0.188***	−0.142***	−0.070***
家庭规模	−0.025***	−0.016**	0.011*	0.003
儿童数量	−0.001	0.007	−0.010	−0.001
老年人数量	0.006*	−0.006	−0.003	0.005
家庭类型（核心家庭=0）				
主干家庭	0.010	−0.013	0.001	−0.028***
联合家庭	0.001	0.051**	0.003	−0.012
破损核心家庭	0.016	0.088***	0.037	0.051*
单人家庭	0.026**	0.091***	0.055*	0.088***
家内是否有本地户口居民	0.010	−0.184***	−0.085***	−0.038*
现住房是否自有住房	−0.012**	−0.081***	−0.056***	−0.013
家庭城乡所在地（农村=0）	−0.002	0.065***	0.119***	0.002
家庭人均净收入的对数	0.000	0.005	0.025***	−0.005*
户主性别（女=0）	0.013***	−0.011	−0.027**	−0.010
户主年龄	−0.002*	−0.012***	−0.005*	−0.009***
户主年龄平方项	0.000***	0.000***	0.000**	0.000***
户主受教育年限	−0.001*	0.000	0.005***	−0.002**
家庭成员平均配合度	−0.001	−0.006	0.000	0.005
家庭成员平均调查兴趣	−0.001	−0.001	−0.006	−0.008***
家庭成员平均调查疑虑	0.000	0.004	0.011***	0.005*
家庭个人问卷完访率	−0.032***	−0.248***	−0.201***	−0.141***
样本量	8540	11299	10349	8938

因全家死亡流失的家庭以空巢老年人家庭为典型：家庭规模小、老年人数量多、单人家庭。相较于全程追踪家庭，失联流失的家庭规模较小、家庭结构更特殊，家庭居住不稳定性对失联流失的影响更强。拒访流失家庭的典型特征是居住在城市、人均收入更高、户主受教育年限高。具备这些特征的家庭或许更注重隐私、时间成本更高，对参与调查的态度更为抵触，这从调查疑虑程度对拒访流失的影响更大中也得以体现。

（四）预测追踪流失

本文上述对流失类型的分析是否有助于预判追踪调查未来的流失状况？本部分尝试运用 2010—2020 轮次 CFPS 家庭的流失类型对 2022 轮次家庭层面的应答情况进行预测，并与该轮次真实追踪结果进行比对，以检验预测的准确性。

表 7 展示的是根据 CFPS 2012—2020 按追踪状态聚类所得的五个组别在 2022 轮次的预测完访率及实际完访率。如前文所言，首期流失和中途流失呈现一去不回的单调流失特性，预测完访率为 0%，从现实情况来看，这两类在 2022 轮次追踪中接受访问的比例也极低，合计不及 5%。冲击流失是由于模式转换导致的短期流失，在 2022 轮次保持电访为主访问模式不变的情况下，根据 2012—2018 模式不转变情况下的平均跨轮逆转率，预测会有 30%的家庭能够逆转挽回，实际上的确有约 36%家庭在 2022 轮次回归。摇摆流失组别的最大特征就是访问状态在完访、失访间来回切换，有近半数家庭在 2022 轮次接受访问也比较符合预期。全程追踪样本中有约 15.9%未接受访问，这个数字与 2020 及之前轮次全程追踪样本平均失访率 15.45%基本相当。也就是说，在访问模式不调整的情况下，全程追踪样本组别的流失率是相对稳定的。

整体而言，预测完访率与实际完访率高度吻合，表明综合运用前期追踪状态信息能够较为准确地预测后续轮次的应答状况。

表 7　　追踪流失类型与 2022 追踪状态

流失类型	2022 追踪状态		完访率（%）	
	失访数	完访数	预测	实际
首期流失	984	16	0.0	1.6
中途流失	1258	97	0.0	7.2
冲击流失	988	553	30.0	35.9
摇摆流失	1411	1218	50.0	46.3
全程追踪	1343	7092	85.0	84.1
小计	5984	8976	60.0	60.0

四、结　论

样本流失是固定样本追踪调查面临的普遍问题与复杂挑战，其原因涉及外部社会环境的变化、受访对象客观特征与主观因素、项目设计及执行因素等多种方面。本文以 CFPS 为例，从截面及追踪两个角度对家庭样本流失的规模、趋势、流失模式分类及影响因素展开了分析。

本文发现，与国际上其他追踪调查的一般规律相一致，CFPS 家庭样本流失的规模总体上呈现为随访问轮次增加而逐渐上升的趋势，且体现为早期流失较快，后续流失趋于稳定（Friedel & Birkenbach，2020）。从 2010 到 2020，CFPS 基线样本家庭中有超过半数持续参与每轮访问，在非全程参与的家庭中，流失可以分为不同类型。当社会外部环境发生变化迫使项目访问模式从面访为主切换至电话访问主导时，会导致一部分在面访模式下可能持续追踪的家庭失访，这种在国

内特定时期产生的冲击流失类型给认识访问模式与样本流失之间的关系提供了一个特殊的经验证据。基于欧洲 ECHP 和中国健康与营养调查（CHNS）的研究发现，非全程追踪家庭中单调流失更为常见（Nicoletti & Peracchi，2005；梁玉成，2011），但本文发现，在 CFPS 中非单调流失是最常见的类型，约有四成的非全程追踪家庭在不同轮次间属于在完访和失访之间切换的摇摆流失。这部分摇摆流失样本在一定程度上维持了 CFPS 的截面应答率。这一发现表明，追踪调查采取将往期失访样本保留在追踪样本库的做法而不是过早放弃连续两轮失访的样本，会明显有助于减少样本损耗。

导致 CFPS 家庭样本流失的两大主要原因是无法获得受访对象的有效联系方式及受访对象主动退出，但这两类流失样本在首次失访后都有一定比例在未来的轮次中回归。进一步分析影响家庭样本非全程追踪及失访原因的因素发现，分裂家庭数量较多、居住状况稳定、收入水平较低的家庭持续参与调查的可能性较高。社会地位较高的家庭更可能单调流失，首期流失与受访对象对调查的配合和疑虑关系较强，冲击流失则更多产生于原先面访调查中容易维护的群体难以适应访问模式的转换。分不同流失原因来看，与失联流失相关的典型特征是家庭规模小、居住流动性强，但这类家庭参与访问的积极性并不低；拒访流失则更多表现为居住在城市、高收入家庭、高学历户主、参访意愿较低等特征。运用 CFPS 数据甄别出的影响样本流失的因素主要集中在居住流动性及参访意愿两方面，未来可以利用这些特征信息研判追踪执行难度并设置激励机制，这对其他追踪调查也有参考价值。最后，本文发现运用 CFPS2010—2020 前六期追踪状态聚类所得的流失模式能够较好预测 2022 第七期的追踪状态，由此进一步说明对往期追踪状态的深入分析和分类能够为评估未来追踪的难度提供有效参考。

本研究对固定样本追踪调查提供的重要经验是，维护受访对象有效联系方式是提升追踪成功率的重要方向。组建受访群体私域社群，非调查季通过适当频次接触或不定期组织各类维护活动保持受访家庭与项目的黏性，设置信息更新奖励等举措将有助于维护受访样本有效联系方式。常规激励举措难以维系高知受访对象持续参与追踪的热情，以适当的形式展现项目成果、体现受访者在数据采集中的价值及贡献，引入保密性更高、侵入性更低的数据采集手段，或许都是保留高知群体的有效手段。当然，执行过程中也可以通过更换访员、更换访问方式、利用相关人员劝说等方式尽力尝试劝访。最后，如果选择性样本流失不可避免，定期评估样本代表性，通过样本更新、补充或替换等方式保障追踪数据质量是重要且必要的。

深入了解样本流失模式对于评估追踪调查数据质量以及采取相应措施来减少流失对数据质量的负面影响非常重要。本研究在这方面进行了初步探索，但依然存在一些可以继续研究的方向。首先，现有研究虽然分析了样本流失影响因素，但仍可以探讨这些因素导致样本流失的具体机制。其次，本文是以 CFPS 项目的样本流失模式作为分析内容，国内同类型的其他调查的流失模式是否也类似还有待进一步论证。最后，本文评估样本流失的模式、演变规律及影响因素，但未涉及样本流失对数据质量的影响，尤其是对研究核心变量的影响还有待评估。未来仍需要不断关注和探索新的方法和技术，通过认真拆解分析样本流失的原因及机制，动态调整追踪调查项目执行方案，采取有效措施切实降低样本流失率，提升调查的质量和效果。

参考文献

[1] 梁玉成. 追踪调查中的追踪成功率研究——社会转型条件下的追踪损耗规律和建议[J]. 社会学研究, 2011(6): 132-153.

[2] 齐嘉楠. 无回答频次的影响因素研究及追踪措施探讨[J]. 西北人口, 2016(6): 1-9.

[3] 任强, 谢宇. 对纵贯数据统计分析的认识[J]. 人口研究, 2011(6): 3-12.
[4] 孙妍, 吴琼, 张春泥. 中国家庭追踪调查: 设计理念及数据运用问题[J]. 调研世界, 2024(1): 4-14.
[5] 孙妍, 邹艳辉, 丁华, 等. 跟踪调查中的拒访行为分析——以中国家庭动态跟踪调查为例[J]. 社会学研究, 2011(2): 167-181.
[6] 王俊, 金勇进, 王亚峰, 等. 追踪调查中的样本更新问题研究——部分国际追踪调查的实践经验总结及思考[J]. 统计研究, 2024(1): 124-134.
[7] Behr A, Bellgardt E, Rendtel U. Extent and Determinants of Panel Attrition in the European Community Household Panel[J]. European Sociological Review, 2005, 21(5): 489-512.
[8] Fitzgerald J, Gottschalk P, Moffitt R. An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics[J]. The Journal of Human Resources, 1998, 33(2): 251-299.
[9] Frankel L L, Hillygus D S. Looking Beyond Demographics: Panel Attrition in the ANES and GSS[J]. Political Analysis, 2014, 22(3): 336-353.
[10] Friedel S, Birkenbach T. Evolution of the Initially Recruited SHARE Panel Sample Over the First Six Waves[J]. Journal of Official Statistics, 2020, 36(3): 507-527.
[11] Goebel J, Grabka M M, Liebig S, et al. The German Socio-Economic Panel(SOEP)[R]. Jahrbücher für Nationalökonomie und Statistik, 2019, 239(2): 345-360.
[12] Groves R M, Couper M P. Nonresponse in Household Interview Surveys[M]. New York: John Wiley and Sons, 1998.
[13] Groves R M, Eleanor S, Amy C. Leverage-Saliency Theory of Survey Participation[J]. Public Opinion Quarterly, 2000, 64(3): 299-308.
[14] James N D. Respondents for Nearly Three Decades: How do Loyal Sample Members Differ from Others?[J]. Survey Research Methods, 2023, 17(1): 15-36.
[15] Lillard L A, Panis C W A. Panel Attrition from the Panel Study of Income Dynamics: Household Income, Marital Status, and Mortality[J]. Journal of Human Resources, 1998, 33(2): 437-457.
[16] Lipps O. Attrition of Households and Individuals in Panel Surveys[R]. SEOP papers on Multidisciplinary Panel Data Research, 2009.
[17] Lynn P, Burton J, Kaminska O, et al. An Initial Look at Non-Response and Attrition in Understanding Society[R]. Understanding Society Working Paper Series, 2012.
[18] Michaud P C, Kapteyn A, Smith J P, et al. Temporary and Permanent Unit Non-response in Follow-up Interviews of the Health and Retirement Study[J]. Longitudinal and Life Course Studies, 2011, 2(2): 145-169.
[19] Nicoletti C, Peracchi F. Survey Response and Survey Characteristics: Microlevel Evidence from the European Community Household Panel[J]. Journal of the Royal Statistical Society Series A: Statistics in Society, 2005, 168(4): 763-781.
[20] Potârcă G, Mills M, Lesnard L. Family Formation Trajectories in Romania, the Russian Federation and France: Towards the Second Demographic Transition?[J]. European Journal of Population, 2013, 29(1): 69-101.
[21] PSID-2021 Main Interview User Manual: Release 2023[R]. Institute for Social Research, University of Michigan, June, 2023.
[22] Smeeding T M. The PSID in Research and Policy[J]. The Annals of the American Academy of Political and Social Science, 2018, 680(1): 29-47.
[23] Uhrig S C N. The Nature and Causes of Attrition in the British Household Panel Survey[R]. ISER Working Paper Series, 2008.
[24] Watson D. Sample Attrition between Waves 1 and 5 in the European Community Household Panel[J]. European Sociological Review, 2003, 19(4): 361-378.

The Patterns and Trends of Sample Attrition in Large-scale Panel Surveys

—A Case Study of China Family Panel Studies (CFPS)

Sun Yan[1] Wang He[2] Pan Xiuming[2] Zhang Chunni[2]

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

Abstract: To better understand the characteristics and trends of sample attrition in household panel surveys in China, this paper analyzes data from the China Family Panel Studies (CFPS) spanning from 2010 to 2022. The results indicate that the scale of cross-sectional sample attrition gradually increases with the number of follow-up waves, and the attrition rate of CFPS household samples initially experiences a rapid increase and then stabilizes. Among the non-fully tracked sample households, the predominant type is swing attrition, where households toggle between non-response and response status across waves. This implies that retaining non-responsive samples can help mitigate sample loss. Changes in interview modes, driven by shifts in the external social environment, can also contribute to rapid sample attrition. Loss of contact and refusal to participate are the primary factors leading to sample attrition. The difficulty of conducting follow-up interviews in subsequent waves depends on the reason a household did not receive a previous interview. For sample attrition due to loss of contact, the challenge lies in obtaining effective contact information, while for samples lost due to refusal to be interviewed, the difficulty is even greater. Households with larger sizes, lower levels of residential mobility, and those from rural and low-income backgrounds demonstrate a greater propensity to participate in the panel survey. Additionally, respondents' attitudes toward the surveys in previous waves also influence subsequent attrition. Practically, utilizing attrition groups classified by previous response status can yield more accurate predictions of follow-up rates in the next wave, thereby providing valuable and effective reference information for sample maintenance and on-site implementation.

Key words: Panel Survey; Sample Attrition; Sequential Cluster Analysis; China Family Panel Studies

（责任编辑：曹　麦）