社会调查中职业问题编码的方式与质量研究

任莉颖[1] 邱泽奇[2] 李 力[1] 严 洁[3]

(1. 北京大学 中国社会科学调查中心，北京 100871；2. 北京大学 社会学系，北京 100871；
3. 北京大学 政府管理学院，北京 100871)

[摘 要] 职业是社会科学研究的重要变量，然而社会调查中的职业编码很容易出现偏差。目前我国社会调查中主要采用访员实地编码与访问结束后由编码员进行集中编码两种方式。基于经验数据分析发现，这两种编码方式的结果存在较大的差异。这些差异一方面受访员职业信息的记录质量、访员编码经验及访员自身特征的影响，另一方面也与不同职业类别的编码难度有关。因此，在社会调查中要注意监控访员职业信息的记录规范，并采用有质控的编码员集中编码的方式来提高职业问题编码的数据质量。

[关键词] 社会调查；职业编码；数据质量

Methods and Quality of Occupational Coding in Social Surveys

Ren Liying[1] Qiu Zeqi[2] Li Li[1] Yan Jie[3]

(1. *Institute of Social Science Survey, Peking University, Beijing* 100871, *China*;
2. *Department of Sociology, Peking University, Beijing* 100871, *China*;
3. *School of Government, Peking University, Beijing* 100871, *China*)

Abstract: Occupation is an important variable in social science research, but mistakes in the coding process of occupations in survey research are. unavoidable. Coding operations can take various forms. They are distinguished as centralized coding and decentralized coding based on their work sites, or as manual coding and computer-assisted coding based on their coding tools. Thus, combining these two dimensions there are four coding methods: manual centralized coding, manual decentralized coding, computer-assisted centralized coding, and computer-assisted decentralized coding. Computer-assisted coding has not been well developed in China, so most Chinese surveys employed the first two coding methods: interviewers carrying out coding during the interviewing process; or experienced coders performing the coding within the survey

[收稿日期] 2011 - 09 - 21 [本刊网址 · 在线杂志] http://www.journals.zju.edu.cn/soc
[在线优先出版日期] 2011 - 11 - 30
[基金项目] 北京大学人文社会科学青年教师科研启动基金项目

[作者简介] 1. 任莉颖，女，北京大学中国社会科学调查中心助理研究员，主要从事社会抽样调查研究方法、政治学研究方法等研究；2. 邱泽奇，男，北京大学社会学系教授，博士生导师，主要从事城乡发展等研究；3. 李力，女，北京大学中国社会科学调查中心博士后研究人员，主要从事卫生事业管理、卫生政策等研究；4. 严洁，女，北京大学政府管理学院副教授，主要从事社会科学定量研究方法等研究。

organization after data collection.

When choosing coding methods, survey practitioners usually have three factors in mind: cost, time efficiency, and coding quality. It is commonly believed that on-site coding by interviewers is cheaper and quicker than coders' centralized coding. However, there have been contradictory attitudes towards the quality of these two coding methods, and there have been very few empirical studies about that. Based on analysis of the occupational information collected by the Chinese Family Panel Studies (CFPS) in 2010, this study compares the results from these two existing coding methods in China and discusses the core factors that affect coding quality.

This study shows that coding results from these two methods differ greatly. Regarding the most detailed coding with 595 categories, only about one-third of the results from these two methods are identical. Even for simple coding with only eight categories, the proportion of identification still makes up only three-fourths.

Interviewers' text recording quality is an important factor that affects coding quality. In addition, interviewers' background and coding experiences are two main reasons for the discrepancies in the detailed coding results. It is also shown in this study that occupational categories have different levels of coding difficulty which also have an effect on coding results.

Administration of quality control over interviewers' on-site occupational coding is difficult in practice. Therefore, in rigorous social surveys, especially when detailed coding results are needed, it is strongly suggested to use the method of centralized coding. Moreover, since the quality of the interviewers' text recording is so important to the collection of accurate and complete occupational information, the following steps are recommended: establish a standard for interviewers' text recoding, strengthen the training of interviewers, and check their performance on a regular basis. It is also important to enhance quality control in the coding process, such as paying more attention to the design of the coding process as well as the supervision of the coders' work. These suggestions can be effectively put into practice in computer-assisted interviewing surveys.

Key words: social survey; occupational coding; data quality

职业是社会科学研究中广为应用的重要变量，对于研究中国改革进程中的社会变迁尤为关键。在国内许多重要研究项目中，如社会流动、社会分层、公民素质等，职业变量都是不可或缺的主要构成或影响因素。

近年来国内对社会经济信息的重视使社会调查业在中国得到蓬勃发展，个人职业状况几乎是每个调查中必定要采集的信息。这些调查或采用访员分散编码，或采用编码员集中编码，对职业信息进行分类处理，然而对于职业编码的质量却鲜有报告或经验性研究。

本文运用“中国家庭动态跟踪调查”(Chinese Family Panel Studies，CFPS)2010 年初访调查中收集的受访人职业信息，对访员手工分散编码和编码员手工集中编码两种方式的结果进行比较分析，期望能为职业编码方式的选择提供经验性依据，并寻求计算机辅助调查(Computer-Assisted Interviewing，CAI)模式下提高职业编码质量的途径。

一、职业问题的编码方式

在严谨的社会调查中,为了采集到详实的职业信息,一般都采用开放式提问,要求访员如实地记录受访人的回答,并在数据采集完毕后,再根据权威或普遍的职业分类标准创建编码列表,培训并组织编码员将文本信息转化为数值型的职业代码,提供给研究者使用。这种编码方式由于使用了专业的编码员,且进行统一指导和及时管理,可以在很大程度上保证编码质量,因而得到了广泛应用。然而这种编码方式的缺点是成本较高,且时效性差。与之对应的另一种编码方式是在数据采集过程中由访员在采访现场或在完成单个采访任务后按照调查机构事先提供的编码表,对问卷中开放问题的回答进行及时分类。这种方式虽然降低了编码成本,增强了时效性,然而由于缺少有效的质控措施,不得不以降低编码质量为代价。

随着计算机技术的广泛应用,人们总是期待可以使用计算机自动编码。一套设计完善的编码系统可以有效地降低成本、减少用时和提高编码信度。计算机自动编码方式可以分为两类:一类是计算机辅助编码(Computer-Assisted Coding, CAC),这种方式需和编码员手工编码结合使用。当编码员遇到编码困难时可以向CAC系统求助,CAC系统会根据编码员输入的信息给出一系列编码建议。另一类方式可以称为全自动编码。编码员将受访者的应答信息直接输入系统中,软件会自动对其分配编码,对于无法匹配编码的信息则由系统退出,转为人工编码。计算机自动编码方式已广泛应用在美国、加拿大、瑞典、澳大利亚等国家的人口普查职业信息编码中。

综上所述,社会调查中职业问题的编码方式有两种分类方法:一种是根据编码的地点及时间分为集中编码和分散编码。一般来说,集中编码都发生在调查结束后,由专业编码员来完成;而分散编码则发生在调查进行中,多由访员来完成。另一种是根据编码工具的使用分为手工编码和自动编码。手工编码中,编码员主要依据自己对职业的理解和对编码列表的掌握情况来选择职业编码;而自动编码则是借助职业编码软件进行全自动或辅助性编码。于是,在这两个维度上形成了四种基本的编码方式(见图1),即手工集中编码、手工分散编码、计算机辅助集中编码和计算机辅助分散编码。

	集中编码	分散编码
手工编码	手工集中编码	手工分散编码
自动编码	计算机辅助集中编码	计算机辅助分散编码

图1 职业问题编码方式分类

计算机辅助编码技术目前在中国尚未得到开发和应用。由于中国当前职业特征多样,职业分类复杂,更是增添了该项技术开发的难度,在短时间内很难应用到实践中来。因此,目前国内社会调查中职业问题编码主要采用的是编码员手工集中编码和访员手工分散编码。

二、职业问题的编码质量

对于职业问题的编码质量早在社会调查兴起之时就引起了西方研究者们的注意。

首先,一系列研究显示,职业变量存在着较为严重的编码误差。例如,瑞典1970年人口普查中关于职业问题的编码误差为13.5%,同一年美国人口普查对职业问题的编码误差也高达13.3%。针对这一问题,两国都采用了新的编码质控程序。在瑞典1975年的人口普查和美国1980年的人

口普查预调查中，都成功地将误差降低了约 8 个百分点[1]238。近几十年来，尽管有多种编码方式的混合使用，但关于职业信息编码质量的评估报告却较少，目前只发现在美国 Research Triangle Institute (RTI) 1991 年的一项研究中报告职业信息的编码误差为 21%[2]315。

研究者们对职业信息编码质量的影响因素及后果进行了探讨。基于传统的编码员手工集中编码，一些研究发现职业信息编码信度(reliability)会受到编码员自身的影响。为了评估瑞典 1970 年人口普查中开放性问题的编码质量，研究者们抽取了一部分样本，邀请 5 名经验丰富的编码专家对这些样本涉及的 8 个开放式问题(其中包括职业问题)进行编码，并计划将其编码结果作为标准，用来评估此次人口普查的编码质量。然而，在对 5 名编码专家的编码结果进行分析后发现，这些编码结果不仅因编码员不同存在着较大的差异，即使同一编码员的编码也有较大的变异性。如职业编码，编码员间的变异比例(between-coder variability)为 28.4%，而编码员自身的变异比例(within-coder variability)在 7.1%至 10.9%之间不等。而且职业问题的编码结果在测试的 11 个问题中变异性最高[3]。

访员实地(采访进程中或采访结束后)对职业问题进行编码的方式也比较常用。研究者们对两种方式的编码质量进行了比较研究，然而意见却很不一致。有研究发现，在职业问题编码上，专业编码员比实地采访员更容易达成一致标准，但总体来说两者的差别不大[4]。有研究证明，在降低编码员关联方差(correlated coder variance)上，使用访员编码要优于使用专业编码员，同时，两种方式的编码精确度没有明显区别[5]。另有研究发现，访员实地编码的变异性平均占应答总体方差的 3%，而编码员编码则只有 0.6%[6]，因此采用访员实地编码方式要谨慎。此外，也有证据显示访员实地编码会对采访行为产生负面影响[7-8]。

国内对于社会调查方法的研究正处于起步阶段。政府、商业和学术的调查机构对于职业信息的编码都设立了一定的规范，但尚未有任何一项研究对编码误差进行公开报告，也没有发现其他有关编码质量的实证研究。可以说，国内对职业编码质量的研究是一块尚未开垦的处女地。

三、数据来源及职业编码方案

(一) 数据来源

以往关于问卷调查中职业编码质量的研究多采用实验方法。这种设计虽然可以很好地控制目的，但却很难将结果推论到真实的调查实践中。同时，实验结果也常常会受到小样本量的限制，致使一些必要的统计分析无法正常进行。

本文所使用的数据来自一个正在进行中的全国概率样本跟踪调查——“中国家庭动态跟踪调查”(CFPS)。CFPS 是国内首次应用计算机辅助面访调查技术(Computer-Assisted Personal Interviewing，CAPI)的全国性综合跟踪调查数据平台，是由北京大学中国社会科学调查中心设计实施的一项旨在通过跟踪收集个体、家庭、社区三个层次的数据，反映中国社会、经济、人口、教育和健康的变迁，为学术研究和政府决策提供第一手实证数据的重大社会科学项目。作为探索性的尝试，该调查在 2010 年初访调查中对职业问题的编码采用了记录详细职业信息和访员分散编码的双重保障方式；在调查结束后，还对收集到的职业问题数据(文字描述)组织编码员进行手工集中编码。

CFPS 实地采访问卷中对于目前有工作的成年受访人共设计了 5 道有关职业的问题：

G303 您现在主要是在哪个机构工作？

G304 您现在工作单位的名称？

G305 请问,您现在主要工作的机构属于?

G306 您的职业是_____。

G307 您的职业属于哪一类?

其中,G303和G305为封闭式选择题,G304和G306为开放式问题,G307则是访员在CAPI系统下根据CFPS职业代码字典进行查询,对受访人的职业和行业进行现场编码。访员在记述G306的回答时,采访系统会给出提示:(1)如果受访者有多份工作,请询问占用时间最多的工作;(2)请详细记录受访者的主要工作,填写具体内容:工作部门+工作职责/工作内容+工作岗位/工种名称。

CFPS进行初访调查问卷的设计时,国家统计局尚未公布最新的职业分类体系(GB/T 6565-2009),而当时的职业分类体系(GB/T 6565-1999)已不能完全体现近十年来我国职业发展的状况。所以,CFPS初访调查的职业代码借鉴了"中国社会跟踪调查"[①]的职业分类标准,在GB/T 6565-1999的基础上进行了修订,包括8大类共计595个职业代码。

(二)访员实地编码

CFPS在2010年度的初访调查共使用了438名访员,这些访员大多数来自本次调查的目标区县,并且都参加了在北京大学举行的为期6天的集中培训,其中包括关于职业代码分类的专门培训。在职业代码分类培训中不仅详细讲解了职业代码的分类框架,还传授了在CAPI系统中快速查找职业代码的技巧,并对一些分类的难点进行了举例说明和现场演示。

作为CAPI环境下访员实地编码方式的首次尝试,这次调查对编码系统的设计采用了简单的查询法,在采访界面上呈现为树形结构(见图2)。访员编码时本着"先大类、后细类"的原则,首先确认受访人的职业属于哪一大类,然后可以逐级点击,最后确认四级代码为最终代码。此外,在访员培训中强调访员在编码感到含糊时要"多追问,问细节",以获取足够的编码信息。同时也建议访员使用键盘而非鼠标的方式进行操作,以减少错误点击,并加快操作速度。

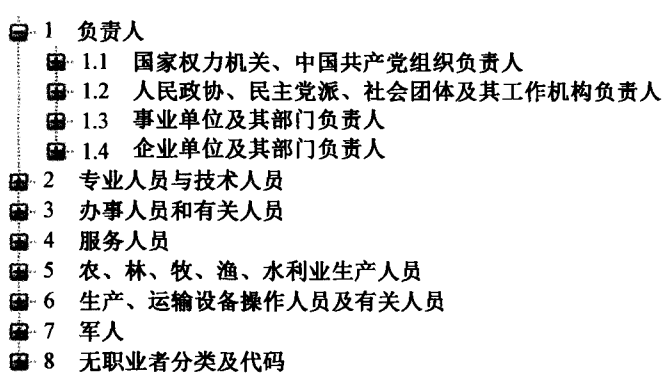

图2 CFPS 2010职业分类编码系统界面显示

(三)编码员集中编码

编码员集中编码工作开始于数据采集完毕后。编码员来自北京大学社会学系高年级本科生和研究生。在编码流程上采用了双向独立验证并判定(Two-way Independent Verification with

① 中国社会跟踪调查(Chinese General Social Survey, CGSS)是中国人民大学社会学系与香港科技大学调查研究中心合作,自2003年开始的全国性综合社会调查项目。

Adjudication)的质控方式[1]240，具体流程如图 3 所示：

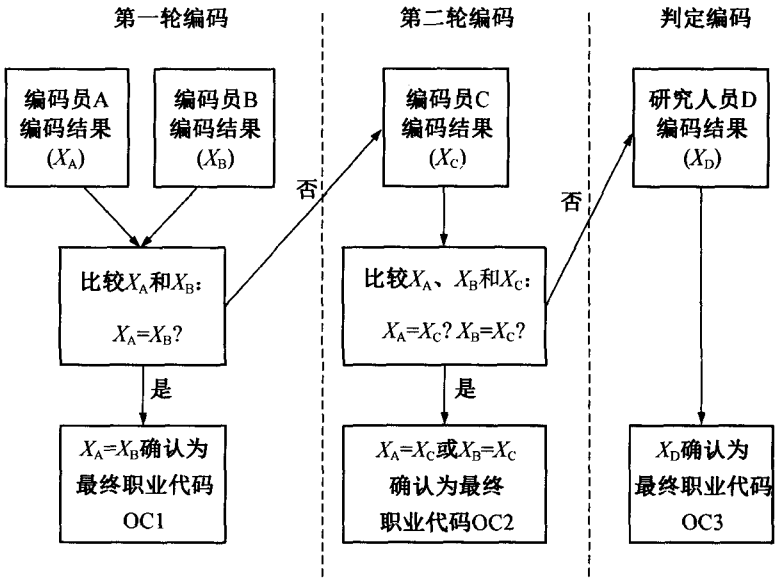

图 3　CFPS 2010 编码员集中编码质控流程图

在第一轮编码中，两个编码员(A 和 B)根据具体的职业描述、工作单位名称和工作机构属性信息，采用背对背的方式，对每一个受访者的职业信息进行编码。在这一轮编码中，如果两人编码结果一致，则确定为最终职业代码(OC1)；如果不一致，则将这些条目提取出来，进入下一轮编码。

第二个阶段选用经验较为丰富的编码员(C)，由其一人对这些不一致的条目进行独立编码，如果其编码结果与前一轮编码结果中的一个保持一致，则确定该编码为最终职业代码(OC2)；如果三人编码结果均不一致，则由专业的研究人员(D)根据编码员编码、访员实地编码以及相关的辅助信息进行判定，从而形成最终的职业代码(OC3)。因此，对于每一个职业问题条目，其最终代码的确定要经过 2—4 个编码员的判断，从而达到对编码质量进行控制的目的。

四、访员实地与编码员集中编码的结果比较

不出所料，编码员集中编码表现出了较好的编码信度。在第一轮编码结束后，两个编码员有 76.4%的结果达成一致，经过第三个编码员的确认编码，一致率提高到 91.4%，最后只有不到十分之一的职业编码需要由研究人员进行最终判定。

然而，编码员的编码结果与访员实地采访时的编码结果却有较大的分歧。我们将所有编码结果根据职业分类标准分为一到四级编码(一级码为最基本的分类，四级码为最精细的分类)，并且以编码员达成一致的一级编码①为分类标准来计算各级编码的一致比例，结果见表 1。

总体来看，两种编码方式在四级码上只有大约三分之一的结果是一致的，随着编码类别的减少，一致的比例也逐渐加大。如在二级和三级码上，有三分之二左右的结果相同，而在职业大类的一级码上，一致率可以达到四分之三。

分职业类别来看，从事农、林、牧、渔、水利业的生产人员在各类有职业信息记录的受访者中

① 职业编码所采用的分类标准共有八大类，但军人和无职业者不适用于该问题，因此没有列入在此项分析中。

人数最多,约占一半。同时,这一类别内职业详细情况的编码一致率也最高。尤其在一级的大类编码上两种方式的编码结果一致率高达99.35%,在最为精细的四级编码上也有一半以上的相同比例。

与之相比,人数比例占第二位的生产、运输设备操作人员及有关人员对其具体职业的编码结果则不甚乐观。在四级码上只有6.05%的编码结果相同,即使在一级的大类码上一致率也只有38.02%。对于其他职业大类内的四级精细具体编码,其结果的一致率也大多低于50%。

虽然有经验的调查人员对访员实地编码方式可以列举出诸多缺点,但却很少能够如此直观地发现访员实地编码结果与编码员集中编码的结果存在如此大的差异。那么到底是哪些因素影响了两种方式的结果不同?我们是否可以认定编码员集中编码的结果要优于访员分散编码呢?

表1 编码员集中编码与访员实地编码结果比较

职业分类	一级编码一致率(%)	二级编码一致率(%)	三级编码一致率(%)	四级编码一致率(%)	N
1. 负责人	43.05	41.79	29.38	16.19	1 266
2. 专业人员与技术人员	58.44	55.67	41.75	26.69	1 049
3. 办事人员和有关人员	56.93	45.62	41.73	20.32	822
4. 服务人员	76.44	52.53	61.28	31.14	1 108
5. 农、林、牧、渔、水利业生产人员	99.35	92.83	85.78	55.05	7 696
6. 生产、运输设备操作及有关人员	38.02	24.99	24.21	6.05	2 809
总体	75.84	67.23	61.79	36.64	14 750

五、影响职业问题编码质量的因素

影响职业问题编码质量的因素可分为三大类。

第一类影响因素来自于访问过程,如访员所记录下的职业信息量、访员编码时的态度以及访员编码的经验等。无论是编码员集中编码还是访员实地编码,其编码质量都会受到访员记录下来的职业信息的影响。一方面,这反映了访员了解受访人职业状况的程度;另一方面,它也是编码员集中编码的重要信息来源和凭据。因此,职业信息越丰富,两者的编码结果一致的可能性越高。本项研究对访员记录职业信息的字数进行了统计,并以此来测量职业信息的丰富程度。数据显示,访员记录职业信息的平均字数为5.76,最大值为24,最小值为1。

访员编码时的态度可以用访员编码的时间来测量,我们假定工作认真的访员所用的编码时间会相对较长。这项指标在传统的纸笔方式的问卷采访中很难获得数据,然而由于CFPS采用计算机辅助采访系统,系统可以自动记录下每道问题的采访用时。本次调查中,访员对职业问题进行实地编码的平均用时为26秒。我们认为访员在编码时所用的时间越长,其编码结果和编码员集中编码的结果越有可能取得一致。

分析发现,访员对职业信息的编码数目从1个到271个不等,但平均下来每个访员大约会对36个受访人的职业信息进行编码。我们假定访员编码的经验会随着编码次数的增加而上升,为此我们对每个访员所完成的需对职业进行编码的问卷依据完成时间进行排序并赋予序号,序号越大意味着该访员的编码经验越丰富,从而与编码员集中编码结果一致的可能性越高。

第二类影响因素与访员自身的背景有关,我们主要选取了访员的性别、年龄和受教育程度三个

变量。本次调查中，有 28.2%的访员是女性。访员的平均年龄是 28.2 岁，最小为 18 岁，最大为 51 岁。其中，拥有大专学历的占 36%，本科及以上学历的占 49%，而高中及以下学历的仅占 15%。经验上认为年龄大的访员、男性访员以及受教育程度较高的访员在职业方面的知识较为丰富，编码质量也会相对较高。

第三类影响因素与职业本身的编码难度有关。表 1 展示了不同职业大类下编码结果的一致率有较大差异，这意味着这些职业大类下的具体编码难度不尽相同。例如对于农、林、牧、渔、水利业生产人员的编码一致率最高，这可能是因为这类职业编码相对容易。因此，在分析中以第五职业大类（农、林、牧、渔、水利业生产人员）为参照组，对不同的职业分类进行控制。

为了考察这些影响因素在不同级别的职业编码上的表现，我们控制了职业类别的影响，分别对每一级别下两种方式的编码是否一致的结果进行了 Logistic 回归分析，结果见表 2：

表 2　编码员集中编码与访员实地编码结果一致性的 Logistic 回归分析

影响因素		一级编码是否一致	二级编码是否一致	三级编码是否一致	四级编码是否一致
访员记录职业信息字数		1.055***	1.101***	1.080***	1.014*
		(0.009)	(0.008)	(0.007)	(0.007)
访员实地编码用时		0.999	1.000	1.001*	0.999
		(0.001)	(0.001)	(0.001)	(0.001)
访员职业编码经验		0.999	1.000	1.004***	1.005***
		(0.001)	(0.001)	(0.001)	(0.001)
访员性别（0＝女）		0.997	1.130**	1.220***	1.272***
		(0.050)	(0.052)	(0.053)	(0.054)
访员年龄		1.004	1.016***	0.995	1.034***
		(0.004)	(0.004)	(0.003)	(0.003)
访员受教育程度	大学本科及以上	1.080	1.028	0.808***	1.542***
		(0.076)	(0.067)	(0.050)	(0.095)
	大专	1.006	0.807**	0.606***	1.035
		(0.070)	(0.052)	(0.038)	(0.065)
样本数（N）		14 750	14 750	14 750	14 750

注：（1）职业大类为控制变量，访员受教育程度中以高中及以下为对照组；（2）表中所报告的数值为发生比（odds ratio），括号中数值为标准误（standard error）；（3）*** 表示 $p<0.001$，** 表示 $p<0.01$，* 表示 $p<0.05$。

分析结果显示，在第一类影响因素中访员记录的职业信息字数至关重要。访员记录下的字数越多，两种方式的编码质量越好，结果一致的可能性也就越大。如在控制其他变量影响的情况下，访员多记录一个字，在二级编码上结果一致的可能性会提高 10 个百分点，一级编码上会提高5.5个百分点，三级编码上会提高 8 个百分点。然而这个变量在四级编码上则效果不太明显，这意味着在精细的职业编码中，除职业信息的详实程度外，其他因素也起着重要作用。

访员职业编码经验在进行初级的职业编码时虽然没有什么作用，但在高级别的职业编码上显示了显著的影响，证明访员编码经验越丰富，在进行精细的职业编码时与编码员集中编码的结果越

接近。但由于访员接触到的职业编码条目较少,编码经验在访员身上并没有显示出明显的优势,然而在本次调查集中编码时,平均每个编码员会对大约3 000条职业信息进行编码,编码员丰富的编码经验意味着更好的编码质量。

相比之下,访员实地编码用时在控制了其他因素的作用后没有显示出明显的影响[①]。

对访员自身背景因素的分析显示,性别和年龄两个变量的作用基本与预期的相同,尤其是在进行精细的高级别的职业编码上,男性访员确实比女性访员表现出更好的编码质量,在精细的四级编码上,控制其他变量的影响,男性访员的编码结果与编码员集中编码结果一致的可能性要比女性访员高出27个百分点。同时,年龄较大的访员也表现出一定的优势,在四级编码上,访员每年长一岁,编码结果一致的可能性会增长3.4个百分点。

然而,访员受教育程度的作用并非像我们预料得那样简单。在较粗略的低级别的职业编码上,教育差别并没有明显地反映在编码数据质量的差异上,甚至在三级编码上,受教育程度高的访员反而更容易出现与编码员集中编码不一致的情形,但在进行精细的四级编码上,大学本科及以上学历的访员显示出了明显的优势,与集中编码结果取得一致的可能性远远大于高中及以下学历的访员。

职业类别在分析中既是影响因素,也是控制变量。与农、林、牧、渔、水利业生产人员的职业相比,其他类别的职业确实显示出较高的编码难度,从而导致两种方式编码结果不一致的可能性较高(结果未报告)。同时,在控制了职业类别后,可以更清楚地看到访员的编码信息、编码经验以及自身素质对编码质量的影响。

根据以上分析可以概括,信息、经验和素质是影响职业问题编码质量的重要因素,三方面因素对于精细或粗略编码上的作用不尽相同。对于粗略的职业编码,访员记录的职业信息最为关键,信息越多,编码结果一致的可能性越高;然而在进行精细的职业编码时,访员自身因素也起了比较重要的作用,编码经验和访员素质都会直接影响编码结果的一致性。

六、结论及提高职业问题编码质量的设想

采用访员分散编码的方式对社会调查研究人员具有很强的吸引力。一方面,这种方式成本低,时效强;另一方面,也可以避免使用专业编码员时所遇到的信息不足及编码员关联方差的问题。

然而,本研究显示,该种编码方式的数据质量令人担忧。和编码员集中编码的结果相比,在精细的四级职业分类编码上,只有大约三分之一的结果相同,即使在最为粗略的一级职业分类编码上,结果相同的比例也只有四分之三。

在精细编码上出现的差异,主要受到访员记录的职业信息量、访员编码经验及访员自身素质的影响;在粗略编码上的不同,则主要与访员记录的职业信息量有关。分析也显示,不同的职业类别也显示出不同的编码难度,造成两种方式的编码结果不一致。

由于没有判断所有职业编码对错的绝对标准,我们不能直接判断哪种编码方式的数据质量更好。然而,研究发现访员的素质和编码经验对于精细的四级编码数据质量非常重要。但在访问实施过程中,对访员因素的控制难度较大,提高访员实地编码质量不易实现。相比之下,采用集中编码的方式可以对编码员进行筛选,编码员不仅会受到编码工作的专业培训,并且可以接触大量的编码条目,从而积累丰富的编码经验,加上编码过程的集中管理和有效的质控手段,我们可以推断编码员集中编码的数据质量要更可靠。因此,在严谨的社会调查中,特别是在精细的职业编码上,建

① 我们怀疑访员实地编码用时过长,可能会在记录职业描述信息时偷工减料,从而间接影响到两种方式编码结果的一致性,然而附加分析发现这种猜测并没有得到数据支持,因此没有反映在模型建构中。

议采用编码员集中编码来获取更好的职业编码数据。

研究结果还肯定了访员记录的职业信息对于职业编码的重要性，因此应该加强对访员记录职业信息的行为规范，尽量采集到准确编码所需的重要信息。同时，在编码员编码过程中也要加强质量控制，从而在信息输入和处理过程两方面来保证编码数据的质量。

在计算机辅助调查中，这些建议可以得到有效实现。具体地说，可以考虑从三个角度来改进职业编码：（1）调查前对访员加强职业编码的培训，使访员明白编码所需的重要信息，确立访员对职业信息的记述规范。（2）在调查执行的同时组织编码员进行集中编码。计算机辅助调查的优势之一就是调查数据可以在采访当天传送到总部，这样可以及时将职业描述信息提取出来，组织编码员开始编码。实时的集中编码有两个目的：一是可以及时发现职业信息记述含混及难以归类的条目，并请访员协助补充信息；二是作为数据质量监控的手段，可以提醒或干预访员遵守职业信息的记述规范。（3）除了采用双向独立验证并判定的编码流程外，还要加强编码员集中编码的质量监控，对编码效率和质量进行定期评估。较为简单的做法就是借用计算机辅助调查系统，将职业分类说明作为帮助文件，以方便编码员查询，然后利用下拉菜单选择代码或直接输入代码的方式进行编码。计算机辅助调查系统可以记录下编码员每个条目的编码用时，同时也可以及时导出数据进行编码结果的比较和判定，从而为定期的质量和效率评估提供数据基础。这样，不仅可以保证职业问题编码的数据质量，还可以加强编码员集中编码的时效性，并且提升编码效率，降低编码成本。

[参　考　文　献]

[1] P. Biemer & L. Lyberg, *Introduction to Survey Quality*, New York: Willey & Sons, Inc., 2003.

[2] P. Biemer & R. Caspar, "Continuous Quality Improvement for Survey Operations: Some General Principles and Applications," *Journal of Official Statistics*, Vol. 10, No. 3(1994), pp. 307 - 326.

[3] L. Lyberg, *Control of the Coding Operation in Statistical Investigations—Some Contributions*, Stockholm: Statistics Sweden, 1981.

[4] P. Campanelli, K. Thomson & N. Moon et al., "The Quality of Occupational Coding in the United Kingdom," in L. Lyberg, P. Biemer & M. Collins et al. (eds.), *Survey Measurement and Process Quality*, New York: Wiley-Interscience, 1997, pp. 437 - 453.

[5] J. Martin, D. Bushnell & P. Campanelli et al., "A Comparison of Interviewer and Office Coding of Occupations," http://www.amstat.org/sections/srms/Proceedings/papers/1995_195.pdf, 2011 - 09 - 20.

[6] M. Collins & G. Courtenay, "A Comparative Study of Field and Office Coding," *Journal of Official Statistics*, Vol. 1, No. 2(1985), pp. 221 - 227.

[7] D. Maynard, H. Houtkoop-Steenstra & N. Schaeffer et al., *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, New York: Wiley, 2002.

[8] F. Fowler & T. Mangione, *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*, Beverly Hills: Sage Publications, 1990.