

并行数据与社会调查质量探讨

任莉颖^a,严洁^b

(北京大学a.中国社会科学调查中心;b.政府管理学院,北京 100081)

摘要:并行数据是关于社会调查过程的数据。文章详细介绍了并行数据的概念、特征、内容和采集方式,并且以中国家庭动态跟踪调查为例,探讨了并行数据在中国社会调查中的应用状况与应用价值。提出了并行数据研究和应用的前景和挑战。

关键词:并行数据;社会调查;调查管理;质量监控;CAPI

中图分类号:C915 **文献标识码:**A **文章编号:**1002-6487(2014)06-0027-05

0 引言

近些年来,计算机技术和信息通信手段的突飞猛进不仅冲击着人们生活的方方面面,也带来了社会抽样调查的技术创新。首先,调查执行的模式变得多元化,通过与通

信技术系统的紧密结合,出现了计算机辅助的一系列数据采集模式,如计算机辅助电访、计算机辅助面访,以及计算机辅助网访等。在计算机辅助调查模式被不断应用的同时,一类关于调查过程的数据逐渐浮出水面,并引起了社会抽样调查研究者的重视——这就是并行数据。

并行数据起初只是作为调查管理之用,后来研究者发

基金项目:国家自然科学基金资助项目(71171004)

作者简介:任莉颖(1973-),女,河北三河人,博士,副研究员,研究方向:社会调查方法与数据挖掘。

严洁(1973-),女,吉林通化人,博士,副教授,研究方向:社会调查方法与数据挖掘。

表1 二级指标权重分配

一级指标代码	二级指标代码	党政机关 微博权重	党政干部 微博权重
A	A1	50	40
	A2	20	25
	A3	20	25
	A4	10	10
B	B1	40	30
	B2	20	20
	B3	20	20
	B4	10	20
	B5	10	10
C	C1	40	30
	C2	25	25
	C3	20	20
	C4	5	15
	C5	10	10
D	D1	35	30
	D2	20	20
	D3	10	10
	D4	10	10
	D5	10	15
	D6	10	10
	D7	5	5

$$Z = \frac{x - \bar{x}}{\sigma} \quad (2)$$

式中, $\bar{x}$ 是全部单项二级指标数据的平均值, σ 是标准差。对得出的Z值进行线性变换,将其映射到[60,100]区间,得到实际单项二级指标标准分。

4 总结与展望

国家行政学院电子政务研究中心基于上述评估体系进行了2011年度的政务微博评估,鉴于目前政务微博客的发展实际,评估活动没有覆盖境内全部政务微博客网站,仅对在新浪网、腾讯网、人民网、新华网4家网站经过认证的32358个党政机构微博客和18203个党政干部微博客进行了分项评分和综合评分,分别给出了综合评分排名前100名的党政机构微博客和排名前100名的党政干部微博客。下一步将逐步扩大微博客评估的范围,细化微博客评估内容,逐步把政务微博客评估工作常态化,以促进我国政务微博客的健康可持续发展。

政务微博绩效高低的最终发言权归属于公众,公众对的满意程度应该是评估政务微博的重要指标,但由于条件和篇幅所限,本文未能就此展开全面和深入的讨论。此外评估不仅需要比较各微博的物理表征指标与服务内涵,还应将政务微博活动同其建设环境联系起来,更客观地评估人在政务微博的能动作用,以及政务微博作为建设项目的投入产出效率。

参考文献:

- [1]闫丰,常亚平.微博研究综述[J].情报杂志,2011,30(9).
- [2]刘宁雯.中国政务微博研究文献综述[J].电子政务,2012,(6).
- [3]国家行政学院电子政务研究中心.2011年中国政务微博客评估报告[R].2012.
- [4]王益民,丁艺,胡红梅.我国政务微博发展现状与对策[J].行政管理改革,2012,(10).

(责任编辑/亦民)

现了并行数据对数据质量的实时监控和改进、对数据质量评估、对测量误差矫正,以及对无回答插补技术等具有非常重要的作用。如使用电话联系记录和联系结果数据来研究拒访现象^[1];使用访员观察记录来评估由于拒访造成的数据误差;使用访问录音文件来筛选说服参与访问效果较佳的访员的声音特征,或用于评估语音交流导致的特定问题上的测量误差;使用采访用时数据和键盘痕迹文件来对网络调查进行改善、用于评估访员绩效,或用于对调查工具的效果进行测试。

对并行数据的研究提升了整个定量数据采集行业改进数据质量的信心。国际上对并行数据的研究刚刚兴起,国内的学术机构、政府相关部门正在努力加快缩小与国际调查领域水平的差距,计算机辅助调查技术也开始在国内得到广泛应用。在这种形势下,对并行数据与数据质量管理的研究成为迫切的现实需求。

本文将讨论并行数据在中国的应用情况及其对社会调查质量的影响,通过梳理并行数据的概念、特征、内容和采集方式,并以中国家庭动态跟踪调查为例,在探讨并行数据在中国社会调查中应用效果的基础上,提出并行数据研究的前景和挑战。

1 并行数据的定义

“并行数据”一词最早由 Couper 在 1998 年美国统计协会的联合统计会议上的研究报告中提出^[2]。他把社会调查的数据分为三大类:调查数据是关于研究现象的数据,也就是采集的调查问卷数据;元数据是关于调查数据的数据,有助于解读调查数据;而并行数据是关于调查过程的数据。自此,这个概念迅速为社会调查的业内人员广泛接受,并且成为研究热点。

有的学者认为并行数据是辅助性数据的一种,而有的学者则对二者做了明确区分,并且这种区分有助于更好地理解什么是并行数据。

Loosveldt 和 Beullens 提出了辅助性变量的三个特征:

(1)辅助性变量来源于社会调查或调查数据采集过程以外,其数值在实地执行过程中保持不变(如注册数据)。

(2)辅助性变量由访员或调查机构在实地执行过程中采集,其数值在执行过程中保持不变(如访员记录的社区特征)。

(3)辅助性变量由访员或调查机构在实地执行过程中采集,其数值的变化不受调查机构与受访对象的互动影响。

简单地说,辅助性变量最重要的特征就是在实地执行过程中数据值保持不变,或者即使有变化,也与调查机构和受访对象之间的互动无关。而并行数据主要记录的就是执行过程中的动态信息,其本身就是在调查机构与受访对象的互动过程中产生。比如访员的个人背景信息,单独看是一种辅助性变量,因为这些数据值不受调查过程的影响。然而在执行过程中,哪个访员采访哪个调查对象是由

调查机构进行分配,这就使得对应某一个调查对象,其访员是不确定的,那调查对象所经历的访员个人背景也是不确定的。所以当附着在样本层面上时,访员背景信息就是一种并行数据。

因此,并行数据记录的是调查过程中的动态变化,并且存储为可供分析的数据信息。

2 并行数据的内容

Couper 在提出并行数据的概念时主要是针对计算机辅助调查模式下采集到的一类新型数据:键盘操作痕迹文件,采访时长,以及详细的联系记录。近几年来,随着学者们的关注增加,并行数据的内容也得到不断的拓展。但普遍认为,并行数据可包括以下几大类信息:

(1)联系记录数据。这是最常见的一类并行数据,主要包括访员每次联系受访人的日期、时间、次数、联系结果等。这类数据在传统的纸笔访问方式中已有应用。

(2)访员观察数据。根据研究者的需求,访员观察数据可以包括多方面内容,如对拒访程度的观察,对社区环境的观察、对采访情境的观察、对受访对象本人及家庭的观察等等。

(3)访问录音数据。在受访人知情或同意的情况下,计算机采访系统可以对采访过程进行全程或部分录音,并且可以分问题条目存储,以便录音数据的查找。

(4)访问痕迹数据。访员或受访人在使用计算机设备进行采访或自答时,采访系统可以自动记录键盘的敲击顺序、鼠标移动的轨迹、每道问题的起始/结束时间,功能键的使用情况等等。

(5)样本调配数据。调查执行管理者要将所有的受访样本分配给多个访员,有时还需要将样本在访员中调配。计算机调查管理系统也可以准确地记录下每个样本对应的访员姓名或代码,以及样本调配的相关信息。

并行数据的内容和采集会受到调查模式、关注的过程,以及研究焦点的影响而呈现多样化。

在纸笔调查的模式下,并行数据多通过访员记录来采集,不仅信息量少,而且数据质量难以保证。如北京大学中国国情调查中心从 1995 年就开始实施的“北京社会经济发展年度调查”,在问卷的封面就包括了以下各类信息的记录:(1)采访完成日期;(2)采访用时;(3)整理问卷用时;(4)采访员编号;(5)采访员签字;(6)验收员签字;(7)验收日期;(8)录入员 1 签字;(9)录入员 2 签字;(10)录入完成日期;(11)第一、二、三次采访日期;(12)每次另约时间;(13)每次接触后的访问状态。在问卷的封底则包括了详细的采访记录,涉及访员对采访对象的兴趣、可信度、理解水平等方面的观察,对采访对象家庭的观察,对采访对象居住环境的观察,以及对采访场景的观察等。^[3]

相比之下,计算机辅助调查模式在微观并行数据的采集上显示出了纸笔调查模式所缺乏的优势。这些并行数据不仅包括在样本层面上的联系记录、访员观察等数据,

而且可以详细到每个问题的采访用时,每个问题的采访录音,甚至到每一次键盘或鼠标操作的访问痕迹数据。在计算机辅助网络调查中,还可以对视觉进行追踪记录。同时,由于这些访问痕迹数据和样本调配数据由系统自动存储,排除了由访员主动采集而带来的录入误差。而对于主动采集的并行数据,如联系记录、访员观察,和访问录音数据,访员也可以借助系统的管理和提示而提高数据质量。

并行数据是关于过程的数据,因而在社会调查中的每一个过程,如抽样、问卷设计、访问管理、以及后期的数据处理过程都会产生大量的并行数据,这时识别并挖掘有用的并行数据是研究者们的任务。具体的做法是首先绘制流程图,列举出主要的环节,并标识各环节间的路径关系。然后,根据流程图的指示,确定各环节可能产生的并行数据,再进行有目的的采集。

在 Groves 等制作的总体调查误差框架图上^[4],不仅指示了误差种类和来源,也显示了测量和抽样两个过程的主要影响环节。每个环节,或者每类误差,都会涉及的不同的并行数据如图1。以抽样过程为例,首先在制作样本列表来选取抽样框的过程中可以产生多种并行数据。对于地址抽样来说,可能会包括绘图员的性别、年龄、职业等背景特征,绘图所需的时间,绘图员观察到的当地居住格局、地貌特征等。采访过程中会产生大量并行数据,如采访键盘操作信息,访员特征、联系记录,访员观察等等。数据处理过程也可以采集到用来编辑数据的键盘操作信息等。

研究者的研究焦点不同。其关注的并行数据也会有所不同。并行数据的种类很多,内容也很丰富,然而哪种并行数据被定义并采集很大程度上取决于研究者的研究兴趣。图1显示不同的误差种类可以选取对应的并行数据进行辅助研究。例如对于关注测量误差的研究者,他们可能会注重采集访员采访时的键盘操作信息,来帮助识别问题设计上的疏误;也可能会关心由于访员特征而影响受访人应答时的心理,以至于带来测量误差。因此,从这个意义上来讲,前面提到的主要并行数据种类更多地是从研究使用角度来判断其重要性。有的并行数据可以有多个用途,所以会被普遍重视和采集;而有的并行数据针对性很强,则会被研究者们刻意采集。并行数据是一个宝藏,会根据研究需求而不断被开采,提供宝贵的辅助信息。

3 并行数据的应用

并行数据近年来得到调查方法研究者的广泛重视,其应用领域也在不断拓展。在最初阶段,研究者们主要在调查结束后对并行数据进行分析,期望对调查中出现的各种误差有所调整,或在将来的调查中能够有效控制。

如在对无应答误差的研究上,图1列举了三类并行数据——访员特征、联系记录,和访员观察。这些并行数据的共同特征是包括了所有受访人——无论是否应答——的相关信息。如联系记录数据,包括了访员与所有受访人的联系时间及联系结果,这就为研究无应答提供了重要的

辅助信息。早在1980年,Drew和Fuller就提出了在联系记录数据基础上对无应答现象进行模型预测^[5]。于是,有学者开始尝试把联系记录的信息应用到权重计算上,来调整由于无应答而带来的估计偏差。或是通过对联系记录或访员观察的分析,寻求降低无应答的途径,如选取最佳联系时机,减少拒访和无联系比例,预测无应答等。

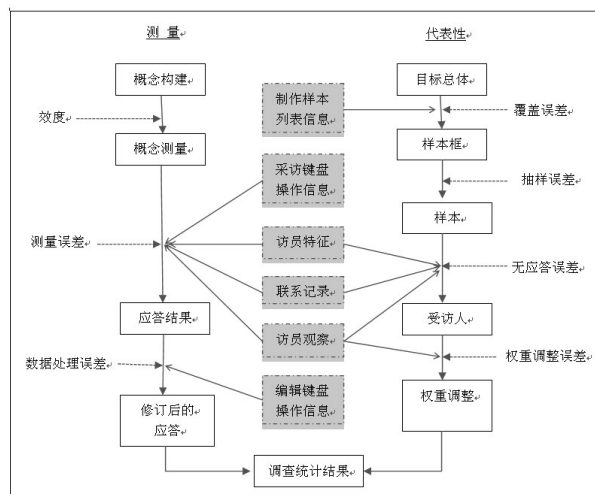


图1 总体调查误差:种类、过程及并行数据

在对测量误差的研究上,联系记录、访员观察和访员特征同样是重要的信息资源。如有学者发现在邮寄问卷调查中,早收到的调查问卷与晚收到的调查问卷会表现为不同的测量误差。访员观察到的采访场景会对受访人的应答有影响,访员特征,如性别、年龄,甚至语音特征都会影响到问题的测量。此外,计算机辅助模式的调查采集到了采访过程中键盘操作的详细信息,这些数据可以帮助设计者在预调查阶段了解到问卷设计中的问题,从而有效地改善问题质量通过对受访人回答问题的时间分析,还可以发现受访人在表达观点时的顾虑。

计算机辅助调查模式的广泛应用及通信手段的迅速发展实现了并行数据的有效采集和及时传输,于是并行数据的应用从调查后的评估和研究转向调查期间的实时监测,继而社会调查研究的焦点也从事后调整转移到过程质量和成本控制。在这方面比较有代表性的研究就是 Groves 和 Heeringa 提出的回应式社会调查设计(Responsive Survey Design)^[6]。Groves 和 Heeringa 指出近些年来社会调查在设计 and 实施过程中面临着许多不确定性,如受访对象越来越不愿意参与调查,于是调查机构采取各种方法来提高访到率,致使数据采集过程的成本增加,而且很难控制预算。与此同时,计算机辅助调查技术得到推广,这种调查模式下可以采集到大量记述调查过程的并行数据,研究者们通过对这些并行数据和调查数据的综合分析,可以在数据采集过程中及时地改变调查设计,达到有效地提高成本效益和测量精度的目的。

回应式社会调查设计将数据采集过程分为几个阶段。在第一个阶段同时采用多种设计方案,如不同金额的受访人酬金,不同次数的联系尝试,或使用不同长度的问卷等。在进行过程中需要采集大量的并行数据,如访员用

在联系受访对象、与受访对象交谈,以及采访受访对象的时间等。研究者们要在对这些并行数据分析的基础上,充分权衡成本和调查误差的代价,来决定数据采集第二阶段所要采用的设计方案。第三或后面几个阶段一般是在成本许可的程度上来进一步降低无应答的情形,获取更准确的统计数据。在整个过程中,并行数据的采集和分析至关重要,直接影响到研究者们对下一阶段所要采用的调查策略。

此外,并行数据还可以应用在访员培训和访员访问绩效的评估,采用统计过程控制方法测量和分析执行过程中的变异情况,以便及时识别异常情况,并对访员行为进行干预,以及缺失信息的比例来监控调查数据的质量等。

在“中国家庭动态跟踪调查(CFPS)”中,并行数据被应用到问卷设计、调查管理、质量监控和数据处理等多个工作环节。下面从问卷设计、调查管理、质量监控三个环节,讨论并行数据的应用情况、效果,以及遇到的困难与挑战。

3.1 关于职业问题的设计与编码

CFPS在2010年初访调查中不仅要求访员记录下职业的详细描述信息,还要根据职业编码字典对职业问题进行现场编码。职业编码字典包括八大类,每类四个级别,共计595个职业代码。这些职业代码在计算机访问系统中设计为树形结果,访员可以逐级展开,最终锁定职业的最终编码。访员实地编码的优势是遇到含混的职业信息访员可以及时追问,同时访员因为了解编码的过程,会在一定程度上提高职业描述信息的记录质量。当然,这种编码方式还能提高数据的时效性,减少编码的成本。然而访员编码的质量如何却一直令研究者们担忧。

于是,在数据采集完毕后,调查中心又从北京大学招募职业编码员,对成人问卷中的个人职业问题重新编码。在编码流程上采用了双向独立验证并判定的质控方式,每一个职业问题条目都要经过2~4名编码员的判断,最后确定这些职业问题的最终代码。

然后,我们将编码员的编码结果和访员的实地编码结果进行比较,发现二者之间有一定的差异,在职业大类的一级码上,一致率可以达到四分之三。然而在二级和三级码上,只有三分之二左右的结果相同。到最精细四级代码上,一致率降到了三分之一。

分析访员实地编码过程,我们从并行数据中选取了几个指标来分析访员编码质量:

- (1)访员记录职业信息的字数;
- (2)访员实地编码的用时;
- (3)访员实地编码的条目次序号;
- (4)访员的年龄、性别和受教育程度

通过对这些并行数据进行分析,我们发现访员记录职业信息的字数对于编码的结果十分重要,同时访员自身的经验和素质也影响到编码质量^[7]。由于在调查过程中对访员因素的控制难度较大,于是根据这些研究成果,我们决定在问卷设计上放弃访员实地编码的问题,在访问系统中

增加访员提示,指导访员如何记述职业信息,同时在培训中强化职业信息的记述规范。

CFPS在2011年的跟踪调查中采用了编码员集中编码的方式。我们在访员实地编码系统的基础上开发了编码员编码系统,这样也可以采集到集中编码过程的并行数据,如每个条目编码的实际用时,编码员修订编码的痕迹数据,三名编码员编码结果的一致率等等。我们在此基础上设计了一些指标用来评估编码员工作绩效,如通过对编码员登记的工作用时和编码实际工作用时的比较,监督编码员的工作投入;对每个编码员的编码条目和一致率的分析,确定编码员的奖励标准等。

3.2 关于调查实施的访问管理

2010年CFPS初访调查的采访对象涉及全国25个省市的649个村居(村居问卷),19986个家户(家庭成员问卷和家庭问卷),还有这些家户中的所有成员(成人和少儿问卷)。调查中心采取集中管理的方式,负责实地执行的工作人员有主任一名,督导12名。2010年度的初访调查共招募了438名访员,这些访员大多数来自本次调查的目标区县,背景各异。如何有效地管理众多的访员,及时了解全国各地的采访进度成为CFPS调查管理的首要问题。

CFPS为此设计并采集了大量的并行数据,主要信息包括以下几类:

(1)样本访问状态,分为完成、联系不上、需要再次联系、拒访、特殊原因无法接受访问、全家外出、搬迁/拆迁等。因为在执行过程中,督导要对样本的访问状态进行掌控,所以还设计了一系列终极代码,表示该样本已成功完成或可以放弃访问。

(2)样本联系时间和次数,记录了访员每次联系受访对象的时间及汇总的次数。

(3)样本分配信息,即样本如何在访员之间进行分配或调配。

(4)数据发送/接收信息,采集完成访问的数据没有成功发送回总部的样本信息。

(5)质量监控和督导的信息,即质量监控的信息在督导和访员之间的反馈环。

基于以上信息,中心建立了每日更新的报告系统。报告系统中为调查管理制作了多个进度表,如针对项目研究人员制作了各类问卷访到率、拒访率的报告;针对执行主管制作了各类问卷的访问进度报告,全国完成访问的家户比例等;针对各级督导制作了每日完访家户比例,每日完访各类问卷总数,采访结果频数统计,日访问量和周访问量等,并且这些结果在省、村居和访员三个层面汇总报告。这样,在调查执行过程中,在总部的管理人员可以及时掌握到各地和各个访员的采访进度等情况。

CFPS要求访员在每天采访结束后立即将采访数据发回总部,然而由于网络原因,有些完成访问的数据没能成功发送,造成样本是完成访问状态,却没有收到对应的调查数据。也有时由于访员的疏忽,对于已经成功访问的样本没有插入表示完成状态的结果代码,这样虽然收到调查

数据,却会因为样本为非完成状态而没有成功提取。针对这些情况,我们必须对样本访问状态数据和调查数据接收数据进行比对,将不一致的情形及时报告给负责督导,然后通知访员采取相应操作,以避免数据丢失。

此外,督导们也很关注访员的联系时间。对于不合理的过早或过晚的联系时间,督导们要对相应访员给以警示。他们也会根据同一时间段联系次数与访到次数的比例,指导访员寻找最佳访问时机。督导们还会根据一段时期内访员完成的工作量来控制访员的工作投入程度。

3.3 数据质量的实时核查

数据质量的核查由于CAPI系统的使用实现了实时操作。在CFPS的实施过程中,由人为因素引起的系统误差是质量监控的主要目标。

系统误差的来源可以分为以下几种:

问卷设计不当产生的系统误差。如由于问卷中的问题设计不适用于受访者,导致受访者很难回答或者大量拒答,或者系统性错误跳转模式,选项不周延、选项分类效度太低等。

末端抽样精度不充分导致的系统误差。如在调查开始后才发现抽中了空宅、非住宅单位,或抽中的住户单位不符合事先设定的样本条件等。

访员行为不规范导致的系统误差。这部分误差来源比较多,比较常见的可以细分为:替换或访错地址;替换或访错住户;替换或访错个人,例如,访员请住户内的成员代答其他成员问卷,或访问了住户内的不符合调查资格的成员,或访员自填问卷等;破坏中立原则,引导受访者选择某些答案;用各种方式催促受访者,使受访者未加思考匆匆作答;通过跳转模式故意回避需要长时间作答的题组;访员自己有意地有选择地填答一些题目或选项;系统地读错题目或选错答案中的代码,或录入错误等。

数据汇总和整理过程中出现的系统误差。这部分误差由于使用了CAPI系统会大大减少,但是在对原始开放型问题的编码过程中,也可能产生系统误差。

CFPS调查质量监控的目标就是尽可能地控制和减少这些误差,具体监控方法和流程等内容依据上述误差的来源有所不同。

以访员臆答为例,在质量监控中,要监控的具体目标是故意为缩短访问时间而自行填答,所选择的监测题目均是访员无法通过自行判断能得出答案的题目。

针对访员臆答,CFPS调查通过电话回访、录音回放,访问过程键盘回放,以及对题目采访用时的统计分析来发现问题:首先,依据每个问题采访用时,通过与合理用时进行比较,识别出疑似臆答样本。第二步,对于疑似臆答样本进行录音回放,确认是否没有提问。接着,对于没有录音的样本,进行电话回访,询问受访者是否提问了某个题目。最后,对于既没有录音也没有联系电话的样本,进行访问过程键盘回放,以确认是否臆答。^[8]

对于确认为访员臆答的问卷则立即由督导反馈给访员,及时纠正其访问行为。从并行数据的使用效果来看,

由于能够及时获得这些数据,因此可以实现对访员的及时干预,数据显示首次干预能够使得访员臆答卷的比例显著降低。首次干预之前臆答卷比例的均值为0.255(标准误差为0.020),干预之后则下降到0.098(标准误差为0.018)。^[8]

4 并行数据研究的前景和挑战

以上例证能够让研究者们看到并行数据在中国社会调查中的应用效果,它对完善研究设计、优化调查管理、提高调查质量具有重要的价值,但是并行数据的采集、应用还存在着许多的困难和未被探索之处。

(1)并行数据研究面临的首要挑战是如何采集并数字化的问题。

以访问录音为例,从录音文件中可以识别出受访对象真实的回答、访员的提问方式、受访对象和采访员的互动模式、访问过程状况等等,但是这些文件面临着如何进行编码、如何智能地构建成数据库的难题。国内一些学者正在尝试通过录音来识别访员不自觉形成的各种诱导模式,进而来讨论其对元数据的影响,为了实现此目的,学者们不仅面临着数万条录音记录,而且面临着如何判断各种诱导模式和进行编码的难题。

(2)并行数据研究面临的第二个挑战是如何及时使用以督导访员行为的问题。

并行数据的价值之一在于是否能被及时应用,在访问过程中,如果能够及时地发现并行数据对元数据的影响,并将不利的影响反馈给访员,那么,才能有效地提高调查质量。但是,电话调查、网络调查的实施周期都比较短,面访调查虽整体周期较长,但是对于个体访员来讲也是一个较短的周期。如果要求在较短的周期内实现及时督导,那么就要求调查管理系统能够及时分析和处理这些并行数据信息。以访问痕迹数据来讲,目前这些文件还都处于一个个独立的文本文件,怎样能够将这些信息与访问质量相关联?这个工作不仅是编码的问题,更重要的是需要研究者们去发现、去检验访问痕迹数据与元数据质量的关联模式。与此相类似的研究工作包括录音文件中反映出来的访问互动模式与元数据的关系、样本联系信息与拒访率、访问环境与敏感问题的真实性等等。在发现关联模式的基础上,要解决的问题是如何实现智能化处理和信息传递与反馈的问题。这个工作需要研究者们、IT技术人员、调查实施人员的多方合作才有可能实现。

(3)并行数据研究面临的第三个挑战是如何应用它改善调查设计的问题。

这个挑战一方面是如何改善研究设计,包括问卷内容、结构、调查工具的设计等等;另一方面的挑战是是否需要或者如何改革调查实施模式,如前文所述,Groves和Heeringa提出的回应式社会调查设计已在海外获得了令人乐观的应用经验,但是对于它在中国的社会调查中是否适用?如何应用?这些问题都需要进行检验。这个过程中,对于并行数据的利用将成为关键的讨论对象,哪些并

行数据与哪一类题目的测量精度紧密相关? 哪些并行数据是受访对象合作的预测因子? 哪些并行数据与调查实施的效率有关联? 等等。研究者们将面临着将并行数据与元数据、与调查管理进行关联研究的大量工作。

并行数据对整个调查数据质量的影响是不可忽视的,它目前尚属于待挖掘利用的宝贵资源,它所涵盖的大量信息究竟能够怎样改善测量精度、能够如何优化调查管理、如何被智能化开发,以及如何能被及时应用将成为研究者们面临的紧迫课题。

参考文献:

- [1]M. Couper, L. Lyberg. The use of Paradata in Survey Research[C].In Proceedings of the 55th Session of the International Statistical Institute, Sydney, Australia,2005.
- [2]R. M. Groves, M. Couper.Non-Response in Household Interview Surveys[M],New York: John Wiley & Sons, 1998.

[3]有关此项目的具体情况请参看杨明等.1995~2004 北京社会经济发展年度调查数据报告[M].北京:北京出版社,2007.

[4]R. Groves, F. Fowler, M. Couper, R. Tourangeau, J. Lepkowski. Survey Methodology[M].New York: John Wiley, 2004.

[5]J. Drew, W. Fuller.Modeling Nonresponse in Surveys with Callbacks [C]. In Proceedings of the Section on Survey Research Methods of the American Statistical Association,1980.

[6]R. Groves, S. Heeringa. Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs[J]. Journal of the Royal Statistical Society, Series A ,2006,(3).

[7]任莉颖,邱泽奇,李力,严洁.社会调查中职业问题编码的方式与质量研究[J].浙江大学学报(人文社会科学版),2011,(5).

[8]严洁,邱泽奇,任莉颖,丁华,孙妍,社会调查质量研究:访员臆答与干预效果[J].社会学研究,2012,(2).

(责任编辑/亦 民)