제2장

총오차 축소를 위한 Paradata 수집방안

김정섭·임경은

제1절 서 론

1. 개요

조사에서 발생 가능한 모든 오차를 총오차(total survey error)라고 하며, 크게 표집오차(sampling error)와 비표집오차(non-sampling error)로 나눌 수 있다. 표집오차는 전수가 아닌 표본조사를 수행함으로써 발생하게 되는 오차로 표본수가 증가하면 작아지는 특성을 가지고 있으며, 발생되는 오차의 정도를 이론적으로 측정할 수 있다. 그러나 비표집오차는 측정이 쉽지 않고 이론적인 설명이나 추정 또한 어려운 실정이다. 비표집오차에는 포함오차(coverage error), 무응답오차(non-response error), 측정오차(measurement error), 처리오차(processing error) 등이 포함되며, 표집오차에 비해 총오차에서 차지하는 비중은 훨씬 크다.

이에 유럽의 여러 나라와 미국, 캐나다 등에서는 단순히 조사에서의 표집오차뿐 아니라 비표집오차도 조사에서 관리해야 하는 오차로 규정하고, “총오차”라는 개념을 사용하고 있다. 이를 위하여 비표집오차에 대한 연구가 꾸준히 진행되고 있으며, 그 중 측정오차 및 무응답오차와 관련하여 조사방법 전반에 대한 자료를 수집하기 시작하였다. 이 때 수집 되는 자료를 “Paradata”라고 한다. Paradata는 실제 조사 과정에서 언제나 존재하게 되며, “조사 관리에 관한 자료”라고 할 수 있다. 이와 같은 paradata의 수집은 궁극적으로 정합성 높은 조사 결과의 생산(총오차 축소)과 조사 비용 축소(조사의 효율화) 그리고 조사 관리의 기초 자료를 마련하기 위한 방안으로 이용될 수 있다. Paradata의 어원은 그리스어의 “Para”로 가까운, 옆, 뒤, 병렬 등의 의미를 가지고 있다. Eurostat에서는 paradata와 관련하여 “Process data”라는 용어를 사용하고 있으며, 자료수집 과정, 공표, 유지, 보존 등 조사의 품질을 향상시키기 위한 전방위적 관리 자료를 의미한다. 따라서

paradata는 process data의 부분집합이라고 할 수 있다. 이와 같이 paradata에 대한 용어는 해당 용어를 쓰는 지역과 paradata가 포함하는 자료의 범위에 따라 다르게 이용되고 있으며, 연구자에 따라 서로 다른 용어로 표현되기도 한다. 일부 연구자들은 paradata와 관련된 용어를 통일해야 한다는 주장을 펴기도 하였으나, 스웨덴 통계청의 Lyberg(2009)는 표준화된 전문용어를 개발하는 일은 시간을 낭비하는 일이며, 자료가 포함하는 범위나 쓰는 목적에 따라 다르게 부르는 것이 좋다는 의견을 제시하였다. 본 연구에서는 조사관리에 관한 자료를 Eurostat에서 지정한 “Process data”가 아닌 미국이나 캐나다 통계청에서 사용하는 “Paradata”로 명명하도록 한다.

조사비용과 총오차 축소를 통한 자료의 품질 향상은 매우 어려운 일이다. 그러나 paradata를 이용하여 실시간으로 자료수집 과정에 대한 의사결정을 내리게 된다면 조사비용을 줄임과 동시에 무응답 오차를 줄이는 것이 가능할 것이다. 즉, 응답을 잘 할 수 있도록 설계된 조사는 paradata를 이용한 정보로부터 가능하다고 할 수 있다.

Couper(1998)는 자료를 수집하는 과정에서의 자료; 예컨대 조사대상자를 만나기 전 몇 번의 전화를 걸었는가, 면접 시간, 비용 관련 자료, 입력 과정에 대한 자료, 조사원의 특성, 그리고 응답률 등을 “Paradata”라고 정의하고, 조사방법론 분야에 이 개념을 처음 소개하였으며, 이후 paradata를 활용하여 무응답오차와 측정오차 등을 축소하는 방안에 대하여 연구를 계속하고 있다. 또한 조사 품질 평가를 위한 paradata 이용의 잠재력과 혜택에 대하여 재고하고 여기서 생성되는 정보를 이용하여 조사를 어떻게 진행할 것인가 하는 문제에 대한 방안도 제안하였다.

Scheuren(2001)은 paradata의 수집이 이용자 지향적이기보다 조사를 수행하는 조사자 지향적이라고 주장하였다. 즉, 인터넷을 활용한 표본조사를 진행하는 과정에서 수집된 paradata는 조사 수행 과정에 대한 개선에는 유용하게 이용될 수 있으나 자료를 이용하는 이용자들에게는 그리 유용한 정보를 제공하지는 못하는 것이다. 즉, 자료수집 과정에 대한 세부적인 정보는 자료를 이용하는 이용자들에게 제공될 수는 있으나, 이용자들이 paradata를 이용하기 위해서는 먼저 paradata가 자료 분석 결과에 어떤 영향을 미치는지 알 필요가 있다. 이용자와 제공자 간의 쌍방향 소통을 개선하기 위해서는 조사 품질에 대한 평가에 관한 새로운 생각이 필요한 것이다.

본 연구에서는 미국 및 캐나다의 paradata 수집 현황을 파악하고 paradata 활용에 대한 정보를 수집하여 우리 통계청에서 시급히 도입해야 할 paradata를 선정하는데 목적을 둔다.

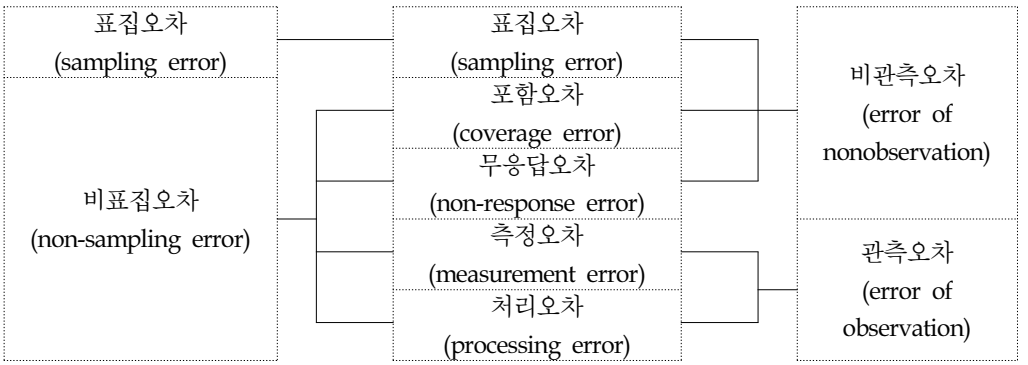
2. 용어

가. Paradata

조사 자료의 수집 과정에 관한 자료를 paradata라고 하며, 모든 조사 수행과정에 언제나 존재한다. 초기 paradata는 조사 관리를 위한 정보를 얻기 위한 목적으로 수집되었으며, “조사 관리에 관한 자료”라고 불리우기도 한다. 현재는 paradata 분석을 통하여 무응답오차나 측정 오차 등, 비표집오차를 줄이는 목적으로 이용되며, 체계적이고 이용 가능한 형태의 paradata를 수집하는 것이 조사 주체와 관련 분야 연구자들의 목표이다.

나. 총오차

총오차는 모든 오차를 포함하는 개념으로 [그림 2-1]과 같이 분류된다.



[그림 2-1] 총오차(Total survey error)

일반적으로 표본조사를 통해 생산되는 통계에서는 표집오차만을 측정하고 있으며, 확률화이론(randomization theory)의 불편성(unbiasedness)에 의해 표준오차(standard error)는 표집오차와 같다. 즉, 표준오차에는 비표집오차가 포함되지 않으며, 비표집오차는 편의(bias)에 영향을 미치게 된다. 그러나 총오차를 연구하는 연구자들은 비표집오차의 측정이 쉽지는 않으나 별도의 연구를 통해 측정이 가능하다는 견해를 가지고 있다.

1) 표집오차(sampling error)

표집오차는 전수가 아닌 표본을 조사하기 때문에 일어나는 오차로 표본 조사의 특성 중 하나이다. 이 때 표본 수가 클수록 표집오차는 작아지게 되며, 이론적으로 그 크기를 측정할 수 있다.

2) 포함오차(coverage error)

모집단 간의 불일치에서 발생하는 오차로 일반화하고자 하는 모집단(population of interest)과 실제로 정의된 목표 모집단(target population), 그리고 실제 조사에서 이용되는 모집단(frame population) 간의 차이에서 발생된다.

3) 무응답오차(non-response error)

모든 조사대상자가 다 응답하지 않기 때문에 일어나는 오차로 응답자의 조사 거절, 연락(접촉) 불가, 질병, 신체 혹은 발달 장애 등에 기인한다. 또한 외국인 노동자의 유입과 국제 결혼의 폭발적인 증가는 언어장벽에 의한 무응답오차 증가의 원인이 되고 있다.

4) 측정오차(measurement error)

조사를 진행하는 과정에서 측정의 결함으로 발생하게 되는 오차로 조사에 이용하는 도구, 조사표 구성, 설문 문항, 조사원의 조사 방식, 응답자의 이해도 등에 의해 발생된다.

5) 과정오차(processing error)

수집된 자료를 처리하는 과정에서 일어나는 오차로 자료 입력 오류 등에 의해 발생된다.

총오차는 조사를 통한 자료의 품질과 깊은 연관성을 가지고 있으며, 이를 측정하여 최소화하는 방안을 마련하는 일은 조사를 수행하는 주체들의 중요한 이슈들 가운데 하나이다. 지금까지는 총오차를 줄이기 위한 노력들이 표본수를 조절하는 등의 표집오차 축소에 집중되어 있었으며, 무응답오차나 측정오차 등 비표집오차를 측정하여 줄이고자 하는 노력은 체계화되어 있지 못했다. Paradata는 이러한 비표집오차를 체계적인 방법을 통하여 객관적으로 측정하고 분석함으로써, 궁극적으로 총오차를 최소화하기 위한 방안 마련의 기본 자료로 이용될 수 있을 것으로 기대된다.

제2절 해외사례연구

조사 자료를 수집하는 과정에 대한 정보인 paradata의 이용은 조사와 관련된 여러 방면에서 깊이 있게 연구되고 있다. 1980년대 이후 현장에서 입력되는 추적파일(trace file)이나 keys pressed 그리고 컴퓨터를 이용한 면접조사의 진행시간, 조사 과정에서의 의미 있는 정보들은 다양한 방법으로 탐색되고 있으며, paradata의 적용 영역은 날로 확대되고 있다. 최근에는 우리나라뿐 아니라 세계 여러 나라에서의 표본조사가 사회·경제·인구·문

화 등 다양한 방면에서 폭발적으로 증가하고 있으며, 이에 따라 표본조사의 품질을 유지하기 위한 paradata에 대한 관심도 점차 높아지고 있다. 이에 최근 미국 통계국에서는 표본조사 설계와 자료수집 과정의 중요한 잠재적 요소들을 새로운 paradata로 수집하는데 매우 강력한 방법으로 여겨지고 있는 CARI(Computer Assisted Recorded Interviewing) 시스템을 미국 통계국 산하의 모든 표본조사에 적용하기 위하여 시범조사와 관련 연구를 꾸준히 진행하고 있다. 본 연구에서는 최근 미국 통계국에서 진행하고 있는 paradata 연구와 적용 그리고 CARI 시스템에 대한 전반적인 사항들을 살펴보고자 한다. 또한 통계 선진국 중 하나인 캐나다의 paradata 관련 연구도 소개하도록 한다.

1. 미국 사례

가. 미국 통계청의 PANDA(Performance and Data Analysis) 시스템

미국은 넓은 지역에 많은 인종이 다양한 문화를 가지고 살고 있어 표본조사에 어려움을 겪고 있다. 따라서 전국을 대상으로 하는 조사를 수행하기 위해서는 많은 표본과 긴 조사 기간이 필요하며, 조사 과정에서 발생하는 비표본오차도 인종 및 문화의 다양성과 더불어 매우 많고 다양하다. 이에 미국 통계국에서는 신뢰성 있고 비표본오차가 적은 국가 통계를 생산하기 위하여 여러 기관을 활용하여 관련 통계를 대신 생산하도록 하고 있으며, 체계화된 조사 과정과 조사 현실을 감안한 조사표 그리고 조사원들에 대한 교육 등을 바탕으로 고품질의 통계 자료를 생산하기 위한 노력을 아끼지 않고 있다. 또한 국가 통계 관련 연구를 위하여 통계청 내에 방법론을 연구하는 두 개의 부서를 운영하고 있으며, 표본조사방법 및 추정 등에 대한 연구를 Research Triangle Institute(RTI)에서 심도 있게 진행하고 있다. 예를 들어 미국의 보건 관련 통계는 미국국립보건센터의 국립보건면접조사로 생산되고 있으며, RTI에서 개발한 CARI를 이용한 시범 조사를 통하여 미국 통계청에서 CAPI 방식으로 조사되는 모든 표본조사에 적용할 계획을 가지고 있다. 또한 RTI에서는 가중값 및 분산 추정 등에 대한 연구를 통하여 국가 통계의 정확도 확보를 위해 노력하고 있다.

미국 통계국의 Ari Teichman(2009)은 국가 통계를 생산하는 기관들의 조사 수행과 자료 분석(Performance and Data Analysis; 이하 PANDA) 시스템 그리고 해당 시스템의 표본조사의 품질 개선 효과에 대한 연구를 진행하였다.

PANDA 시스템은 2007년 미국주택조사(American Housing Survey)에서 시행된 바 있으며, 시스템의 주요 목적은 조사원들에게 조사의 주요 개념을 명확하게 전달하고, 허위 및 부실조사에 대하여 엄중히 경고하기 위한 것이다. PANDA 시스템은 면접 대상 및 지역, 하루 중 면접이 진행된 시간, 면접 결과, 이상치, 그리고 그 외 여러 사항들과 관련된

paradata를 기반으로 상당한 수준 이상의 보고서를 제공하고 있다. 조사원들에 의해 위조가 가능한 사항들 중에는 빈 가구 비율이 너무 높은 경우, 크기가 작은 가구의 수(e.g. 단독 가구)가 많은 경우, 일반적이지 않은 시간(자정~오전 8시)에 이루어진 면접, 면접 시간이 지나치게 짧은 경우 등이 포함된다. 우리나라에서도 이와 같은 보고서가 지방청 자체적으로 발간된 바 있다. 경인지방통계청에서는 2009년 지역별 고용조사를 대상으로 조사원들의 부실조사 건에 대하여 사후 분석을 실시하였으며, 그 결과 지정된 표본조사 구역을 벗어나거나 지정된 구역 내 임의의 가구(가구원)를 선정하여 조사한 경우, 실제 한 가구를 여러 단독가구로 분리하여 조사한 경우 등의 부실 조사 사례가 발견되었으며, 조사 내용을 허위 기재하거나 내용에 대한 착오를 일으킨 경우 등도 나타났다. 이와 같은 조사는 재조사와 사후조사를 통하여 해당 조사원이 조사한 조사표를 전체 결과 집계에서 제외하거나 수정하는 과정을 거친 바 있다. PANDA 시스템의 관리보고서에는 행정구역/지역별 총계, 누적/주단위 보고서, 그리고 조사원 당 평균 면접 개수 등이 포함되며, 면접조사에 관한 사항으로는 조사대상자의 연봉을 묻는 질문 등 민감한 질문에서 발생한 무응답 비율, 20분 미만에 완성된 면접 수, 자정부터 오전 8시 이전에 완성된 면접 수, 항목무응답 수 등이 포함되어 있다.

또한 PANDA 시스템은 현장 관리자가 면접의 세부정보를 검사할 추적파일을 다운받을 수 있도록 되어 있다. 예를들어 현장 관리자는 자정부터 오전 8시 사이에 면접이 이루어진 경우에 잠재적인 위조가 일어났음을 알 수 있으며, 추적파일로부터 이 부분과 관련된 정보를 제공받을 수 있는 것이다. 또한 현장 관리자는 부실조사와 관련된 잠재적인 문제를 해결하기 위하여 조사원 재교육이나 재면접의 필요 등을 식별하기 위한 정보를 해당 시스템을 통하여 검색할 수 있다.

나. 국립건강통계국 건강면접조사의 Paradata File

미국 국립질병통제예방센터(Centers for Disease Control and Prevention; CDC) 산하 국립건강통계국(National Center for Health Statistics; NCHS)에서는 매년 전국적으로 건강면접조사(National Health Interview Survey; NHIS)를 수행하고 있다. 건강면접조사는 정부에서 이용하는 랩탑 컴퓨터를 사용한 CAPI로 이루어지며, 면접조사 후 필요한 경우에 한해 전화조사로 후속 조치를 취하고 있다. 조사는 35,000 가구 이상, 약 87,500명 정도를 대상으로 이루어지며, 표본 가구로 선정되면 해당 가구 내의 아동과 성인을 임의로 한 명씩 선정하여 조사를 수행하게 된다. 이 때 가구와 가족에 대한 질문에는 가족 내 성인이 답하게 되며, 가구 내 아동이 있는 경우 표본으로 선정된 아동과 성인이 건강과 관련된 세부적인 항목에 대해 조사에 참여하게 된다. 1957년 처음 실시된 건강면접조사는 연간으로 조사가 이루어지며, 원자료(micro data) 또한 매년 홈페이지를 통해 제공하고 있다.

건강면접조사를 수행하고 있는 국립건강통계국에서는 면접조사의 효율성과 조사 과정에서 발생 가능한 측정 오차, 그리고 무응답 오차를 축소하기 위한 일환으로 paradata를 수집하여 코드화하는 작업을 2006년 처음 실시하였다. Paradata 파일은 표본조사 과정에 대한 자료로 이루어져 있어 건강 관련 자료는 포함되어 있지 않다. 그러나 paradata 자료를 이용한 분석 시에는 paradata 파일 내에 포함된 자료만을 이용한 분석과 함께 건강면접조사를 통해 얻어진 자료와 연계하여 분석하는 방법이 모두 가능하다. 또한 paradata는 조사 과정에 대한 자료이므로 조사가 끝난 이 후 약 6개월 간의 코드화 작업이 필요했으며, 건강면접조사의 자료 파일과 가구 코드가 일치하도록 입력되었다.

1) 건강면접조사의 Paradata

가) 접촉이력계기(The Contact History Instrument; CHI)

현장 조사원들에 의해 만들어지는 각각의 접촉 시도에 대한 정보를 수집하는 것으로, 표본가구에 조사원이 접촉한 횟수 등을 나타낸다. 조사원은 접촉의 성공 여부와 관계없이 CHI 안에서 정보를 기록해야 하며, 과거에는 도스 기반(CASES)으로 기록하던 것을 지금은 윈도우 기반(Blaise)으로 하여 자료를 수집하고 있다. 해당 자료 수집 방법은 미국 센서스국(Census Bureau)에서 이용하는 방법을 국립건강통계국에 맞게 조정하여 이용하고 있다.

CHI에 수집되는 정보로는 접촉 횟수나 접촉 시도를 직접 방문하여 했는지 아니면 전화로 했는지 여부와 접촉 시도 결과 등이 있으며, 만약 접촉에 성공했다면 조사원은 접촉 시도가 어떤 과정을 통하여 이루어졌는지를 기록하게 된다. 예를들어 처음 응답자가 조사를 꺼리는 것처럼 보였다면 그 이유가 무엇인지(바쁨, 무관심 등), 면접 성공을 위해 사용한 방법(전략)은 무엇인지 등을 기록하게 되며, 만약 접촉에 실패했다면 접촉에 실패하게 된 이유와 앞으로의 면접 전략 등을 기술하게 된다. 즉, 방문 당시 빈집이었는지, 초인종 또는 노크를 얼마나 많이 시도했는지 등이 포함되며, 재방문 약속을 잡기 위해 이용한 방법(메모, 약속 카드 등) 등 조사 성공을 위해 취한 조치가 무엇인지 등이 포함된다.

이와 같은 CHI 자료는 각 조사 단위(가구)에 대한 파일과 접촉(방문)에 대한 파일로 각각 구분된다. 조사 단위(가구)에 대한 파일에는 해당 가구의 접촉 시도 횟수, 접촉 성공 여부, 조사 거절에 대한 이유 등의 확인, 1인 가구 등에 대한 면접 전략 등이 포함되어 있으며, 접촉(방문)에 대한 파일에는 접촉을 시도한 날짜와 시간, 특정 방문 결과에 대한 설명 등이 포함된다. 2006년 국립건강통계국에서 처음 구성된 paradata 파일은 조사 단위(가구)에 대한 파일이었으며, 이 후 2007년과 2008년에는 접촉(방문)에 대한 파일도 추가되었다.

나) 표본조사 전 · 후의 추가 정보

조사원이 조사를 수행하면서 거치게 되는 일련의 절차와 상황들 외에 표본조사 전·후의 확인 사항들도 paradata에 포함될 수 있다. 표본조사를 수행하기 전에는 표본으로 선정된 가구가 실제 면접조사를 수행하기에 적합한 가구인지 확인하는 과정이 필요하며, 만약 조사 대상에 포함되었으나 조사에 대한 거절이나 일시적 부재로 인하여 조사 참여가 불가능한 경우에는 해당 가구를 표본조사 대상에서 제외하게 된다. 이 때 조사 대상 가구의 조사 가능 여부 확인 결과와 거절이나 조사 불능(일시적 부재)인 가구를 조사 대상에서 완전히 삭제할 것인지 아니면 단위무응답으로 인정할 것인지 등에 대한 결정 등도 paradata로 이용된다.

면접조사 후에는 조사원들을 대상으로 하는 일련의 질문을 통하여 면접 방식(방문, 전화 등), 응답자의 협조성, 항목무응답이 발생하는 원인 등을 추가적으로 조사하여, 면접 조사가 완벽하게 이루어지지 않은 이유에 대한 정보를 얻게 된다. 이렇게 얻은 추가 정보 또한 paradata로 이용될 수 있다.

다) 기타

접촉이력에 대한 측정이나 표본조사 전·후의 추가적인 정보 이외에 전체 면접에 소요된 시간이나 분야(가족 및 가구 사항, 아동, 성인 등)별 소요 시간, 감사추적기록(audit trail)[1], 키스트로크파일(keystroke file)[2], 추적파일(trace file)[3], 각 문항별 소요 시간, 면접을 수행한 날짜와 시간(조사를 끝낸 시간 - 시작한 시간) 등이 있으며, 전반적으로 큰 제약이 없는 상태에서 자료가 수집되고 있다.

2) Paradata 파일의 일반적 정보

2006년의 조사 단위(가구) 파일에는 125개 변수와 44,264개의 개체가 들어있으며, 거절, 부분 조사와 기타 다른 형태의 무응답 유형 등이 포함되어 있다. 조사 대상 내에서의 무응답은 4,100개가 발생했으며, Type A로 표기한다. 또한 건강면접조사의 응답률에는 계산되지 않지만 paradata 파일에는 포함되어 있는 대체 조사 자료는 9,994개로, Type B로 표기한다.

조사 결과 유형에 따른 분포는 다음과 같다.

---

1) audit trail(감사추적기록): 시스템에서 일어나는 일을 저장해 놓은 파일로 log를 통해 남긴 흔적(작업의 투명성을 제공하고 신뢰성을 향상)

2) keystroke file(키스트로크파일): 키보드를 친 모든 정보를 담고 있는 파일

3) trace file(추적파일): Server에서 에러가 발생했을 때 문제의 원인을 찾기 위해 사용되는 파일(Control 파일의 손상을 대비하기 위한 파일로 문제의 원인을 기록하는 파일)

〈표 2-1〉 건강면접조사의 조사 결과 유형

조사 결과 유형	빈도
조사 대상 내	
면접 개체	
201-조사가 완벽하게 이루어진 경우	24,323
203-부분 조사가 이루어졌으나 이용하기에 충분한 경우	5,847
Type A 개체(무응답)	
213-언어문제	63
215-부분 조사가 이루어졌으며 이용하기에 불충분한 경우	438
216-부재, 재접촉 시도가 있는 경우	891
217-일시적 부재, 후속조치가 불가능한 경우	204
218-거절	2,156
219-기타	348
조사 대상 외	
Type B 개체	
299-군인들로만 구성되었거나 평소와 다른 주거지에 기거하는 경우, 가구주가 인종/민족성 등을 이유로 조사에서 배제시켜줄 것을 요구하는 경우 등[4]	9,994
합계	44,264

3) Paradata File 구성

가) 목록

Paradata File Description Document

Data set : ASCII 형식의 자료 제공

Variable Summary Report : 변수 목록, 변수 특성, 문항 번호, ASCII 파일 내 변수 위치 등에 대한 정보 수록

Variable Layout Report : 변수의 보편적인 특성, CHI로 부터의 변수 소스, 이용 가능 영역, 질문 목록, 응답코드 등(변수 요약 보고서에 비해 보다 구체적인 내용 포함)

Variable Frequency Report : 각 변수에 대한 빈도(%) 제공

대부분의 변수들이 조사원에 의해 스스로 기록되었으므

4) 면접 전에 건강면접조사(NHIS) 표본에 포함된 가구가 비밀보호 등을 이유로 조사를 거절하는 경우, NHIS는 가구주 명단 중 새로운 가구를 추출하여 이용하게 된다. 이 때 면접조사는 가구주 명단에 흑인, 아시아인 또는 히스패닉이 한 명 이상 포함되어 있는 경우에만 지속된다. 그 외에 면접조사는 마쳤으나 가구주가 해당 조사 결과를 배제시켜주기를 요청하는 경우가 있을 수 있다.

로 자료에는 “거절”, “모름” 등의 응답 결과가 포함되어 있을 수 있음 그러나 이와 같은 응답 결과는 전체 자료의 2% 대로 그리 큰 부분을 차지하지는 않는 것으로 보임

Sample SAS(SPSS, STATA) program[5] : 다양한 형식의 예제 프로그램 제공

나) 내용

표본 설계, 분산 추정 방법, 시간 측정 자료, 접촉 가능성, 협조성, 접촉 전략, 부분 조사되거나 중단된 면접과 관계된 변수, 측정 방법 등과 조사 지역, 개체(가구, 가구원) 특성 코드, 면접 구역 등의 조사 단위 정보가 포함되어 있으며, paradata 파일에 포함된 자료를 활용한 개별적 분석과 함께 건강면접조사 결과와 연계된 분석도 가능하다. 또한 연간 자료로 누적되는 paradata 파일을 이용한 추세 분석 등도 가능하다.

Paradata 파일에 포함된 일부 변수는 건강면접조사의 건강 관련 자료 파일에도 포함되어 있다. 예를들어, 전화 사용 및 중단 등에 대한 항목은 두 파일에 모두 포함되어 있다. 이 때 불충분하게 부분 조사된 개체나 다른 Type A 포함 개체, 그리고 Type B 포함 개체는 paradata 파일에는 포함되어 있으나 건강면접조사의 건강 관련 자료 파일에는 포함되어 있지 않다. 이에 몇몇 연구자들은 두 자료의 커버리지 문제에 대한 연구를 진행하고 있다.

4) 가중치 및 분산

Paradata 파일에 포함된 가중치는 가구 추출확률에 대한 설계가중치이며, 무응답 조정이나 사후층화보정은 고려하지 않는다. 일반적으로 가중치는 모집단에 대한 추정에 관심이 있는 경우에만 이용되며, paradata 파일을 건강면접조사 자료와 연계하여 이용하는 경우에는 건강 관련 자료 파일에 포함된 가중치를 이용한다.

표본 설계와 연관성이 큰 분산 추정은 Research Triangle Institute(이하 RTI)가 제공하는 SUDAAN[6] 방법을 이용하게 된다. 건강면접조사의 표본은 층화, 집락, 다단계표본추출 등 복합표본설계를 통해 얻어지며, 이러한 표본추출 과정에서 최적화된 분산 추정식은 테일러급수선형모형(Taylor series linearization method)을 이용하는 SUDAAN이라고 할 수 있다(RTI, 2004).

---

5) 2006년에는 SAS program만 제공되었으나 2007년부터는 SPSS, STATA program도 함께 제공

6) SUDAAN is a software product designed and developed by RTI statisticians for analyzing clustered data arising in many applications, including complex sample surveys, randomized experiments, and epidemiological studies.
- Taylor series linearization (GEE for regression models)
- Jackknife (with or without user-specified replicate weights)
- Balance repeated replication (BRR)

5) 활용 연구

가) Paradata 파일을 이용한 분석

Paradata 파일만으로 분석된 경우

2006년 건강면접조사의 paradata 파일이 구성된 후 실제 paradata를 활용한 여러 형태의 분석이 이루어졌다. 앞서 설명한 바와 같이 paradata 파일을 활용한 분석은 paradata 파일 자체만 이용하여 분석하는 경우와 건강면접조사 자료와 연계하여 분석하는 경우로 나누어 볼 수 있다. 이 중 paradata 파일만을 이용한 분석으로는 완벽하게 이루어진 면접조사에서의 평균 접촉시도 횟수나 전체 면접 시기 중 해당 면접조사가 진행된 시점(초기, 중기, 말기), 면접조사의 시작 시간(아침, 점심, 저녁)과 성공적인 면접조사를 위한 조사원의 전략 등에 대한 것이 있다. 해당 특성들을 파악한 결과는 이후 건강면접조사를 성공적으로 수행하기 위한 전략을 세우는데 중요한 정보원으로 이용되었다.

건강면접조사와 연계 분석한 경우

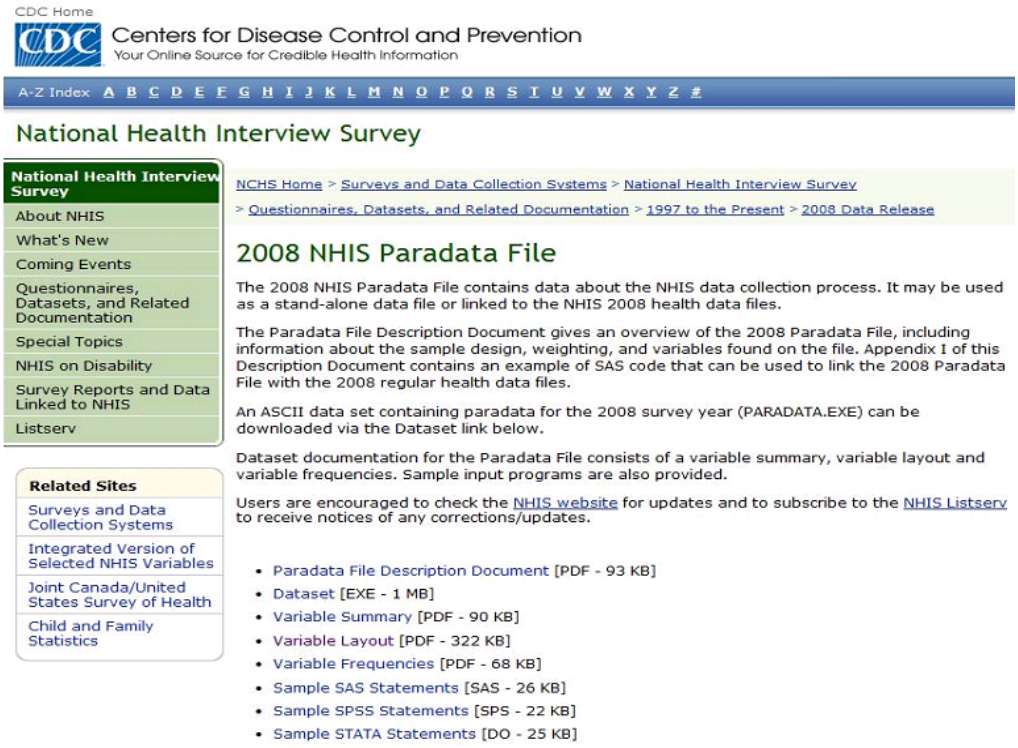
건강면접조사의 paradata와 건강 관련 자료는 같은 코드를 사용하여 입력되었으므로 두 자료는 통합하여 이용하는 것이 가능하다. 따라서 두 자료를 통합한 후 paradata 정보와 건강 관련 자료와의 연관 관계에 대한 연구가 진행되었다. 먼저 가구주의 인구·사회·경제적 특성에 따라 면접조사 과정에 차이가 있었는지 여부가 검증되었으며, 건강결정요인에 대한 면접 방식의 영향을 모형화하는 연구도 진행된 바 있다.

나) 향후연구과제

표본조사에서의 컴퓨터 활용은 점차적으로 증가하고 있는 추세이다. 조사 과정에서의 컴퓨터 활용은 조사원이 조사 도구로서 이용하는 것이나, 응답자가 조사원 없이 스스로 응답함으로써 조사에 참여하는 경우(eg. 인터넷조사)로 확대되고 있다. 이 때 컴퓨터를 활용하여 조사를 진행하게 되면 의도하지 않았으나 조사 과정에 대한 다양한 정보를 자료로 남길 수 있다. 컴퓨터에서 프로그램을 실행하게 되면 실행 과정에 대한 정보가 시스템상에 남게 된다. 이와 같은 정보는 로그(log), 감사추적기록(audit trail), 키스트로크(keystroke), 추적파일(trace file) 등과 같은 곳에 자동으로 흔적을 남기게 된다. 따라서 조사 시작 시기, 기능키 이용, 면접에서의 언어 사용 변경, 응답 간 시간, 응답 변경 등과 같은 접촉(방문) 및 응답 과정 관련 파일이 생성되며, 이로부터 추가적인 자료 확보가 가능할 것이다. 즉, 컴퓨터를 이용한 조사에서 시스템상에 남게 되는 조사 과정에 대한 흔적들을 DB화할 수 있다면, 이용 가능한 paradata의 영역이 크게 증가할 것으로 판단된다.

또한 면접조사 전·후와 조사에서의 시간과 관련된 추가적인 정보 등을 확대하여 조사한다거나, 자료 이용자들로부터의 피드백 과정에 대한 정보를 수집하여 paradata로 이

용하는 것도 고려해 볼 수 있을 것이다. 그러나 이용자들로부터의 피드백은 해당 내용을 측정하고 수치화하는 과정에 대한 객관적 기준 마련이 어렵고 판단 기준이 모호하기 때문에 그 결과에 대한 해석에 있어서는 주의해야 할 것으로 여겨진다.



출처 : http://www.cdc.gov

[그림 2-2] 2008 NHIS Paradata File

다. 캘리포니아의 건강면접조사(California Health Interview Survey; CHIS)

캘리포니아의 건강면접조사(이하 CHIS)는 주 단위 데이터의 부실 및 카운티 단위 데이터의 부재를 해결하려는 목적으로 2001년에 시작되었으며, 이후 2년 주기로 조사가 진행되고 있다. 조사 결과는 주 전체 및 카운티별 주민 건강 상태, 만성 질환, 사고, 상해, 운동 영양 상태, 의료 보험 사용 실태, 의료 서비스 이용 등에 대하여 공표되며, 자료는 일반에게 무상으로 제공된다. 무작위 전화걸기(Random Digit Dialing; 이하 RDD) 방식으로 진행되는 CHIS는 시설 거주민을 제외한 전체 캘리포니아 거주민을 대상으로 하며, 표본 가구당 성인 1명과 청소년 또는 아동을 1명씩 조사하게 된다. 그러나 해당 조사는 조사가 진행되는 지역(캘리포니아)의 특성상 인종이 매우 다양한 지역으로, 하나의

조사표와 단순화된 조사 방식으로는 원활한 조사 진행이 어려운 상황이다. 이에 실제 조사를 진행하는 Westat에서는 영어, 스페인어, 중국어(북경어, 광동어), 한국어, 베트남어 등 총 6개 언어로 조사를 진행하고 있으며, 비도시 인구와 한인 및 베트남인에 대해서는 과대표본(oversample)을 추출하고 있다. 그럼에도 불구하고 소수 인종을 포괄하는 조사의 특성상 발생할 수 있는 다양한 비표본오차를 줄이기 위하여 paradata의 수집 활동을 매우 활발하게 진행하고 있다. 예를들어 무응답오차를 줄이기 위한 방편으로 다양한 외국어 이용과 함께 사전 협조 문서를 보내는 등의 활동을 진행하고 있다. 측정오차를 줄이기 위한 방편으로는 조사 과정에서 응답자들이 응답하기 어려워하는 질문이나 질문순서 등에 대한 정보를 이용하여 지속적으로 조사표를 개선하고 있으며, 조사 시간, 조사방법, 질문을 읽는 형태 등 조사방법과 관련된 다양한 사항에 대하여 조사원 교육을 시행하고 있다. 또한 조사표에 체크하고 이를 코딩하면서 발생할 수 있는 오차를 줄이기 위하여 CAI(Computer Assisted Interviewing)를 도입하고 있다. CAI의 도입은 다양한 paradata를 보다 원활하게 수집할 수 있는 발판으로 이용되고 있다.

라. Computer Audio Recorded Interviewing(CARI)

CARI는 면접을 수행하는 동안 응답자의 동의하에 면접 과정을 침해하지 않는 범위내에서 조사원과 응답자의 음성을 녹음하는 것으로, 둘의 상호작용에 대한 잠재적인 측면을 다양하게 검토하고 분석할 수 있게 한다. 즉, 면접 중 녹음된 부분을 살펴보면, 조사원과 응답자 사이의 면접과정이 과연 믿을 만한 것인지 아닌지를 측정할 수 있게 된다. 따라서 기존의 타임스템프나 자료 양상의 변화 그리고 자판을 누르는 행동 등과 같은 감사추적기록이나 추적파일과는 거리가 있다. 조사 과정을 관리하기 위한 새로운 방법, CARI는 RTI에 의해 개발된 이래, 많은 연구자들과 국가 통계 기관의 조직적인 연구과정을 통하여 크게 성장하고 있다. CARI는 면접 장소와 면접자의 면접수행 품질을 평가하는데 촛점을 맞춘 품질 보증(QA)용으로 이용이 가능하며, 기존의 조사 방식에 쉽게 적용할 수 있도록 개발되었다.

최근 미국 센서스국의 Arceneaux(2007)는 CAPI 조사에 CARI 활용 가능성을 평가하기 위한 연구를 진행하였다. 2006년 1~2사분기에 걸친 건강면접조사에 Household Wellness Study(이하 HWS)를 위한 질문 등을 포함하여 CARI 면접을 수행하였다. CARI 시스템을 개발한 RTI와 미국 보건복지부(U.S. Department of Health and Human Services)는 CARI 이용의 실현가능성을 증명하기 위한 CARI 시스템의 품질, 모니터링 가능성, 다른 소프트웨어와의 호환성, 적절한 파일 크기로의 압축 가능성 등 여러 연구를 진행해오고 있었으며, 본 시범조사를 통하여 미국 통계국에서 수행하고 있는 모든 CAPI 조사에 CARI를 적용할 수 있는지 여부를 검증하기 위한 추가적인 현장 검증 과정을 진행하고

자 하였다. 본 연구는 미국 인구통계국 내 통계방법론과와 인구조사과 그리고 현장조사과, 기술지원과 등이 공동으로 진행하였다.

본 연구는 2006년 2월 1일부터 4월 11일 사이에 조사된 CARI HWS 현장조사와 같은 시기에 조사된 건강면접조사 결과를 이용하여 분석하는 형태로 진행되었으며, 필라델피아, 디트로이트, 켄사스 등 3개 지방사무소를 대상으로 하였다. 현장조사는 임시 조사원이 아닌 상주 조사원이 수행하였으며, CARI 조사 결과 자체의 품질을 측정하는 부분과 건강면접조사와 비교하는 부분으로 나누어 연구를 진행하였다.

분석 결과, CARI는 생산된 자료에는 영향을 미치지 않았으며, CARI로 인한 하드웨어적 시스템 오작동 비율은 1.8%로 비교적 CARI 시스템이 조사 수행에 이용되는 하드웨어에 기술적인 문제를 일으키지 않는 것으로 나타났다. 또한 조사가 진행되는 동안 녹음기는 조사원들에게 발견되지 않았으며, 녹음기가 오작동하는 경우도 발생하지 않았다. 그러나 녹음된 오디오 상태는 85.6%만이 고품질 녹음으로 판단되어 오디오품질평가시스템을 재설계할 필요가 있음을 확인하였다. 미국 통계방법론과에서는 오디오품질이 고품질로 평가되는 비율이 96% 이상이어야 CARI 시스템의 오디오 품질이 받아들일 만한 수준이라고 기준을 제시한 바 있다. CARI HWS 현장조사에서 CARI에 완전히 협조한 응답자 비율은 88.7%였으며, 면접의 일부만을 녹음하는데 동의한 부분 협조 비율은 3.0%, 완전히 거절한 비율은 8.4%로 각각 나타나 전반적으로 응답자들은 CARI를 선뜻 받아들이는 수용적 자세를 취한 것으로 나타났다. 이에 비해 조사원들의 39%는 편안하다, 29%는 반대한다는 의견이 혼합되어 있었다.

〈표 2-2〉 응답자(CARI 조사 협조)

	건강면접조사(CARI HWS)
완전 협조	88.7%
부분 협조 (면접의 일부만 녹음)	3.0%
완전 거절	8.4%

〈표 2-3〉 조사원(CARI 조사 사용)

	건강면접조사(CARI HWS)
편안함	39%
반대함	29%

마지막으로 CARI HWS 현장조사와 NHIS의 응답률과 거절률을 비교한 결과, CARI HWS 현장조사의 응답률은 81.4%, 거절률은 15.3%로 건강면접조사의 90.1%와 6.5%에 비해 응답률은 낮고 거절률은 높은 것으로 나타났다.

〈표 2-4〉 조사원(CARI 조사 사용)

	건강면접조사(CARI HWS)	건강면접조사
응답률	81.4%	90.1%
거절률	15.3%	6.5%

연구 과정에서 나타난 여러 결과를 바탕으로, 미국 센서스국에서 이루어지는 모든 CAPI조사에 CARI를 적용하는 문제에 대해서는 추가적인 연구가 더 필요하다는 결론이 내려졌다. 이 후 2008년 미국 국가선거설문조사(Lupia, et al., 2009)에서는 선거의 사전 사후 면접에 CARI를 모두 적용하였으며, 해당 조사에서의 CARI 파일은 RTI에서 면밀히 분석된 바 있다. 이들 외에도 CARI를 적용하기 위한 사전 연구가 다양한 조사에서 다양한 방법으로 이루어지고 있으며, CARI가 미국 통계청의 모든 CAPI조사에 적용되는 시기도 곧 도래할 것으로 여겨진다.

2. 캐나다 사례

가. 캐나다 통계청의 Paradata 연구

캐나다 통계청에서는 paradata를 이용하여 자료수집 과정에 대한 연구를 진행하고 있다. 실제 자료수집 과정은 자료의 품질을 결정하는데 가장 의미있는 영향을 미치며, 분기별로 전체 조사 비용의 반 이상이 조사 진행 과정에 투여되고 있다. 따라서 조사 과정을 이해하고 관리하며, 그 안에서 새로운 효율성을 발견하는 것이 매우 중요한 연구 과제로 다루어지고 있다. 이에 캐나다 통계청에서는 전화조사와 면접조사를 수행하기 위한 전화번호나 접촉 관련 정보 및 관리 정보 등을 저장해두는 paradata 창고를 개발하였다. 자료창고(data warehouse)의 paradata 관련 주요 비용은 모든 조사에 개별 표기되며, 비용 부담을 최소화하기 위해 중앙집권화하여 관리된다. 이러한 paradata 시스템은 순차적으로 관련 연구의 잠재력을 향상시킬 뿐 아니라, 조사 비용 분석에 관한 비용 정보를 적시에 제공할 수 있을 것으로 예상하고 있다. 캐나다 통계청에서는 CATI에 대한 집중과 RDD, 횡단면, 장기 조사, 사회 및 농업 등 다양한 조사에서의 paradata 이용에 대한 연구를 진행하고 있으며, 연락 시도 횟수, 시스템 작업, 접촉비율, 전화 거는 패턴, 그리고 생산과 비용의 관계 등에 대하여 분석이 이루어졌다. 또한 전략적으로 다음과 같은 문제들을 연구하는데 paradata 시스템을 이용하면 조사 개선을 위한 좋은 기회를 얻게 될 것이라고 예견하였다.

첫째, 사전조사 정보를 본조사가 이루어지는 동안의 정보와 함께 이용

둘째, 첫 번째 접촉 이후 조사 성공을 위한 이용 가능한 방법
셋째, 적극적인 관리와 조정된 자료 수집을 결합하여 응답자 설계 프레임워크 개발
넷째, 진행률 메트릭스를 기반으로 자료수집 기간 중 자료수집에 필요한 사항 예측

LaFlamme(2009)는 생산과 비용의 관계에 대한 연구에서 보다 유용한 표본조사의 생산성 지표(indicator)를 제공하는 방안에 대해 연구하였다. 표본조사에서 전화 거는 횟수를 분석함으로써 장기 조사에서 최대 3.1%~4.2% 정도의 잠재적 비용 절감 효과를 보인 것이 한 예이다. 또한 CATI 조사에서의 응답설계를 이용하여 표본조사의 생산성과 비용에 대한 지속적인 정보를 기반으로 하여 자료수집 과정을 동적으로 수정하는 방법을 연구하였으며, 전화 거는 과정은 면접자와 접촉하고 면접 진행의 가능성을 최대로 할 수 있도록 모형화되고 관리되었다.

나. 캐나다 통계청의 POINT(Pace of Interview) 시스템

Mike Maydan(2009)는 점점 복합해지는 조사 환경에서 얻어지는 다양한 자료수집 방법과 활동, 그리고 기타 여러 정보를 포함하고 있는 보고 시스템 등이 현재 캐나다 통계청에서 운용되고 있다고 밝혔다. 그는 매우 다양하고 복잡한 형태로 흩어져있는 자료를 조정하고 통합하기 위하여 CATI(Blaise telephone history file)와 CAPI(시간, 절차, 그리고 세부 결과 등)에서의 사례 수준의 paradata를 기반으로 표준화된 표본조사 보고서를 작성하였다. 이와 같은 표준화된 보고서 작성 시스템은 표본조사의 관리자나 감독자들에게 응답/무응답 비율, 무응답 원인 조사(follow up), 거절로의 전환, 추적 자료 등에 대한 정보를 제공하고 있다.

감사추적기록과 같은 내부용 paradata는 조사원들의 자료수집 행동을 평가할 수 있는 정보로 이용 가능하며, 조사원에 대한 평가 기준을 수립하고 적용시키는 기술을 향상 시키는데 이용될 수도 있다. 캐나다 통계청에서는 이와 같은 paradata의 비교적 새로운 적용을 POINT 시스템이라고 명명하였다. POINT 시스템은 조사원들의 재교육이나 조사 과정을 효율화하기 위한 객관적인 측정을 위한 것으로, 면접 과정(매 분 현장의 변화)과 무응답 항목(모르는 것인지 응답을 거절한 것인지 등을 구분) 등을 측정 도구로 이용하게 된다.

POINT 시스템에 의한 paradata 근거 보고서에는 면접자 수, 전화 시도 건수, 현장 변경 건수, 면접 과정, 항목무응답, 그리고 각 지역과 조사원에 따른 불규칙한 조사 행동 등이 포함되어 있다. 또한 감사추적기록을 이용하면 자료의 손실을 기술적으로 보완하거나, 사후검사를 통한 유효성 검사와 확인, 그리고 사후 조사 분석 및 예산에 대한 보고서에 필요한 시간 등도 활용이 가능하다.

제3절 국내 Paradata 수집 현황

1. 통계청

표본조사는 모집단의 일부를 선택하여 조사한 후 그 결과를 통해 모집단의 특성을 살펴보는 것을 목적으로 한다. 따라서 조사방법론을 근거로 하는 표본추출 방법을 이용하여 모집단을 잘 대표할 수 있는 표본을 추출함으로써 표본의 대표성을 유지하고, 표본의 적절한 크기를 유지하여 모집단 전체를 조사하지 않고 일부 표본을 조사함으로써 발생하는 표본오차를 최소화하게 된다. 그러나 표본조사 과정에서 발생되는 비표본오차는 이론적으로 그 정도를 측정하기 어렵다. 비표본오차는 조사표, 조사 진행의 수행 과정, 조사원의 특성, 응답자의 특성 등 다양한 원인에 의해 발생할 수 있으며, 그 정도를 측정한다는 것은 매우 어려운 일이다. 이와 같은 비표본오차는 대규모 조사에서 더 많이 발생하게 되므로, 일부 리서치 회사나 학계에서 수행하는 표본조사에 비해 국가기관이 수행하는 표본조사에 더 많은 영향을 미치게 된다.

이에 통계청에서는 각 지방청 단위에서 표본조사의 비표본오차를 최소화하고 조사의 품질을 향상시키기 위한 노력으로 사후조사를 시행하고 있다. 사후조사를 통하여 부실 조사를 찾아내고, 부실 조사원에 대한 사후 처리에 그 결과를 이용하기도 한다. 또한 조사가 잘 되는 시간이나 조사 방식 등에 대한 정보를 수집하여 실제 조사에 이용하기도 한다. 이러한 활동들은 포괄적인 의미에서 paradata의 수집과 활용이라고 볼 수 있다. 그러나 이와 같은 활동들은 각 지방청 단위에서 개별적으로 이루어지고 있으며, 그 기준 또한 통일 되어 있지 않아 수집된 자료가 체계적으로 정리되고 있지 못한 상태이다. 더불어 통계청 차원에서의 paradata 수집 및 활용에 대한 움직임은 아직 없는 상황이다.

Paradata를 수집하고 활용하는 것은 원활한 조사의 수행과 조사된 자료에 대한 품질을 높일 수 있는 가장 체계적이고 효과적인 방법이라고 할 수 있다. 또한 일관된 기준으로 수집된 paradata는 비표본오차를 축소하는데 유용하게 이용될 수 있다. 예를들어 응답거절 횟수나 당시 방문 시간, 접촉한 사람 등에 대한 paradata가 수집된다면 해당 케이스의 특성을 파악하여 응답 거절을 최소화할 수 있는 방안을 마련하는데 이용할 수 있을 것이다. 또한 무응답이 많이 발생하는 문항이 있다면 조사표를 재설계하는 등의 과정을 통하여 무응답을 줄일 수 있는 방안을 마련하는데 도움이 될 것이다. 즉, 일관된 기준으로 수집된 paradata는 무응답오차와 측정오차를 줄이는데 매우 유용하게 이용될 수 있다.

가. 충북통계사무소의 “Rapport 형성 프로젝트”

래포(Rapport)란 사람 사이의 상호신뢰관계를 나타내는 심리학 용어이다. 래포는 면

접조사 시 필요한 관계를 형성하여 응답 의무인식, 불신해소, 불안감 최소화 및 유대관계를 형성하여 불응가구를 예방하여 비표본오차를 축소시킬 수 있다.

현재 우리나라의 표본조사 환경은 가구형태의 다양화로 인하여 점점 열악해지고 있으며 응답자와의 면접 자체가 어려운 경우도 많아지고 있다. 통계청은 야간면접 및 휴일면접 비율이 전체 응답자와의 면접에서 상당한 부분을 차지하고 있다고 발표한 바 있다.

〈표 2-5〉 응답자와의 면접(전국)

(단위 : %)

조사업무명	조사대상	평일면접	야간면접	휴일면접
가계조사 경제활동인구조사 농어가경제조사	100.0	59.8	32.5	7.7

출처 : 통계청 '지역통계 WORKSHOP', 2007.11

또한, 통계청 발표자료에 따르면 가계조사에서 2002.10~2002.12 사이에 표본개편으로 연동표본의 불응률이 46.3%로 매우 높게 나타났으며, 가계조사 연동표본 불응사유의 가장 큰 이유는 사생활로 외부에 노출되는 것을 꺼리기 때문인 것으로 나타났다. 즉, 원활한 표본조사의 수행과 연동표본의 유지를 위하여 조사원과 응답자 사이의 유대강화가 매우 중요하다고 할 수 있다.

〈표 2-6〉 가계조사에서의 불응률(전국)

(단위 : %)

가계조사	2002	2002.10~ 2002.12	2003	2004	2005	2006
불응률	17.4	46.3	20.6	15.7	16.5	16.7

출처 : 통계청 '지역통계 WORKSHOP', 2007.11

〈표 2-7〉 가계조사 연동표본 불응사유(전국)

(단위 : %)

사생활	정부 불신	항상 바쁨	현재 바쁨	무대응	문전 박대	불응 의사	방법 불만	조사 불신	기타
25.9	16.7	14.0	12.8	6.8	6.5	6.5	5.4	3.6	1.5

출처 : 통계개발원 '통계조사의 정확성과 효율성 제고방안', 2008.4

현재 불응가구 관리방안으로 사후설득 방안은 많은 논의가 이루어지고 있으나 연동표본에 대한 전략 및 예방책은 부족한 실정이다. 이러한 연동표본 불응가구 예방을 위해

충북통계사무소에서는 'Rapport(래포) 형성 프로젝트'(2008,10)를 수행하였다.

충북통계사무소에서는 조사 불응예방을 위한 방안으로, 응답자와의 관계형성을 통한 불응가구 예방 가능성을 발견하여 '찾아가는 설명회'를 통한 래포형성 계획을 추진하였다. 면접이 어려운 아파트 조사구를 대상으로 관할 동사무소 및 관리사무소와의 협조체계를 구축하여 응답자 설득을 위한 방안을 마련하기 위한 설명회를 개최하였다.

연동그룹별 설명회 개최율은 다음과 같다. 2번 연동 개최율은 56.7%, 3번 연동 개최율은 80.5%로 나타났으며, 사회통계조사의 개체율은 10.0%로 나타났다.

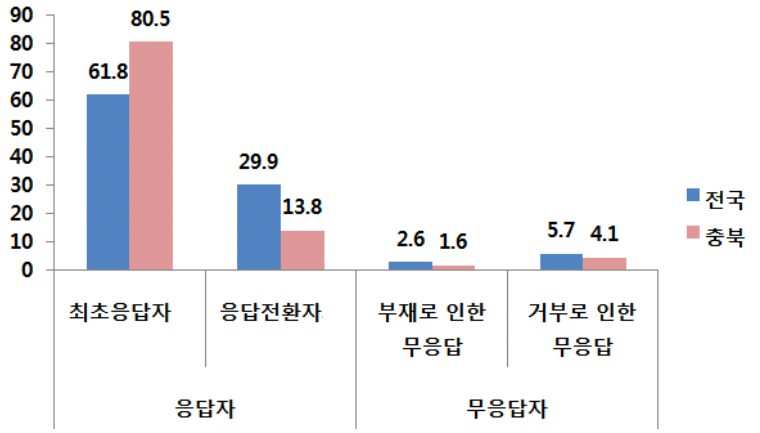
〈표 2-8〉 연동 그룹별 설명회 개최현황

(단위 : 가구, 명, %)

	2번 연동	조사구	3번 연동	조사구	사회통계	조사구
총 가구수	171	9	164	9	600	40
개최 가구수	97	5	132	7	60	4
면접 응답자수	160		340		240	
개최율	56.7		80.5		10.0	

출처 : 충북통계사무소 'Rapport 형성 프로젝트', 2008.10

설명회 실시 조사구의 응답현황 결과, 응답자 중 한번도 거부하지 않고 조사에 참여한 최초응답자의 비율은 전국에 비해 18.7%나 높게 났다.



[그림 2-3] 신표본 설득에 대한 응답현황

〈표 2-9〉 신표본 설득에 대한 응답현황

(단위 : %)

		전국[3]	충북[4]
응답자	최초응답자[1]	61.8	80.5
	응답전환자[2]	29.9	13.8
	소계	91.7	94.3
무응답자	부재로 인한 무응답	2.6	1.6
	거부로 인한 무응답	5.7	4.1
	소계	8.3	5.7
총합		100.0	100.0

1) 조사에 대해 거부없이 최초 대면 시 조사에 응답한 자
2) 최소 한 번 이상 응답거부를 하였지만 최종적으로 조사에 응답한 자
3) 통계개발원 '통계조사의 정확성과 효율성 제고 방안'(2007.10)
4) 충북통계사무소 'Rapport 형성 프로젝트'(2008.10)

사회통계조사 설명회 개최 여부별 연동표본 불응률 결과, 설명회를 개최한 조사구의 불응률은 개최하지 않은 조사구보다 불응률이 절반 이하로 낮게 나타남을 알 수 있다.

〈표 2-10〉 연동 그룹별 불응현황(사회통계조사)

(단위 : 가구, %)

	개최	비개최		
	2번	8번	9번	1번
조사대상가구수	177	165	168	183
신규불응가구수	5	10	8	13
연동그룹별 불응률(%)	2.82	6.06	4.76	7.10
개최여부별 불응률(%)	2.82	6.00		

출처 : 충북통계사무소 'Rapport 형성 프로젝트', 2008.10

불응가구에 대한 대체율은 다음과 같이 나타났다. 설명회를 개최한 조사구가 개최하지 않은 조사구보다 대체율이 22.59% 낮게 나타났다.

〈표 2-11〉 사회통계조사 설명회 개최여부별 평균 대체율

(단위 : 가구, %)

	계	개최	비개최
대상가구수	600	60	540
대체가구수	164	4	158
대체율*	27	6.67	29.26

* 대체율 : (대체가구수/대상가구수)*100
출처 : 충북통계사무소 'Rapport 형성 프로젝트', 2008.10

이와 같은 결과를 종합해보면, 응답자와의 래포형성이 최초응답비율을 높이고, 신규 표본에 대한 불응률을 낮추는 효과를 가져온다는 것을 알 수 있다.

나. 동북지방통계청의 "통계조사 대상처 응답 성향조사" -광업제조업 동향조사-

현장에서 품질 좋은 자료 수집을 위해서는 자료수집 과정에 대한 실태파악 등 통계 품질향상을 위한 다각적인 현장조사 모니터링이 필요하다. 최근 전자조사방식의 활성화로 면접조사에 의한 자료수집은 줄어드는 추세이며, 이에 따라 응답자의 중요성 및 이들에 대한 통계인식과 응답태도를 바탕으로 한 현장조사 관리의 필요성이 부각되었다.

동북지방통계청에서는 사업체부문 현장조사 관리 자료로 활용하기 위해 '조사대상처 응답성향조사 결과 보고서(2009.11)'를 작성하였다.

사업체부문 광업제조업동향조사 1,106 사업체 중 805개 사업체에 대해서 우편(581개), FAX(114개), 직접방문(110개)을 통하여 자료를 수집하였다. 응답자의 특성을 파악하기 위해 개인속성(성별, 연령, 직위, 종사자 수 등)에 관한 항목과 통계에 대한 인식도 및 응답환경 항목을 조사하였다.

조사 결과, 통계에 대한 법적 의무 인지여부에서는 모른다 59.1%, 알고 있다 40.9%로 각각 나타났으며, 개인속성별로 살펴보면 여자의 68.6%, 임원급 이상의 76.9%, 종사자 50명 미만 사업체의 63.4%가 통계 조사응답에 대한 법적의무에 대해서 모르고 있는 것으로 나타났다.

〈표 2-12〉 통계응답에 대한 법적 의무 인지 여부

(단위 : %)

		알고 있다	모른다
전 체		40.9	59.1
성별	남	45.0	55.0
	여	31.4	68.6
직위별	사원	34.6	65.4
	대리급	45.8	54.2
	팀(부서)장급	44.7	55.3
	임원급 이상	23.1	76.9
종사자별	50명 미만	36.6	63.4
	50~99명	37.5	62.5
	100~199명	43.4	56.6
	200~299명	60.9	39.1
	300명~499명	54.3	45.7
	500명 이상	54.5	45.5
	기타	26.7	73.3

통계조사 응답 시 어떤 생각에서 응답하는가에 대해서는 통계는 중요하기 때문 42.5%, 정

부에서 하는 일이므로 39.0%, 응답하기 싫지만 조사원의 끈질긴 설득 때문10.7%, 다른 사업체도 응답하기 때문 3.7% 등의 순으로 나타났다. 개인속성별로는 남자의 47.5%, 50대 이상의 47.8%, 팀(부서)장급의 47.2%가 통계는 중요하다고 생각하는 비율이 높게 나타났다.

〈표 2-13〉 통계조사 응답 시 생각

(단위 : %)

		통계는 중요하기 때문	정부에서 하는 일이므로	응답하기 싫지만 끈질긴 설득 때문	다른 사업체도 응답하기 때문	모르겠다/ 무응답
전 체		42.5	39.0	10.7	3.7	4.1
성별	남	47.5	36.3	10.0	3.2	3.0
	여	31.0	45.3	12.2	4.9	6.5
연령별	30세 미만	35.3	41.4	8.3	9.8	5.3
	30~39세	41.8	40.6	9.9	2.1	5.7
	40~49세	45.5	38.4	11.2	3.0	1.9
	50세 이상	47.8	29.0	17.4	2.9	2.9
직위별	사원	43.5	36.2	11.0	4.9	4.5
	대리급	37.0	45.0	9.7	3.4	5.0
	팀(부서)장급	47.2	36.2	10.3	3.2	3.2
	임원급 이상	35.9	41.0	17.9	2.6	2.6

담당공무원의 안면 유·무에 따른 협조 용이성은 상관없다 49.4%, 안면이 있는 경우 48.0%, 안면이 없는 경우 2.6%로 나타났으며, 개인속성별로는 성별에서는 차이가 없었지만, 비교적 젊은층인 30세 미만의 59.4%, 사원의 54.1%가 안면이 있는 조사원에게 협조가 용이한 것으로 나타났다.

〈표 2-14〉 담당공무원의 안면 유·무에 따른 협조 용이성

(단위 : %)

		상관없다	안면이 있는 경우	안면이 없는 경우
전체		49.4	48.0	2.6
성별	남	49.1	48.2	2.7
	여	50.2	47.3	2.4
연령별	30세 미만	38.3	59.4	2.3
	30~39세	49.9	49.3	0.9
	40~49세	54.1	40.3	5.6
	50세 이상	50.7	49.3	0.0
직위별	사원	44.3	54.1	1.6
	대리급	46.2	51.3	2.5
	팀(부서)장급	55.0	41.1	3.9
	임원급 이상	61.5	38.5	0.0

조사항목 중 응답하기 어려운 항목 유무에는 있다 22.9%, 없다 77.1%로 나타났다. 응답하기 어려운 항목(1순위)은 품목별 생산·출하·재고액이 75.5%로 가장 높았으며, 생산량 9.2%. 재고량 6.5% 등의 순으로 나타났다.

〈표 2-15〉 응답하기 어려운 항목(1순위)

(단위 : %)

품목별 생산·출하·재고액	생산량	재고량	조업일수 및 시간	출하량	종사자수	비고사항
75.5	9.2	6.5	3.8	2.2	1.1	1.6

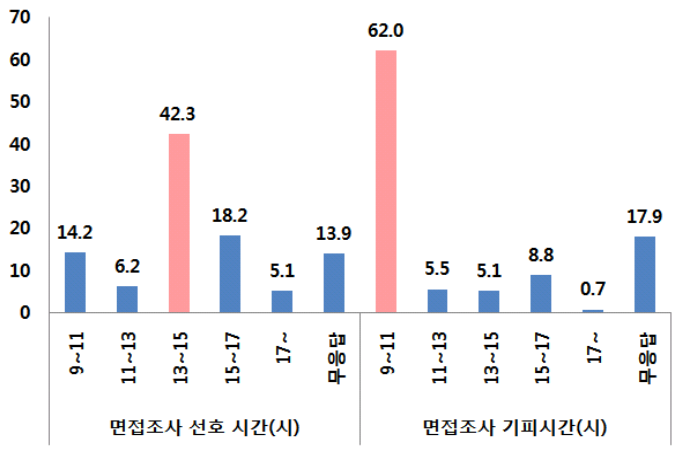
응답하기 어려운 이유(중복응답)는 회사 사정상 공시자료에 맞춰서 집계하기 때문에 32.6%, 회사 영업상 비밀 28.8%, 통계청과 사업체의 품목개념이 다름 27.2% 등의 순으로 나타났다.

〈표 2-16〉 응답하기 어려운 이유(중복응답)

(단위 : %)

회사 사정상 공시자료에 맞춰서 집계	회사 영업상 비밀	통계청과 품목개념이 다름	단위 환산이 어려움	분기 및 반기 결산	기타
32.6	28.8	27.2	25.0	25.0	8.7

면접조사 선호시간은 13~15시 42.3%, 15~17시 18.2%, 09~11시 14.2%, 무응답 13.9% 등의 순으로 나타났다. 면접조사 기피시간은 09~11시 62.0%, 15~17시 8.8%, 11~13시



[그림 2-4] 면접조사 선호·기피시간(면접조사 사업체만 해당)

5.5%, 13~15시 5.1%, 무응답 17.9% 등의 순으로 나타나, 대체적으로 오전 시간대의 방문보다는 오후 시간대의 방문을 선호하는 것으로 나타났다.

〈표 2-17〉 면접조사 선호·기피시간(면접조사 사업체만 해당)

(단위 : %)

면접조사 선호 시간(시)						면접조사 기피시간(시)					
9~11	11~13	13~15	15~17	17~	무응답	9~11	11~13	13~15	15~17	17~	무응답
14.2	6.2	42.3	18.2	5.1	13.9	62.0	5.5	5.1	8.8	0.7	17.9

적정면접 소요시간은 10분 이내 52.9%, 10~20분 이내 32.5%, 20~30분 이내 9.1%, 무응답 5.1%, 30분 이상 0.4% 순으로 나타나, 응답자의 85% 이상이 20분 이내의 면접시간을 선호하는 것으로 나타났다.

〈표 2-18〉 적정 면접 소요시간(면접조사 사업체만 해당)

(단위 : %)

10분이내	10~20분 이내	20~30분 이내	30분 이상	무응답
52.9	32.5	9.1	0.4	5.1

다. 울산사무소의 "CATI를 이용한 가구조사 개선방안 연구"

1) CATI조사 적용 사례

국내에서는 한국석유공사가 석유유가조사통계시스템에 CATI를 처음 도입(2003)하였으며, 매주 주유소, 정유사, 충전소로부터의 유가를 CATI를 이용하여 조사하였다.

국내와 마찬가지로 해외에서도 공식통계에 대한 CATI 시스템의 적용은 빠르지 않았으며, 영국의 National Center for Social Research에서 최초로 CATI를 이용한 공식통계 연구를 시작하였다. 이 후 시험적으로 전 세계에 걸쳐 27개의 조사기관(미국:18개)에서 해당 기관들의 조사나 전화조사에 CATI 시스템을 사용하였다(1987). 1990년 Statistical Methodology에서 밝힌 The Federal Committee의 보고에 의하면 1980년대 말에 전 세계적으로 설치된 CATI의 수는 1,000개가 넘었으며, 1988년 미국에서는 51개의 CATI센터를 보유하고 있는 것으로 보고되었다.

2) CATI를 이용한 시험 조사

우리나라의 통계청에서는 동남지방통계청 울산사무소 CATI조사 연구회와 통계개발

원이 2007년 CATI를 이용한 시험조사를 공동으로 실시한 바 있다. 경제활동인구조사실시 기간에 1차(2007.8.20~24)와 2차(2007.9.17~21)로 나누어 울산광역시 4개 군.구(동구, 북구, 중구, 울주군)에 거주하고 있는 울산사무소 관할 경상조사구 중 전화조사를 원하는 가구(약 500가구)를 대상으로 조사를 실시하였으며, 해당 조사구의 경제활동인구조사 담당자(조사원)가 심층면접조사와 CATI 조사를 병행하여 취업자 부문(조사 항목 14개), 실업자 부문(조사 항목 15개), 비경제활동인구 부문(조사 항목 14개)에 대해서 각각 조사를 실시하였다.

1차 및 2차 시험조사에서 면접조사와 CATI 내용 비교결과, 대체적으로 일치하는 것으로 나타났다.

〈표 2-19〉 1차 및 2차 시험조사 분석결과 - 면접조사/CATI 내용 비교

(단위 : 명)

구분	1차			2차		
	총 응답자	불일치	일치	총 응답자	불일치	일치
취업여부	351	0	351	400	0	400
직장유뮤, 직업		1	350		0	400
1주 실업		0	351		0	400
4주 실업		0	351		0	400
일한 시간		0	351		1	399
추가취업희망여부		0	351		0	400
취업희망 여부		1	350		0	400
전직 유무		1	350		1	399
종사자수		2	349		0	400
취업시기		8	343		1	399

면접 및 CATI 조사방법 비교 결과, CATI 직접조사 참여율이 면접조사 직접조사 참여율보다 3.5% 높은 것으로 나타나, CATI 조사 방법이 가구원별 직접조사 참여율을 높일 수 있을 것으로 예상된다.

〈표 2-20〉 조사방법 비교

(단위 : 명, %)

구분	총 가구원	직접조사	간접조사
면접조사	751(100.0)	363(48.3)	388(51.7)
CATI	751(100.0)	389(51.8)	632(48.2)

면접조사 희망가구 및 불응가구에 대한 CATI 기법 적용 시도 결과, 희망가구에서의 CATI조사 시도는 실패(2가구)보다는 성공(34가구)이 많았으며, 불응가구에서는 성공(1가구)보다는 실패(4가구)가 많은 것으로 나타났다.

〈표 2-21〉 면접조사 희망가구 및 불응가구에 대한 CATI 기법 적용 시도

(단위 : 가구)

적용사례	계	성 공	실 패
면접조사희망가구 → CATI조사시도	36	34	2
불응가구 → CATI조사로불응설득시도	5	1	4
총 계	41	35	6

CATI 조사방식의 도입으로 방문 횟수를 단축(4~5회 → 1~2회)시킬 수 있었으며, 이로 인하여 조사원의 조사구 내 이동시간(145분 $\pm\alpha$)을 크게 줄일 수 있었다. 비용 측면에서는 면접조사가 CATI방식에 비해 2.4배 정도 더 소요되었다.

〈표 2-22〉 CATI 조사방식 도입 효과

구분	이동시간 (왕복)		조사 시간 (평균)		휘발유 (왕복)		전화통화 요금 (원)	합계 (원)
	시간 (분)	금전적 가치 (원)	시간 (분)	금전적 가치 (원)	소모량 (l)	가격 (원)		
면접조사	145.2	23,232	164.4	26,304	7.0	10,920	0	60,456
CATI	0	0	143.1	22.896	0	0	2,390	25,286

* 본 조사 소요시간 기준임(보조조사표 배부, 답례품 배부 등 소요시간 제외)

** 비용 산출 시 측정 기준

- 1분(60초) = 160원, 시간을 돈으로 환산하는 방법 : V = W((100-t)/100)/C
  (V:시간당 가치, W:시간당 임금, t:세율, C:생활비)
- 1L 가격 : 1,560원(2007.8.), 1L당 12km(1500cc 승용차 기준)
- 전화통화요금 : 3분당 50원으로 환산

녹취 기능에 대한 반응에서는 '하든 안 하든 상관없다' 를 포함하여 긍정적으로 응답한 응답자가 21명으로 부정적으로 응답한 응답자 6명에 비해 많은 것으로 나타났다.

〈표 2-23_1〉 녹취에 대한 반응_1

(단위 : 명)

꼭 해야한다	하면 좋겠다	상관없다	안하는게 좋다	할 필요없다
4	6	11	5	1

CATI 조사 방식을 예상치 못한 응답자들은 조사 방식에 대한 당황과 부끄러움으로 인하여 약간의 거부감을 보이기도 하였으나, 충분한 설명으로 이해가 가능하였으며 실제 조사의 불응으로 이어질 만한 거부감은 보이지 않았다.

〈표 2-23_2〉 녹취에 대한 반응_2

(단위 : 명)

당황해했었다	부끄러워했다	재미있어했다	심한 거부감 형성	약간의 거부감 형성	응답을 거부함	상관없다
7	5	3	3	6	1	2

시험조사 직후에 CATI 시험조사에 참여한 조사관 18명을 대상으로 CATI 이용에 관련된 설문조사를 실시하였다.

CATI 시스템 사용 편리성 설문조사 결과, 1차에서는 편리했다(61%) 2차에서는 익숙해서 편리한 편(56%)이 가장 높았다.

〈표 2-24〉 CATI 시스템 사용 편리성

(단위 : 명, %)

1차			2차			
편리했다	그저 그랬다	불편했다	익숙해서 편리한 편	많이 편리	1차보다 편리하지만 그래도 불편	많이 불편
11(61)	7(39)	0(0)	10(56)	4(22)	4(22)	0(0)

향후 CATI를 이용한 통계조사 이용 바램 결과, 1차 및 2차 모두 바란다의 비율이 높았으며, 1차에서 CATI 조사의 도입을 '바라지 않는다' 2명에서 2차에서는 1명으로 변화하였다.

〈표 2-25〉 향후 CATI를 이용한 통계조사가 개발되어 조사에 이용되기를

(단위 : 명, %)

	1차	2차
바란다	16(89)	17(94)
바라지 않는다	2(11)	1(6)

CATI 시험조사에 참여한 응답자(1차:210명, 2차:182명)를 대상으로 한 통화음성만족도 결과, 총 응답자가 1차와 2차에 차이가 있지만 '잘 안 들렸다'고 응답한 사람이 줄어들었다.

〈표 2-26〉 통화음성 만족도(1차/2차)

(단위 : 명, %)

	1차	2차
잘 들렸다	198(95)	172(96)
보통이다	6(3)	3(2)
잘안들렸다	5(2)	3(2)

조사소요시간 만족 결과, 1차, 2차 모두 '적당했다'가 가장 높을 비율을 차지하였다.

〈표 2-27〉 조사소요시간(1차/2차)

(단위 : 명, %)

	1차	2차
적당했다	191(91)	164(91)
그저그랬다	16(8)	11(6)
길었다	2(1)	2(3)

전화조사와 면접조사 선호 결과, 1차 및 2차 응답자간에 총응답자의 차이가 있었다. 면접조사를 계속적으로 원하는 가구수는 차이가 없었지만, 이를 전체비율로 봤을 때는 28%에서 32%로 증가하였다.

〈표 2-28〉 원하는 조사방법(1차/2차)

(단위 : 명, %)

	1차	2차
전화조사	151(72)	124(68)
면접조사	58(28)	58(32)

전화조사를 원하는 시간대는 오전 9시~12시와 오후 3시~6시를 원하는 사람이 가장 많았으며, 1차와 2차의 전체적인 비율은 큰 변화가 없었다.

〈표 2-29〉 원하는 전화조사 시간대(1차/2차)

(단위 : 명, %)

	1차	2차
07:00~09:00	5(2)	5(3)
09:00~12:00	77(37)	60(34)
12:00~15:00	29(14)	19(10)
15:00~18:00	67(32)	60(34)
18:00~21:00	30(15)	34(19)

이상의 결과를 종합하면, CATI를 이용한 조사가 현장조사에서의 실제 조사시간을 감소시킬 수 있으며, 조사 개선, 품질관리 및 효율적인 업무 처리에 도움이 될 수 있을 것으로 판단된다.

통계청에서는 맞벌이 가구나 1인 가구 등 만남이 어려운 가구들에 대한 대체 조사방법 등으로 CATI 조사를 추진하고 있으며, 2009년을 기준으로 CATI의 평균 적용비율은 전체의 9.8%를 차지하고 있다. 현재 경제활동인구조사, 가축동향조사, 집세조사, 어업생산동향조사 등 4종은 CATI 조사방법으로 조사 중이다. 2010년(하반기)에는 어류양식동향조사에도 CATI를 도입할 예정이다.

CATI 조사 방식의 도입 시에는 녹취파일의 관리와 활용방안 등의 연구를 통하여 조사의 품질을 높이는데 노력을 기울여야 하며, 조사시스템의 해킹방지를 위한 보안프로그램 강화에도 힘써야 할 것이다.

2. 리서치업체

2008년 말 시장조사 및 여론조사업 사업체수는 270개로 전년에 비하여 2.2% 감소하였으며, 종사자수와 매출액은 각각 6,289명과 5,480억 원으로 16.0%, 24.8% 증가하였다. 관련 사업체수가 2.2% 감소했음에도 불구하고 매출액이 크게 성장한 이유는 경영환경 변화에 적극적으로 대처하고 시장과 고객의 수요를 정확히 파악하기 위한 기업과 정부의 수요가 크게 증가했기 때문인 것으로 보인다. 다음은 전문·과학 및 기술서비스업조사 보고서 중 시장 및 여론조사업에 관련된 부분을 발췌한 것이다.

〈표 2-30〉 시장 및 여론조사업 사업체수, 종사자수, 매출액

(단위 : 개, 명, 억원, %)

산업분류	사업체수			종사자수			매출액		
	2007년	2008년	증감률	2007년	2008년	증감률	2007년	2008년	증감률
시장조사 및 여론조사업	276	270	-2.2	5,422	6,289	16.0	4,392	5,480	24.8

이와 같이 시장 규모가 확대되고 있는 시장 및 여론조사업체 중 비교적 규모가 큰 한국 리서치와 한국 갤럽의 비표본오차 관리 상태를 살펴보고, paradata에 대한 인지 정도와 paradata 수집의 필요성에 대한 인식 정도를 알아보았다.

가. 한국 리서치

서울에 본사를 둔 한국리서치는 부산, 대구, 광주, 대전 등 4개 지방에 사무소를 설치하여 운영하고 있으며, 전북, 제주, 강원의 경우에는 사무소는 없으나 관리자가 있어 전국 조사 시 조사 수행을 지원하는 형태로 운영되고 있다. 상주하는 정규직원은 220명 내외이며, 연간 수행조사건수는 약 700건으로 이 중 20~30% 내외를 정기 조사로 진행하고 있다. 다음은 비표본오차를 축소를 위해 한국 리서치에서 수행하고 있는 조사 과정에 대한 관리 사항이다.

1) 조사원 관리

조사원 관리의 시작은 조사원에 대한 교육이라고 할 수 있다. 표본조사를 직접 수행하는 조사원은 실제 조사에 투입되기 전 조사방법 및 여러 돌발 상황에 대한 대처 방법 등을 사전에 교육받게 된다. 특히 면접 및 전화조사의 경우에는 조사원과 응답자의 상호작용에 의해 조사가 진행되므로 조사원에 대한 교육의 필요성이 매우 크다고 할 수 있다. 한국 리서치에서는 총 5단계 조사원 교육을 실시하고 있으며 그 과정은 다음과 같다.

1단계(조사일반교육) - 조사의 필요성 및 전반적인 내용 설명
2단계(Role Play) - 역할면접 실시, 역할면접 실시 후 연구원 및 수퍼바이저가 보완사항 지적/해결 방안 재교육
3단계(평가 및 재교육) - 자료수집 시작 후 본조사가 시행되는 중에 1-2회 정도 실시
4단계(시범조사) - 실제 field에서의 시범조사로 조사원 test
5단계(평가 및 탈락 여부 확정) - 시범 조사 평가 자료를 바탕으로 탈락 여부 확정

일반적으로 3단계까지의 교육을 진행하고 있으며, 조사 내용이 민감한 사항이거나 조사 의뢰자가 별도로 원하는 경우에는 5단계까지의 교육과정을 수행하게 된다.

조사원들의 면접 관할 지역은 가능한 조사원의 거주 지역을 중심으로 인근 지역을 배분하는 것을 원칙으로 하고 있다. 그러나 이 부분에 있어서는 개인 비밀보호에 문제가 발생할 소지가 있으며, 응답자가 정직하게 답하기를 꺼려 거짓응답을 할 가능성이 있다는 비판을 받기도 한다. 그러나 응답자와 조사원이 아는 사람일 경우 거짓 응답(매출액, 수입, 학력 등)을 방지할 수 있는 장치가 될 수 있으며, 조사원과 응답자 사이에 유대관계가 있는 경우 조사를 좀 더 수월하게 진행할 수 있다는 장점 때문에 이와 같은 방법을 유지하고 있다.

2) 조사 과정 관리

면접조사의 경우 자료수집 담당 연구원 및 수퍼바이저에 의해 한국 리서치 자체 개발 시스템으로 관리되고 있으며, 1일 통제 시스템과 주간 통제 시스템 등이 운용되고 있다.

1일 통제 시스템은 조사원이 해당 수퍼바이저에게 1일 조사 활동을 보고하고 통제를 받는 것으로, 조사 진행 사항의 보고(성공 부수, 조사 중 문제점 등), 조사 진행상의 문제점 및 해결방안, 조사원의 근태와 관련한 내용, 검증 결과에 대한 통보 및 주의 당부 등이 포함된다. 이 때 특정 조사구에 문제가 발생할 경우 본사의 해당 조사 관련 연구자 및 관리자들이 회의를 통해 문제를 해결하게 된다.

주간 통제 시스템은 프로그램을 활용하여 조사원 별 주간 진행 상황을 통제하고 조사원이 한 주에 할 수 있는 부수에 크게 못 미칠 경우 조사 진행을 독려하거나, 과대하게 많이 한 경우 해당 사유를 파악하여 조사 전반의 진행을 통제하고 조사의 질을 확보할 목적으로 이용된다. 이와 더불어 실사 진행 중 연구원 및 수퍼바이저와 조사원이 만나 조사 관련 전달 사항, 애로사항 등을 이야기 할 수 있는 간담회를 실시하여 면접원의 소속감과 조사 전반에 대한 이해도를 높이기 위한 활동을 이어가고 있다.

전화조사의 경우에는 전화조사 담당 연구원 및 수퍼바이저의 전일 모니터링에 의한 관리가 수행되고 있으며, 1일 단위로 조사 전체 진행 상황 및 거절, 중단, 차후 재조사 요청 등의 상황을 검토하게 된다. 상주하는 전화조사원은 50명 내외이며, 이들을 모니터링 하는 모니터 요원은 3명, 이들을 관리하는 관리담당은 2명 정도로 약 10%의 관리 요원을 활용하고 있다.

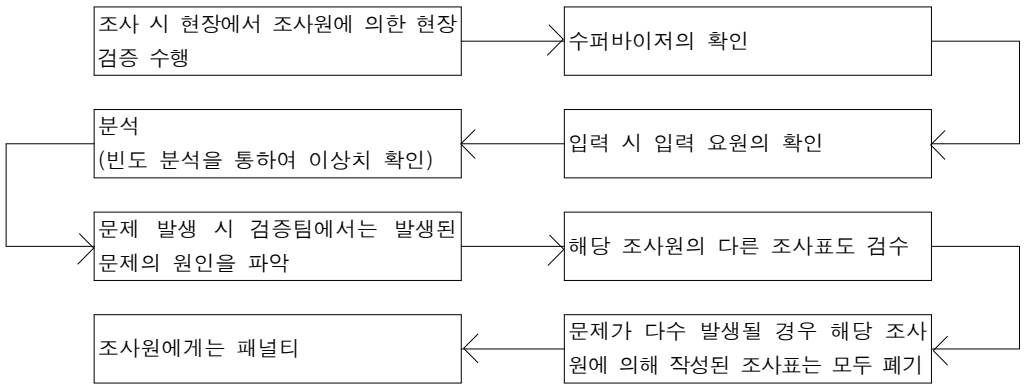
인터넷조사의 경우에는 응답자가 자기기입식으로 조사에 응하기 때문에 조사원이 응답자의 조사 참여에 직접적으로 관여할 수 없다. 따라서 연구자가 수시로 진행상황 및 문항별 응답상황을 확인하는 활동을 하고 있다.

조사 관리 과정은 H-PQI(Process Quality Investment)를 이용하고 전산화하여 관리되고

있으며, 의뢰인의 요청이 있는 경우 접촉 횟수 등 조사 과정에 대한 자료를 sheet로 별도 기록하게 된다. 따라서 조사 과정에 대한 응답자 성향별 분석이 가능하다. 그러나 실제 조사 과정 관리에 대한 분석은 따로 이루어지고 있지 않다. 조사 불응의 경우 개인차가 크고 응답자의 특성에 의한 불응은 많지 않은 것으로 판단하고 있으며, 조사 과정에 대한 관리를 통해 자료의 품질을 향상시킬 수 있다는 것은 모두가 인정하고 있는 사항이기 때문이다. 즉, 관리 중 발견되는 문제는 그때그때 해결하며, 종합적인 피드백 활동은 없는 것으로 파악된다.

3) 사후관리

조사 종료 이후 응답자에 대한 대상 검증 및 입력된 자료의 빈도 검토 후 조사가 잘못 진행되었거나 결과값이 이상한 경우, 표본 구성이 모집단과 다른 경우 등의 문제가 확인되면 재조사를 실시하게 된다. 사후관리과정은 다음과 같다.



[그림 2-5] 사후관리과정

사후관리를 위하여 별도의 검증팀이 전체 설문 내용 중 중요 항목을 전화를 통해 확인하고 있으며, 이 때 검증팀이 자료수집 부서와 접촉하는 것을 불허함으로써 검증의 정확성을 보장하게 된다. 만약 여러 단계의 확인 과정을 거치는 도중 문제가 발견되는 경우에는 에러 유형(결번, 설문 내용 틀림, 대상이 아님 등)을 파악한 후 조사표를 폐기하고 재조사를 수행하게 된다. 만약 해당 조사원이 수행한 모든 조사표에서 문제가 발견되는 경우에는 해당 조사원이 조사한 모든 조사표는 폐기되며, 조사원에게는 패널티(이후 조사에서의 조사원 활동 불가 등)가 부여된다.

4) 기타

표본으로 결정된 조사 대상자의 e-mail을 알고 있는 경우에는 조사 이전에 e-mail로 조사에 대한 사전 통보를 하게 되며, 전화로 방문 약속 후 방문 조사를 수행하게 된다. e-mail을 모르는 경우에는 먼저 전화로 조사 관련 내용을 보내준다는 메시지 전달과 함께 e-mail을 확보하고 일정 기간이 지난 후 전화로 방문 약속을 잡게 된다. 전화 또는 e-mail 정보가 모두 없는 경우에는 사전에 직접 방문하여 조사에 대한 안내를 실시하게 되는데, 만약 집이 비어있다면 문에 안내문을 부착하는 등의 활동을 하게 된다.

표본으로 선정된 가구가 조사 협조를 "거절"하는 경우, 최소 2회 이상 방문하여 조사 참여를 독려하고 있으며, 3회째는 시간과 요일을 달리해서 다시 방문하고 있다. 그럼에도 불구하고 대상 가구가 계속 조사를 거절하는 경우에는 표본 대체를 수행하게 된다. 표본 대체는 동일한 성격의 표본층을 대상으로 미리 준비된 예비 표본에서 추출된다. 조사가 어려운 특이 모집단(장애인, 외국인 노동자 등)의 경우에는 조사 자체가 어렵기 때문에 공문, 홍보물, 답례품 등을 이용하여 연구자가 직접 접촉을 시도하게 된다. 한국 리서치에서는 특이한 상권에 대한 조사를 수행하면서 해당 상권이 밀집한 특정 지역을 선정하여 전수 조사하고 연구자의 사전 조사에 의한 유의추출이나 눈덩이 추출방법 등을 이용하여 표본을 추출하여 이용한 경험이 있다.

무응답이 발생한 경우에는 표본 조사 결과만이 관심 사항일 경우 무응답 인정 후 결과를 분석하게 되며, 모집단 추정이 목적인 경우에는 무응답을 인정하지 않게 된다. 이 때는 층의 평균을 이용한 평균 대체 방법을 주로 이용한다.

나. 한국갤럽

1979년 세계 최대의 조사네트워크인 Gallup International에 가입하였으며, 1980년부터 전국 실사네트워크를 구축하여 월 20여 회 이상의 전국규모조사를 실시하는 등 연간 약 1,000건 이상의 국내 최다 조사실적을 보유하고 있다. 한국갤럽은 전국 6대도시(서울, 부산, 대구, 대전, 광주, 전주)에 정식직원이 상근하는 지사를 보유하고 있으며, 강원 및 제주에는 실사 네트워크를 구축하여 운영하고 있다. 다음은 비표본오차 축소를 위해 한국갤럽에서 수행하고 있는 조사 과정에 대한 관리 사항이다.

1) 조사원 관리

조사원에 대하여 기본 소양 및 현장조사방법 등과 함께 해당 조사 내용에 대한 교육을 수행하고 있다. 이 때 조사원들은 조사의 난이도에 따라 조사 경험을 바탕으로 배치되며, 가급적 조사원 거주 인접 지역으로 배분하게 된다. 한국 리서치에서와 마찬가지로 비밀보호나 거짓응답 등의 문제점이 제기되었지만, 조사원 확보의 어려움과 경험 많은

조사원들의 지속적인 조사 활동 지원 등을 이유로 인접 지역 배치는 지속적으로 이어지고 있다. 인접 지역에서의 조사 활동은 조사원의 사전 지식과 인맥 등으로 인해 조사 활동을 수월하게 하는 장점이 있다.

2) 조사 과정 관리

전화조사의 경우 의뢰인의 요청이 있는 경우에 한해 녹취 후 제공하게 되며, 면접조사가 이루어지는 현장보다는 비표본오차가 작아 따로 녹취하거나 조사 과정을 면밀히 검토하지는 않고 있다. RDD 방식에 의한 CATI의 경우 조사 과정에 대한 자료가 log 형태로 남아있으나, 조사 과정에 대한 분석을 목적으로 DB화하지는 않은 상태이다. 서울에 상주하고 있는 전화조사원은 약 100명 내외이며, 이들을 모니터링하는 모니터 요원은 2~3인으로 구성되어 있다.

조사 과정을 체계적으로 관리하기 위해서는 많은 비용이 필요하며, 조사 관리 자료를 이용한 분석에서 얻을 수 있는 정보와 적용 범위가 비용 대비 효과적일지에 대한 시각에는 회의적인 것으로 보인다.

3) 사후관리

조사표를 에디팅하는 과정에서 에디팅 요원은 조사표에 이상은 없는지 여부를 일차적으로 확인하게 된다. 이때 논리적인 오류가 발생되면 작위성 여부에 따라 조사원에 대한 재교육을 실시할 것인지, 아니면 해당 조사원의 조사표를 모두 폐기할 것인지를 결정하게 된다. 에디팅하는 과정 이후에는 통계자료처리팀이 다시 한번 논리적인 오류 및 비체계적 오류를 제거(Data Cleaning)하게 된다.

4) 표본 관리

한국갤럽에서는 일반 표본조사와 함께 패널조사도 수행하고 있다. 패널조사에서는 고정된 패널을 유지하는 것이 매우 중요한 사항이므로 조사 거절이나 미접촉 등의 문제가 발생할 경우 3~5회 재방문을 권유하고 있다. 그러나 일반 표본조사의 경우에는 3회 정도 접촉을 시도하고, 조사가 불가능한 경우에는 방문 횟수, 거절 사유 등을 기록하게 된다. 그러나 이러한 정보가 따로 자료로 만들어지지는 않고 있다.

특이 모집단의 경우에는 조사 발주 시 모집단을 제공받는 경우가 많으며, 제공이 불가능한 경우에는 조사 수주 자체를 포기하고 있다. 즉, 관리가능한 수준의 조사 요구(모집단틀 제공) 시에만 조사를 수행하고 있다.

또한 무응답을 인정하지 않고 있음으로 무응답이 발생할 경우 100% 대체 방법을 이용하며 표본 대체 방법은 건별로 다른 방법을 적용하고 있다.

제4절 Paradata 수집 방안

Paradata의 수집 방법은 조사방법의 다양화와 정보화에 맞추어 크게 변화하고 있다. 종이와 연필을 이용하여 면접조사를 수행하던 시기에는 해당 조사표에 내검하는 단계마다 다른 색의 볼펜을 이용하여 잘못된 부분이나 재확인해야 할 부분들을 표기하였으며, 그 표기들이 바로 paradata의 역할을 수행하였다. 이러한 장치들은 누가 몇 번에 걸쳐 조사표를 고쳤는지 여부를 확인하기 위한 것이며, 원칙적으로 변경된 사항들을 자료로 남겨두도록 하였다. 그러나 대부분의 조사표들은 최종 결과만이 입력(coding)되어 최종 조사 결과까지의 과정(조사 과정 포함)을 알 수 없는 경우가 다반사였다. 이 후에는 면접조사에서 면접 내용을 녹음하는 형태로 paradata는 수집되었으나 조사 내용이 녹음되고 있다는 상황 자체가 조사원과 조사대상자 모두에게 영향을 미친다는 문제가 지적되었다. 또한 현장에서 제대로 녹음이 되지 않거나 녹음 테이프를 분실하는 등 체계적인 관리에 문제가 있었다.

조사방법에 컴퓨터를 이용하기 시작하면서 paradata 수집에는 혁신적인 변화가 일어났다. 컴퓨터를 이용할 경우 조사 과정에서 생기는 여러 상황(입력 내용을 바꾸거나 조사를 진행한 시간, 시작 시간, 끝난 시간 등)이 조사 진행과 동시에 자동적으로 남을 수 있게 된 것이다. 조사원 또는 입력요원이 자료 입력 과정에서 값을 수정한다면 해당 값을 수정하는 이유를 제시해야 하며, 제시된 이유는 paradata의 일부로 별도 관리된다. Biemer and Caspar(1994)는 자료입력 과정을 자료로 만들어 평가함으로써 조사 과정의 품질을 개선하기 위한 증거로 제공하였다. 물론 컴퓨터 시스템에 자동으로 남게 되는 paradata는 분석을 위한 자료가 아니기 때문에 매우 지저분(messy)하여 실제 분석 자료로 바로 이용하기는 어렵다. 또한 시스템상에 남겨진 조사원과 응답자의 행동이나 사건 등이 어떤 특정한 생각이나 의도를 가지고 있는지 등에 대한 정보는 알 수 없기 때문에 이용상에 한계를 가지게 된다. 그러나 표본조사에 이용하는 소프트웨어와 하드웨어가 비약적인 발전을 이루기 시작하면서 활용 가능한 paradata의 종류와 양은 크게 확대되었으며, 해당 자료를 바탕으로 얻을 수 있는 정보의 질 또한 크게 향상되었다. 즉, 소프트웨어를 이용하여 적은 비용으로 방대하면서도 양질인 paradata의 수집이 가능해졌으며, 발견된 문제점을 해결하기 위한 개선 활동도 시스템을 이용하여 쉽게 처리가 가능해졌다.

컴퓨터를 이용한 표본조사는 대국민을 상대로 하는 국가 통계뿐 아니라 보건, 복지, 사회, 경제 통계 등 다양한 부문의 다른 여러 표본조사로 점차 확대되고 있다. 또한 조사 도구로써의 컴퓨터의 역할은 인터넷을 이용한 표본조사의 확대로 이어지고 있다. 인터넷을 이용한 조사는 조사 대상자가 자기기입방식으로 조사에 참여하는 경우가 대부분이므로 이전 조사방법에 비해 자기-통제(self-administration)로 진행된다. 따라서 수집되는

paradata도 이전 조사원-통제(interviewer-administration)와는 다른 모드로 이루어지게 되며, 분석된 paradata의 활용에도 차이를 가지게 된다. 인터넷을 이용한 표본조사를 생각해보자. 인터넷을 이용한 조사에서는 조사 진행 과정에서 발생되는 다양한 형태의 paradata를 시스템상에서 실시간으로 얻을 수 있다. 조사를 시작한 시간과 마친 시간을 이용한 조사 진행 시간(조사에 걸린 총시간)이나 문항 당 소요 시간, 조사가 진행된 시간 등에 대한 자료가 확인 가능하다. 이와 같은 paradata는 이후 인터넷을 이용한 조사를 개선하는데 이용될 수 있다. 문항당 소요 시간이 너무 길다면 조사대상자가 응답하기 어려운 사항일 가능성이 크므로 질문을 쉬운 형태로 바꿀 수 있으며, 조사에 걸린 시간이 너무 길다면 조사대상자가 조사를 마칠 때까지 집중하여 참여할 수 있도록 독려하는 방안을 마련해야 할 것이다. 또한 조사가 한밤중에 이루어지는 경우가 많다면 해당 조사대상자들의 조사를 도울 전화상담(Call Center; Hot Line)을 24시간 운영하는 방안에 대해 고민해야 할 것이다. Paradata가 조사원이 없는 상태에서 진행되는 자기-통제 형태의 조사에서 응답자의 행동을 파악하고 그들의 성향을 알아내는데 매우 유용한 정보원 역할을 하고 있는 것이다. 이에 인터넷조사 관련 연구자들은 paradata를 분석하여 인터넷조사를 재설계(redesign)하는 문제에 많은 시간을 투자하고 있으며, 실시간으로 측정오차를 줄이기 위한 paradata 분석의 유용성을 입증하는 연구를 진행하고 있다. 실제로 자기기입식으로 진행되는 인터넷조사에서의 paradata는 측정오차를 최소화하고 무응답오차를 줄이는데 유용하게 이용될 수 있으며, 응답자의 성향을 보다 면밀히 파악하여 성향가중치 계산에 활용함으로써 보다 개선된 형태의 추정치 계산을 가능하게 할 것이다.

이제 paradata 수집 방안에 대하여 자세히 살펴보자. 수집방안은 조사원-통제 조사방법과 자기-통제 조사방법으로 나누어 제시된다.

1. 조사원-통제 조사

가. CATI 방식에서의 모니터링

CATI(Computer Assisted Telephone Interviewing) 방식을 이용한 조사에서는 전화 조사원들을 모니터링(Monitoring)함으로써 paradata를 수집할 수 있다. CATI 방식을 도입한 초기에는 모니터할 조사원의 선정이 모니터 또는 감독관의 판단으로 이루어졌으나, 최근에는 표본추출 방식을 도입하여 모니터할 조사원을 추출하고 있다. 또한 과거에는 조사원의 행동을 관찰하고 코드화하여 이용하는 것이 전부였다면 요즘에는 조사원들의 입력 단계도 추가하여 조사하고 있다. 이와 같은 모니터링 자료를 활용하여 통계적 단계조절 차트를 작성하고 이를 바탕으로 조사 과정을 피드백하고 평가하는데 이용할 수 있다. CASIC(Computer Assisted Survey Information Collection)에서는 효율적이고 품질 높은

조사를 수행하기 위한 조사 과정의 자료를 실시간으로 제공할 수 있으므로 모니터링을 기반으로 하는 평가시스템의 능률을 향상시키고 보다 확장시킬 수 있다.

나. 재면접

재면접(reinterviews)은 비용이 많이 드는 평가 방법이지만 응답편차(variance)와 응답편의(bias)의 추정과 조사원의 변조를 방지하는데 매우 유용하게 이용될 수 있는 방법이다. 과거 종이와 연필을 이용하여 조사를 진행하던 시기에 비해 CAI(Computer Assisted Interview) 방식에서는 재면접의 과정이 매우 간소화되었으나, 재면접의 필요성이 없어진 것은 아니다. CAI 방식에서는 표본조사와 자료 생성이 현장에서 매우 빠르게 이루어지며, 재면접의 효율을 최대로 하기 위하여 표본을 선정하는 것이 허용되고 있는 추세이다. 즉, 어떤 변수에 대한 응답이 보통의 경우와 달리 너무 적거나 또는 너무 큰 비율에 속하는 값이 나오면 재면접 표본으로 뽑힐 확률 또한 그에 비례하여 부여하는 것이다.

다. 조사수행 및 자료 생성 과정 측정

과거에 조사 활동(수행)과 자료 생성 과정을 측정하기 위해서는 조사 관리자가 종이로 조사된 조사표를 일일이 살펴보아야 했다. 면접이나 거절, 비접촉 등에 대한 정보와 개방형 질문이 얼마나 명확하게 조사 되었는가, 무응답 항목 수, 검너뛰 오류(skip errors) 등 조사원 조사 활동을 평가하는 작업들이 이루어졌으며, 이들 과정을 요약하여 보고서로 작성하였다. 즉, 조사 활동 및 자료 생성 과정에 대한 측정은 일반적으로 조사원의 활동에 기반한 것이었으며, 해당 자료를 수집하기 위해서는 많은 비용과 시간이 필요하였다.

그러나 CASIC 도입 이후에는 자료가 전자적 형식으로 수집되며, 매일 업데이트되고 테이블 형태 등으로 수시 제공하는 것이 가능해졌다. “잘 모름” 응답 수, 거절, 응답 전환 횟수, 내검 실패, 조사 진행 시간(length), 조사 수행 시간(언제 조사를 수행했는가), 이상치나 이상한 패턴의 응답 등에 대한 정보를 자료 입력과 동시에 알 수 있다. 또한 이와 같은 자료는 조사 비용과 관련된 자료로도 활용이 가능하다. 즉, 각 조사원이 얼마나 많은 조사를 수행할 수 있는가를 알 수 있을 뿐만 아니라 의미 있는 조사 결과를 얻기 위하여 각 면접당 필요한 비용 등을 추정할 수 있게 된 것이다.

무엇보다 중요한 것은 조사 과정 전반에 걸쳐 얻어지는 관련 자료는 실시간으로 수집 가능하고 매일 확인이 가능하므로, 조사 활동을 조절하고 소프트웨어 등을 이용하여 품질을 강화하는데 이용할 수 있다는 것이다.

라. 시간 측정

조사와 관련된 시간을 측정하는 전통적인 방법은 조사표를 작성하기 시작한 시간과 마친 시간을 확인하는 것이었다. 이와 같은 시간 자료는 연구자에 의해 전체 응답 시간으로 계산되어 이용될 수 있으며, 응답에 필요한 시간 등을 조절하는데 이용되었다. 조사 관련 시간을 측정한 자료를 이용하면 조사가 너무 빨리 끝나거나 너무 길게 진행된 경우 면접 또는 전화 조사 과정에서의 문제가 있었음을 알 수 있다. 즉, 너무 빨리 응답이 이루어진 문항의 경우, 조사원이나 응답자가 문항을 끝까지 읽지 않았을 가능성이 크기 때문에 문항당 평균응답시간(benchmark(또는 gold standard) time)을 제공하는 등의 대책을 마련할 수 있다.

마. 추적자료이용

많은 표본조사 주체들은 CASIC 이용자들에게 안내데스크를 제공하고 현장에서의 문제를 수집하고 있다. 여기에는 소프트웨어와 하드웨어 문제가 모두 포함된다. 조사원 관련 문제뿐만 아니라 시스템, 장비 등과 관련된 문제들도 관련 정보를 적절하게 관리함으로써 유용하게 이용할 수 있다. 예를 들어 노트북을 이용한 조사에서 노트북이 3년 정도 조사에 이용되었다면, 배터리를 바꾸거나 장비를 업데이트 하거나 하는 작업 등이 필요하다. 즉, 현장에서의 소프트웨어와 하드웨어 문제에 초점을 맞춘 조사 진행 조절 사항을 자료화하여 관련 정보를 관리하면 보다 원활한 조사 진행이 가능하다.

바. 조사원 기록

대부분의 CAI 소프트웨어 시스템에서는 조사원이 각 문항 수준 또는 전반적인 수준에 대한 입력 노트를 작성하게 된다. 이 때 장비의 상태 등을 기록한 조사원의 노트를 활용하면 유용한 정보를 많이 얻을 수 있다. 즉, 조사원의 노트는 조사 과정에 대한 자료의 source로 활용하기에 더 없이 좋은 정보원인 것이다.

사. CARI

CARI(Computer Assisted Recorded Interviewing)는 면접을 수행하는 동안 조사원과 응답자의 음성을 바로 녹음하는 것으로 둘의 상호작용에 대한 잠재적인 측면을 다양하게 검토하고 분석할 수 있게 하는 것으로 타임스템프나 자료 양상의 변화 그리고 그 외의 자판을 누르는 행동 등과 같은 audit trails나 trace file과는 거리가 있다. CARI의 개발과 적용은 수년에 걸쳐 제기되었으며 기술적인 발전을 이루어왔다. 초기 CARI는 면접 장소

와 면접자의 면접수행 품질을 평가하는데 촛점을 맞춘 품질 보증(QA)용으로 이용되었으며, 그 적용 범위는 점차 확대되어 가고 있다. 미국 센서스국에서는 CAPI 조사에 CARI를 적용한 현장 평가를 수행(2006)하였다. 당시 CARI가 적용된 CAPI 조사는 423건이었으며, CARI가 생산된 자료에 영향을 미치지는 않는 것으로 나타났다. 또한 현장 적용에 있어 기술적인 문제없이 적절하게 작동하였으며, 응답자들은 CARI를 선뜻 받아들이는 수용적 자세를 취하였다. 이에 반해 조사원들은 39%는 편안하다 그리고 29%는 반대한다는 의견이 혼합되어 나타났다. 이는 조사 과정에 대한 정보가 조사원 평가에 반영될 수 있으며, 그들이 모니터링되고 있다는 부담이 작용했기 때문인 것으로 보인다. (Arceneaux, 2007)

아. 키스트로크 파일(추적감사기록 또는 추적파일)

많은 CAI 소프트웨어 시스템에서 자동으로 생성(log) 되는 키스트로크 파일(keystroke file)은 조사원이나 조사대상자가 조사를 수행하며 응답 버튼을 누를 때마다 자동으로 입력된다. 이 파일에는 매우 많은 양의 정보가 들어있으며, 조사 과정의 흔적이 모두 남게 됨으로 응답 수정 횟수나 효율적이지 못한 개방형 응답의 입력 등 평가하기 어려운 조사원 관련사항을 평가할 수 있다는 장점이 있다. 그러나 해당 파일이 분석하기 좋은 형태로 구조화되어 있지 않아 자료를 요약하여 분석 가능한 형태로 가공해야 하는 어려움이 있으며, 프로그램 호완성의 문제가 발생할 가능성이 높다는 단점이 있다. 조사 수행 프로그램을 기반으로 하는 키스트로크 파일이 생성되며, 윈도우 등 일반적으로 이용하는 프로그램과 호완이 되지 않는 경우가 종종 발생한다.

2. 자기-통제 조사

가. 응답자의 조사 수행 시간

자기-통제 조사에서는 응답자가 원하는 시간에 조사를 시작하고 끝낼 수 있으며, 조사에 응하는 중에 다른 활동에 제약을 받지 않게 된다. 따라서 응답자들은 조사에 응하는 동안 집중하여 조사에 응하는 경우도 있지만 일부 응답자들은 개인적인 활동이나 다른 작업을 병행하는 등의 활동으로 조사에 집중하지 않는 경우도 발생하게 된다. 이 경우 응답자가 실제 조사에 집중하지 않으므로 인한 조사 결과의 신뢰도 저하는 피할 수 없는 일이며, 불성실 응답이 발생할 가능성이 크다. 따라서 응답자들의 조사 수행 시간을 확인하고 분석 가능한 형태의 자료로 변환하여 이용하면 불성실 응답의 발생 여부와 해당 특성이 조사 결과에 미치는 영향에 대한 분석이 가능할 것이다.

나. 문항 당 및 전체 설문 완성 시간

응답자의 조사 수행 시간과 밀접한 연관관계가 있는 것으로, 각 문항에 응답하는데 걸린 시간과 전체 설문을 완성하는데 걸린 시간을 DB화하고, 이를 조사의 품질, 신뢰성 등을 측정하는 자료로 이용하게 된다. 만약 특정 문항에 대한 응답 시간이 지나치게 길거나 짧다면 해당 문항은 조사표 재설계 과정에서 재정비를 거쳐야 할 것이다. 만약 특정 문항에 대하여 응답자가 이해하기 어렵거나 오해할 소지가 있다면 해당 문항을 수정하거나 삭제하는 등의 조치가 필요하며, 조사표 내 문항의 재배치도 고려해 보아야 할 것이다.

다. 응답 변경 문항, 건수, 경로

자기 통제 조사에서는 조사원 통제 조사에 비해 응답자가 설문 문항에 보다 신중하게 생각하고 솔직하게 답변하는 성향을 가지게 된다. 이와 같은 장점은 민감한 사안이나 개인적인 질문에 매우 유용하게 이용된다. 그러나 신중하게 생각할 수 있다는 장점은 응답의 잦은 변경과 연결 가능하다. 따라서 어떤 문항에서 응답 변경이 발생했는지, 몇 차례 응답 변경이 이루어졌으며, 어떤 경로로 변경이 이루어졌는지 등에 대한 정보가 필요하다. 이를 통하여 응답자의 응답 성향을 파악할 수 있을 것이며, 응답자가 신중한 응답을 위하여 보인 응답 패턴을 파악함으로써 응답자의 속내를 보다 면밀히 파악할 수 있는 단서로 이용할 수 있을 것이다.

라. 키스트로크 파일(추적감사기록 또는 추적파일)

대표적 자기-통제 조사인 인터넷조사의 경우 응답자의 조사 수행 과정에 대한 모든 흔적이 정보로 수집되며, 해당 정보들을 분석하기 좋은 형태로 구조화하여 저장하는 작업이 프로그램상에서 가능하다. 즉, 조사 과정에서 발생하는 부산물들을 분석 가능한 형태의 자료로 변환하여 저장하는 부분까지 프로그램화하여 운영함으로써, 수집된 자료를 간단한 테이블 등으로 그려 확인하거나 다른 관련 자료와 연계하여 분석하는 작업 등이 쉽게 이루어질 수 있다.

제5절 Paradata의 활용 방안

조사과정에서의 paradata 수집은 새로운 것이라고 할 수 없으며, 모든 조사 과정에 존재하고 있다. 그러나 질문지를 사용하여 조사하던 전통적인 방식을 벗어나 PDA나 컴퓨터 등을 활용하는 등 조사방법이 다양한 형태로 확대됨에 따라 조사 과정에서 발생하는 paradata에 대한 세부적인 부분과 범위, 그리고 수집 방법 등에 대한 관심도 급격하게 확장되고 있다. 또한 최근에는 조사에서의 응답률이 감소하는 추세이며, 무응답 편의에 의한 조사 결과의 신뢰도 저하는 우려할 만한 수준으로 증가하고 있다. 여기에 다양한 측정 방법으로 인한 측정오차와 급증하는 조사비용은 조사를 수행하는 측과 조사 결과를 이용하는 측 모두에게 중요한 관심 대상이 되고 있다. Paradata는 다양한 분야에서 수집 가능하며, 활용 범위 또한 매우 넓다. 면접조사의 경우 면접을 수행하는 조사원 관련 사항이나 조사 대상자의 특성, 질문의 형태나 위치, 조사 도구 및 이용하고 있는 시스템의 종류에 따라 조사 과정에서 수집 가능한 paradata는 매우 많다. 또한 수집되는 자료는 체계적으로 정리되어 있는 경우가 거의 없어 이용하고자 하는 목적에 따라 선별하고 필요에 맞게 입력하는 작업이 필요하다. 목적에 맞게 선별된 paradata는 분석 과정을 통하여 본래의 역할을 수행할 수 있게 된다.

그렇다면 실제 수집 가능하고 실용성이 큰 paradata는 어떤 것이며, 우리 현실에 맞는 paradata를 어떻게 수집하고 이용해야 하는가? 조사 과정에서 자연스럽게 발생하는 방대한 양의 paradata를 조절하고 이용하기 위해서는 paradata를 수집하기 위한 도구 및 기술과 더불어 paradata를 수집하는 명확한 목적이 필요하다.

1. 조사 관리 측면

Paradata의 기초적인 분석은 조사 환경을 이해하는데 좋은 정보로 활용될 수 있으며, 이후 조사 관리 과정에 대한 개선과 조사원들의 재교육 등을 통하여 궁극적으로 표본조사 자료의 품질 향상에 기여할 수 있다. 그렇다면 paradata로 부터 얻은 정보를 조사 관리에 활용할 수 있는 방법에 대하여 고민해볼 필요가 있다.

우리나라는 분산형 통계를 생산하는 국가로 모든 국가 통계가 통계청에서 집중적으로 생산되지는 않는다. 최근 들어 유사·중복 통계를 통합하여 예산 낭비를 막고 보다 정확한 통계를 생산하기 위한 움직임들이 활발히 이루어지고는 있으나, 아직도 일부 국가통계는 여러 부처와 한국은행 등에서 생산되고 있다. 현재 통계청에서 생산하고 있는 통계는 인구, 사회, 경제 분야에 걸쳐 총 52개의 통계(2010.6)가 생산되고 있으며, 이 중 10개의 통계(가공8, 보고2)를 제외한 나머지는 조사를 통해 작성되고 있다. 통계청에서 수

행하고 있는 조사는 인구총조사나 전국사업체조사 등 전수와 표본 추출에 의한 표본조사로 나뉘어지며, 표본조사는 동일한 표본을 이용하는 경우와 새로운 표본을 이용하는 경우로 나뉜다. 예를들어 경제활동인구조사의 경우에는 한번 표본으로 뽑힌 가구는 3년 동안 조사에 참여하게 되며, 해당 가구는 각 지방청의 정규직 조사원에 의해 조사된다. 즉, 동일한 조사원이 동일한 가구를 매월 방문하여 조사를 진행하는 것이다. 이에 반해 지역별 고용조사의 경우에는 1년에 한 번 조사가 수행되며, 경제활동인구조사에 비해 많은 표본을 추출하여 조사함으로 임시조사원들을 활용하여 조사를 수행하게 된다. 이 때 임시조사원을 관리하는 문제는 조사의 품질과 직결되는 문제라고 할 수 있다.

본 연구원이 추가로 파악한 자료에 의하면, 임시조사원들의 부실조사 문제는 그대로 간과하기에는 전체 조사에서 차지하는 비중이 작지 않다. 따라서 사후검증을 통해 부실조사를 진행한 임시조사원을 색출하고 해당 조사원이 조사한 조사표들을 표본에서 제거하거나, 해당 조사원을 다시 고용하지 않는 등의 조치보다 원천적으로 부실조사를 예방하는 방법을 고민할 필요가 있다. 이와 같은 부실조사의 예방은 조사 관리 과정을 체계화하고 효율화하는 과정을 통해 이룰 수 있을 것이다.

조사 관리 과정에서의 paradata의 활용 중 대표적인 것은 조사원들의 방문 기록을 활용하는 것이다. 조사원이 응답 대상을 접촉하기 위한 방문 관련 기록들을 통하여 조사원이 실제 조사를 수행했는지 여부를 모니터링할 수 있다. 또한 재방문수나 성공적인 조사를 위하여 취한 조치(메모, 협조 공문, 사전 전화 등)가 무엇인지 파악이 가능하며, 이는 성공적인 조사를 위한 전략 마련에 훌륭한 정보로 이용될 수 있다. 즉, 응답 대상을 방문하기 전 사전 조치는 어떤 것이 효과적이며, 언제 방문하고, 조사에 소요되는 시간은 어느 정도가 적당한지, 그리고 대상을 만나지 못한 경우 몇 번을 더 방문하는 것이 효율적(비용, 시간 등 고려)인지 등을 paradata를 통하여 알 수 있으며, 이를 전략적으로 이용하면 조사를 관리하는데 큰 도움이 될 것이다.

조사원의 허위·부실조사를 근절하고, 응답자들의 응답부담을 최소화하며, 실제 조사 성공률을 높일 수 있는 조사의 수행이 어떻게 이루어질 수 있는지에 대한 총체적인 해답이 paradata 안에 있다고 할 수 있다.

2. 조사 설계 측면

조사 설계 측면에서 paradata를 어떻게 활용할 것인지를 논의하기에 앞서 paradata의 역할이 자료수집 과정이나 자료의 품질을 직접적으로 평가하는 기준이 아니라 보조하는 역할임을 상기할 필요가 있다. 즉, paradata를 통하여 문제가 왜 발생했는지 여부보다는 어디서 발생된 것인지를 파악하여 간접적으로 어떤 문제가 발생했는지를 파악하는데 보

조역할을 하도록 해야 한다. 예를들어, 조사표에 포함된 각 문항당 응답 시간이 paradata로 수집된다면, 조사대상자의 인지 수준에서 혹은 조사표의 구성이나 문항의 위치, 문구 등에서 문제가 발생했음을 알게 될 것이다. 즉, 특정 질문에서의 응답시간이 대체적으로 길게 나타났다면 해당 문항에서 문제가 발생하였음을 알 수 있다. 이 때 발생된 문제가 어떤 것인지에 대한 정보도 함께 수집이 가능하다. 응답자가 질문의 의도를 인지하지 못할 수도 있고, 문구가 잘못됐을 가능성도 있으며, 조사표의 흐름상 올바른 위치에 있지 않을 수도 있다. 우리는 paradata를 통하여 조사표의 어느 부분에서 문제가 발생하고 해당 문제가 어떤 것인지에 대한 정보를 얻게 되는 것이다. 이러한 문제가 왜 생겨났는지에 대한 원인 분석과 해결방안 마련은 발견된 문제를 바탕으로 paradata와 원자료 등을 이용하여 총체적으로 해결하게 될 것이다. Paradata는 문제가 어디서 발생했으며, 어떤 문제인지에 대한 충실한 정보를 제공하게 되며, 이는 해당 문제의 원인을 파악하고 해결방안을 마련하는 실마리로 이용될 것이다.

3. 총오차 축소 측면

CAI(Computer Assisted Interview) 시스템의 도입 초기, paradata의 관심은 키스트로크 파일, 추적파일, 감사추적기록 등이었다. 조사 진행 시 응답을 입력하는 과정에서 컴퓨터 시스템에 의하여 자동적으로 생성되는 부수적인 정보들은 에러 진단이나 잘못된 부분을 고치는 등의 기술적인 목적을 위하여 이용되었다. 이에 paradata를 연구했던 초기의 연구자들은 묻고 답하기 과정을 고려한 질문을 준비하고, CAI 시스템에서 조사원과 응답자가 어떻게 소통할 것인가를 결정하기 위하여 paradata를 이용할 것을 제안하였다. 그러나 이제는 paradata를 수집하고 가공하고 분석함으로써 무응답오차나 측정오차 등을 줄여 궁극적으로 총오차를 축소하는 목적으로 전환되어가고 있다. 즉, paradata 연구 초반에는 조사방법을 개선하는데 이용할 목적이 대부분이었다면, 지금은 총오차를 축소하여 조사의 품질을 향상시키는데 그 주된 목적이 있다고 할 수 있다.

조사 과정에서 수집되는 다양한 형태의 paradata는 무응답의 원인이 무엇이며 어떤 해결 방안을 제시할 수 있는지에 대한 정보를 제공할 수 있다. 대표적 paradata인 조사원의 특성이나 조사 행태, 응답자의 특성, 그리고 조사표의 상태나 조사 수행 과정에 대한 정보는 단위무응답이나 항목무응답 등의 원인이 되는 특성이 무엇인지에 대한 정보를 포함하게 된다. 미국의 경우 여러 인종과 민족이 넓은 지역에 흩어져 살고 있기 때문에 응답자의 특성에 맞는 조사원 이용이 조사의 성공에 여부에 큰 역할을 하는 것으로 알려져 있다. 인종차별은 불법이나 아직도 유색인종을 적대시하는 백인들이 존재하며, 미국에 거주하고 있으나 영어를 사용하지 못하는 사람들도 많다. 따라서 응답자의 특성에

맞는 조사원을 파견하고 조사표를 응답자가 자유롭게 이용할 수 있는 언어로 준비하는 등의 작업을 통하여 무응답을 축소하기 위한 노력을 기울이고 있다.

우리나라의 경우에도 최근 국제결혼으로 인한 다문화가족의 확대[7]와 이들의 언어 문제 등이 조사 환경에 큰 변화를 일으키고 있다. 따라서 보다 원활한 조사를 위하여 다양한 언어의 조사표를 준비하거나 조사원에 대한 교육을 개선하는 등의 작업이 필요하다. 또한 해당 특성을 갖는 표본을 과대추출하는 등의 노력을 통하여 표본조사 결과의 정확도를 높일 수 있을 것이다. 또한 면접기간 중 응답률이 가장 높은 기간이나 조사 수행 시간, 조사 시 이용한 조사표의 형태 등은 단위(항목)무응답을 줄이기 위한 조사방법 개선에 좋은 정보를 제공하게 될 것이다. 조사표 내에 항목무응답이 자주 발생하는 항목에 대해서는 조사 설계 과정의 개선을 통하여 궁극적으로 항목무응답을 줄일 수 있는 방법을 강구할 수 있으며, 단위무응답인 응답자들의 특성을 따로 분석함으로써 성공적인 조사 수행을 위한 전략 마련의 기초자료로 이용할 수 있다. 즉, 조사 과정에서 얻어지는 여러 현상에 대한 정보와 조사원과 응답자의 특성들은 조사의 참여도를 높이기 위한 전략을 세우는데 이용될 수 있으며, 이는 궁극적으로 무응답을 감소시켜 무응답 오차를 줄이는 역할을 수행하게 될 것이다.

무응답 축소를 위한 paradata의 활용 외에 측정오차를 줄이기 위한 활용 방안도 점차 주목받고 있다. 측정오차는 표본 조사를 진행하는 과정에서 측정의 결함으로 발생하게 되는 오차로 조사에 이용하는 도구, 조사표 구성, 설문 문항, 조사원의 조사 방식, 응답자의 이해도 등에 의해 발생된다. 이미 설명한 바와 같이 측정오차의 발생 원인들은 모두 paradata의 범위에 포함된다. 즉, paradata를 통하여 조사에 이용하는 PDA나 넷북, 종이 조사표 등 다양한 조사 도구가 응답에 미치는 영향을 측정하게 되며, 조사표의 구성과 응답에 어려움을 일으키는 문항 등에 대한 정보 수집이 가능하다. 또한 조사원이 기록한 조사 방식이나 응답자의 이해도 등은 측정오차를 줄이기 위한 개선의 기반 자료로 활용될 수 있을 것이다.

무응답오차와 측정오차가 축소되면 해당 오차를 포괄하는 비표집오차가 축소되며, 이는 궁극적으로 총오차의 축소로 이어진다. 또한 앞서 설명한 바와 같이 무응답오차나 측정오차의 축소는 조사 관리와 조사표 설계 측면에서의 paradata 활용과 깊은 연관관계를 가지고 있다. 조사 과정을 효율적으로 관리하고 조사표 등을 개선해 가는 과정은 궁극적으로 비표집오차를 축소하여 총오차의 축소로 연결된다고 할 수 있다.

---

7) 2004년 외국인과의 결혼이 전체 결혼의 11.2%로 10%대를 돌파한 이래 지금까지 해당 추이는 계속 유지되고 있는 실정임

4. 통계 품질 측면

궁극적으로 paradata를 수집하고 분석하여 그 결과를 실제 조사 과정에 적용하는 일은 통계 품질을 향상시키기 위한 목적으로 이루어진다. 이 때 통계 품질을 측정하기 위한 품질의 차원(ONS Guidelines for measuring statistical quality)은 관련성(Relevance)[8], 정확성(Accuracy)[9], 시의성(Timeliness)[10], 정시성(Punctuality)[11], 접근가능성(Accessibility)[12], 명확성(Clarity)[13], 비교가능성(Comparability)[14], 일관성(Coherence)[15] 등으로 분류되며, 이 중 paradata와 관련된 품질의 차원은 정확성과 접근가능성 그리고 명확성 등이 될 것이다. 포괄적인 범위에서 보자면 조사 과정에서 새롭게 파악되는 이용자의 요구사항을 반영하여 조사표를 개선하는 과정을 거치고, 이러한 일련의 과정을 통하여 이용자 요구에 맞는 통계를 생산하게 된다면 paradata는 통계 품질의 차원 중 관련성을 향상시키는데 일조할 수 있을 것이다. 또한 조사원들의 안정된 조사방법과 측정 도구(조사표 등)의 적절한 사용은 통계의 비교가능성과 일관성에도 영향을 미칠 것으로 예상된다.

이와 같이 Paradata 즉, 조사 관리 과정에 있어서의 수많은 부산물들은 체계적이고 객관적인 과정을 통하여 자료화될 수 있으며, 이를 통하여 효율적인 조사 관리와 조사비용의 축소, 나아가 비표본오차 축소를 통한 총오차 축소 효과를 가져올 수 있다. 이는 정확한 통계를 생산하고 생산된 통계의 개념을 명확히 하며 자료수집 과정과 생산 통계의 생성 과정을 투명하게 함으로써 통계의 품질을 향상시키게 되는 것이다. 고품질의 국가통계 생산은 국가 통계에 대한 신뢰를 두텁게 하며, 나아가 국가 경쟁력에도 큰 도움이 될 것이다.

통계청에서 생산하고 있는 통계들은 소수의 총조사를 제외한 대다수가 표본조사에 의해 생산된다. 또한 국가 통계의 특성상 표본의 크기가 매우 크며, 이에 따른 비표집오차는 객관적으로 측정된 바 없으나 무시하지 못할 정도의 크기일 것이다. 비표집오차는 조사원, 조사표, 조사과정, 측정도구, 응답자, 조사 환경 등 조사와 관련된 모든 항목에서 발생할 수 있으며, 해당 오차를 적절히 제어하지 못한다면 아무리 큰 표본을 추출하여 표집오차를 줄였다 하더라도 품질이 높고 신뢰할 수 있는 통계를 생산하는 것은 어려운

8) 통계자료가 포괄범위와 내용에 있어서 이용자의 요구사항을 충족시키는 정도
9) 추정값과 참값의 근접성
10) 공표시점과 그 자료를 조사하는 시점 사이의 시간 경과 정도
11) 공표한 날짜와 사전에 계획된 공표날짜 사이의 시간 지체 정도
12) 이용자가 데이터에 손쉽게 접근할 수 있는 정도와 활용가능한 통계표와 그 통계가 어떻게 만들어 졌는지에 대한 정보 이용 가능성
13) 그 자료가 어떻게 만들어졌는지에 대한 정보, 예시 및 부수적인 사용상의 조언 등에 대한 결과의 충분성
14) 시간의 흐름과 영역에 따라 자료가 비교되는 정도
15) 서로 다른 출처, 작성 방법에 따라 작성된 통계자료지만, 동일한 사회 현상을 반영하는 경우 각 통계자료가 얼마나 유사한지를 나타나는 정도

일이 될 것이다. Paradata 연구 초반에는 조사 방법을 개선하는데 이용되었다면, 현재는 Paradata를 수집하고 가공하고 분석함으로써 무응답오차나 측정오차 등을 줄여 궁극적으로 총오차를 축소하여 조사의 품질을 향상시키는 방향으로 목적이 전환되고 있다.

제6절 결론 및 향후과제

총오차를 축소하기 위한 다양한 노력은 표집오차를 측정하고 이를 줄이는 방법에 집중되어 왔으며, 비표집오차는 측정이 어렵고 축소 방법 또한 쉽지 않은 일로 치부되어 왔다. 따라서 총오차 축소를 위한 노력에는 늘 한계가 존재하였으며, 일정 부분 이상의 오차 축소는 불가능하였다. 그러나 표본조사를 통한 통계 생산에 있어 총오차는 통계의 신뢰성 확보 등 통계 품질과 직결된 문제로, 이를 해결하기 위하여 많은 연구자들은 비표집오차를 축소하는 방안에 관심을 기울이기 시작하였다. 이에 표본조사 과정에서의 비표집오차 축소 노력이 이어졌으며, 이러한 노력과 표본조사 과정에 대한 객관화된 자료가 필요하게 되었다. 이러한 자료는 비표집오차를 측정하고 이를 축소하기 위한 방안을 마련하는데 이용될 수 있으며, 실제 분석 가능한 표본조사 과정에 대한 자료를 통하여 비표집오차를 축소하고, 궁극적으로 자료의 신뢰성과 품질 확보를 기대할 수 있게 되었다. 이러한 표본조사 과정에 대한 자료를 paradata라고 하며, 이는 표본조사 과정에서 얻어지는 부가자료이다. 이 자료를 이용하면 응답자의 패턴을 이해하는데 도움이 되며, 조사원의 조사 방법, 조사표의 문제점, 조사 방식에 대한 응답 형태 등 다양한 분석이 가능하다. 또한 조사 방법이 PDA나 컴퓨터를 이용하는 형태로 발전하고 있는 요즘에는 이러한 paradata의 수집이 보다 원활하게 이루어질 수 있으므로 자료 확보가 매우 용이하고 활용 범위 또한 매우 넓다.

우리나라에서는 통계청뿐 아니라 여러 통계생산기관과 민간 리서치업체 등 표본조사를 시행하는 여러 기관에서 특별히 조사 방법에 대한 자료를 수집하고 있지 않다. 그러나 비표집오차를 최소화함으로써 총오차를 축소할 수 있는 자료수집 방안을 마련하기 위한 노력은 다방면으로 이루어지고 있다. 또한 실제 paradata라는 명목으로 자료를 수집하여 이용하고 있지는 않으나, 관련 자료는 표본조사 과정 속에서 계속하여 수집되고 있다. 대표적인 예로 CATI를 이용한 조사에서는 조사 과정에 대한 녹음이 이루어지고 있으며, 인터넷을 이용한 조사의 경우 키스트로크 파일 등이 자동으로 저장되게 된다. 해당 자료들을 별도의 목적으로 DB화하여 이용하고 있지는 않으나 이용 가능한 paradata가 이미 존재하고 있는 것이다. 앞서 살펴본 것과 같이 paradata는 자체적인 분석 목적으로도 이용되지만 원자료와의 연계 분석으로 그 활용도를 더 넓힐 수 있다. 예를 들어

CATI 분석에서의 녹취록을 모니터링함으로써 조사가 잘 되는 시간을 파악하여 해당 시간에 집중적으로 조사를 수행하고, 조사원들의 조사 방식을 모니터링함으로써 조사원 교육에 이용할 수 있을 것이다. 또한 조사 진행 상황에 대한 정보를 통하여 불성실응답을 구분하고 이를 조정하는 노력도 필요하다.

Paradata를 활용할 수 있는 분야는 매우 넓으며, 궁극적으로 총오차 축소를 통한 표본조사 결과의 품질 향상을 제고할 수 있다. 또한 조사 과정에 대한 모니터링이 가능함으로 조사원 교육이나 조사표 개선 및 조사 방식에 대한 개선 과제를 해결하는데 실마리를 제공할 수 있을 것이다. 우리나라는 아직 paradata라는 조사 과정에 대한 자료를 체계적으로 수집하여 이용하지 않고 있으나 통계 선진국에서는 이미 관련 연구가 많은 연구자들에 의해 이루어졌으며, 연구 범위나 활용 범위에 있어 매우 발전되어 있다. 따라서 이들의 자료 수집 과정이나 활용 방법, 그리고 실제 paradata를 통한 표본조사 방법에 대한 개선 결과를 벤치마킹하여 우리의 표본조사 방식을 보다 개선하고 객관적으로 평가할 수 있는 계기를 마련할 필요가 있다. 이러한 노력은 표본조사 방법을 통하여 생산된 통계에 대한 신뢰성을 높이고 고품질의 통계를 생산하는데 결정적인 역할을 하게 될 것이다.

Paradata 수집을 위한 방법은 앞서 소개한 바와 같이 다양하며, 이미 수집되었으나 활용하고 있지 않은 paradata 또한 다수이다. 먼저 활용되지 못하고 있는 paradata를 분석 가능한 형태의 자료로 정리하는 작업이 선행되어야 할 것이며, 해당 자료의 활용분야도 다양하게 발굴하여야 할 것이다. 또한 눈부시게 발전하는 IT 기술과 더불어 현재 이용하고 있는 조사 방법을 벗어나 한 단계 더 진화된 조사 방법의 도입도 지속적으로 고려하여야 할 것이다.

대표적인 예로 휴대폰은 멀티미디어기기(디지털카메라, MP3, 캠코더, TV 등)로 진화하면서 차세대 대표 정보기술(IT)산업으로 발전하고 있다. 휴대폰으로 인해 다양한 정보 및 지식들을 편리하게 공유할 수 있게 되었으며, 특히 최근에는 비약적인 기술 발전과 다양한 애플리케이션(application)의 추가 등으로 휴대폰은 빠르게 진화하고 있다. 그 중 스마트폰(아이폰, 옴니아폰, 안드로이드폰 등)은 휴대폰 진화의 대표주자라 할 수 있다. 스마트폰은 핸드폰에 PC성능을 탑재한 것으로 휴대폰이 제공하는 일반 모바일 기능과 함께 PC의 정보처리, 파일관리, 일정관리 등의 기능을 제공한다. 이동 중 인터넷 통신, 팩스 전송 등이 가능하며, 모바일 인터넷을 기반으로 다양한 서비스를 공간 제약 없이 사용할 수 있다.

최근 각 기관에서는 모바일 오피스를 위하여 스마트폰을 사원들에게 지급하는 등 스마트폰을 이용한 업무활용방안에 관심을 보이고 있다. 또한 스마트폰의 GPS기능으로 조사원들을 관리할 수 있는 도구로도 제안되었다. 최근 브라질에서는 스마트폰을 이용

하여 인구조사를 실시할 예정이라고 발표하였다.[16] 조사원들의 현장에서 집계한 각종 수치를 실시간으로 사무실로 전송하여 자료 조사에 필요한 시간을 단축시킬 수 있을 것으로 전망하였으며, 스마트폰을 이용하여 전산 처리되는 만큼 수작업에서 발생되는 오류도 줄일 수 있을 것으로 예상하였다. 이처럼 스마트폰으로 조사원들을 실시간으로 관리하여 paradata(조사원 방문 횟수, 조사원 방문 시기 등)를 수집함으로써 비표집오차를 줄일 수 있을 것이며, 궁극적으로는 조사의 품질을 높일 수 있을 것으로 기대된다.

16) LG전자는 2010년 상반기에 브라질 정부산하 기관인 브라질 국립지리통계원(IBGE; Instituto Brasileiro Geografia Estatistica)에 스마트폰 15만 대를 공급하였다.

참고문헌

경인지방통계청, 2009. “정확한 연간조사 자료수집을 위한 사후점검 및 관리체계 개선 방안-2009년 지역별 고용조사를 중심으로-”, 자체보고서

동북지방통계청, 2009. “통계조사 대상처 응답 성향조사”, 자체보고서

충북통계사무소, 2008. “Rapport 형성 프로젝트”, 자체보고서

노규성 회장, 2010. “스마트폰 기반 모바일정부와 스마트행정 구현 방안”, 글로벌스마트오피스지원센터 한국디지털정책학회

Arceneaux, Taniecea A., 2007. "Evaluating the Computer Audio-Recorded Interviewing (CARI): Household Wellness Study (HWS) Field Test". Proceedings of the Survey Research Methods Section, American Statisical Association.

Couper, M. P. (1998) "Measuring survey quality in a CASIC environment." Joint Statistical Meetings of the American Statistical Association, Dallas, TX.

Hicks, Wendy., Brad Edwards, Karen Tourangeau, Laura Branden, Drew Kistler, Brett McBride, Lauren Harris-Kojetin and Abigail Moss, 2009. "A System Approach for Using CARI in Pretesting, Evaluation and Training" FedCasic Conference, Washington DC.

LaFlamme, François. (2009) "Overview of CATI Data Collection Research Focused on Developing Operational Strategies for Process Improvement". FedCasic Conference, Washington DC.

Lars Lyberg. (2009) "The Paradata Concept in Survey Research". NCRM Paradata Network, London.

Lupia, Arthur, Jon A. Krosnick, Pat Luevano, Matthew DeBell, and Darrell Donakowski, 2009. User’s Guide to the Advance Release of the ANES 2008 Time Series Study. Ann Arbor, MI and Palo Alto, CA: the University of Michigan and Stanford University.

Mabry, Patricia L. and G. Stephane Philogene, 2009. "Systems Science Methodologies To Protect and Improve Public Health". IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Nashville TN.

Maydan, Michael J. (2009) "Using Paradata to Monitor Survey Quality in Statistics Canada’s Regional Data Collection MIS Reports". FedCasic Conference, Washington DC.

Scheuren, Fritz. (2001) "Macro and Micro Paradata for Survey Assessment" 1999 NSAF Collection of Papers Washington, D.C.: The Urban Institute. Assessing the New Federalism Methodology Report No.7.

Taylor, Beth.L. 2009 "The 2006 National Health Interview Survey (NHIS) Paradata File: Overview and Application". FedCasic Conference, Washington DC.

Teichman, Ari. 2009. "Panda: Using Paradata to Improve Data Quality". FedCasic Conference, Washington DC.