단기훈련보고서

통계품질평가에 관한 연구

2002년 10월

통 계 청
통계주사 정 동 욱

차 례

▣ 해외훈련 개요 …… 2

▣ 훈련기관 개요 …… 3

▣ 본 론

Ⅰ. 개 요 …… 5

Ⅱ. 통계품질평가

1) 개 황 …… 7
2) 평가대상 및 기준 …… 8
3) 요 약 …… 11

Ⅲ. 통계조사의 품질향상을 위한 방안

가. Survey Quality …… 13
나. 조사표설계 …… 16

▣ 수집자료

1. Research on Survey Data Quality …… 27

2. Improving Quality of Survey …… 42

해 외 훈 련 개 요

1. 훈련국 : 미국

2. 훈련기관명 : 웨스텟 (WESTAT)

3. 훈련분야 : 통계품질평가에 관한 연구

4. 훈련기간 : 2002. 4. 6 ~ 2002. 10. 4

훈 련 기 관 개 요

WESTAT
1650 Research Blvd., Rockville, MD 20850 U.S.A
(301) 251-1500

Westat is an employee-owned research corporation serving agencies of the U.S. Government, as well as businesses, foundations, and state and local governments. In addition to our capabilities as a leading statistical survey research organization, Westat has developed skills and experience in custom research and program evaluation studies across a broad range of subject areas. Westat also has the technical expertise in survey and analytical methods, computer systems technology, biomedical science, and clinical trials to sustain a leadership position in all our research endeavors.

Westat's research, technical, and administrative staff of more than 1,500 is located at our headquarters in Rockville, Maryland, near Washington, DC. An additional 1,100 staff members are engaged in data collection and processing at Westat's survey processing facilities, at our

Telephone Research Center facilities, and throughout our nationwide field interviewing operations. Westat also maintains research offices near our clients in Bethesda, Maryland; Los Altos, California; Raleigh, North Carolina; Atlanta, Georgia; and Houston, Texas.

Demonstrating technical and managerial excellence since 1961, Westat has emerged as one of the foremost contract research organizations in the United States.

Westat Has Conducted Studies on a Diverse Range of Topics: health conditions and expenditures; academic achievement and literacy; medical treatments and outcomes; exposure assessments; program participation; employment and earnings; and respondent knowledge, attitudes, and behaviors. We combine the relevant program area expertise with the capabilities to perform major survey research projects: Study Design and analysis, Methodology, Survey Data Collection and Information Technology.

기관명	WESTAT (웨스텟)
소재지	- (주소) 1650 Research Blvd., Rockville, MD 20850 - (전화) 301-251-1500 - (인터넷) www.westat.com
연 혁	1961년 설립 이후 표본조사설계, 실지조사, 자료처리, 결과집계 등 통계조사 관련 업무를 수행하는 선두 연구업체로 자리잡고 있음
주요기능	- 조사기획 및 표본설계 - 자료처리를 위한 프로그래밍 - 실지조사 실시 - 결과자료를 이용 최종 보고서 작성
조 직	- Rockville 본부에는 1,500명의 직원이 있으며. 사무소에는 1,100명의 자료수집을 위한 직원이 있음 - 표본설계관련 통계Group, 자료처리Group, 기타 관리부서 등

Ⅰ. 개 요

지금은 정보화 시대라는 말을 많이 한다. 최근 몇 년 사이 인터넷보급, 무선통신 이용확대, PC성능 급성장 및 보급률 증가, Compact Disc의 보급 등은 정보의 이동속도와 이동량의 거대한 물결을 일으키는 21세기의 중요한 변화를 가져왔다.

통계조사결과에 대한 이용자의 요구 또한 정보화시대에 발맞추어 다양화 되어가고 있는 추세이다. 예를들면, 이용자에게 제공되는 형태(매체)의 다양화, 분석자료(통계표)의 다양한 형태 요구, 결과자료 및 메타정보에 대한 정보 요구 등이다. 이에따라, 통계조사 결과자료도『정확성』만을 의미하고 강조하던 면에서 이제는 하나의 제품으로 정도 즉, 품질(quality)을 평가하고 나아가 평가된 통계자료의 품질의 인증표시에 대한 필요성까지 논의하는 『통계품질평가에 관한 연구』가 시작되었다.

제품이란 수요와 공급의 논리에 의해 시장에서 평가된다. 통계조사 결과자료에 대한 평가도 예외는 아니다. 수요가 없는 제품은 생산자로 하여금 기획, 제조, 판매 등 총체적 부문에 대한 평가와 재정비 등 전략적 변화를 유발시킨다. 최근의 품질은 이용자(고객) 중심의 이용적합성과 품질의 다차원성을 강조하여 “명시적 혹은 묵시적 수요를 충족시키는 기능에 영향을 미치는 재화 및 용역의 총체적 특성”으로 정의되고 있다.

통계품질평가에 관한 연구는 통계조사결과의 이용확대와 보급을 위해서는 통계품질을 향상시키고 결과자료에 대한 품질을 이용자에게 공개함으로서 의사결정시 발생되는 오류를 최소화하고 통계시장의 활성화를 기하고자 한다.

그러나, 통계품질평가와 관련하여 용어정의, 평가범위 설정, 평가기준과 결과의 수치화 등 평가에 어려움이 있다.

그래서 나는 통계품질평가에 대하여 다음과 같이 기술하고자 한다.

1) 통계품질평가에 대한 개황 및 평가기준, 평가대상

2) 통계조사의 quality 향상

3) 조사표설계에 관한 사항 등을 통해,

주로 통계품질을 향상시키기 위한 부문에 대하여 언급하고자 한다.

Ⅱ. 통계품질평가

1) 개 황

조사통계(Survey Statistics)는 다양한 매체, 예를들면 보고서와 같은 인쇄물, 디스켓, CD-Rom, 인터넷을 통한 제공 등과 같이 이용자의 요구에 적합하도록 발간된다. 이와같은 통계자료는 의사결정(Decision Making)을 위해 자주 사용된다. 그러므로 자료의 부실은 결국 이용자의 의사결정시 중대한 오류를 범할 수 있게 할 수도 있다. 좋은 결과자료를 만들기 위해서는 많은 비용이 소요되며, 한정된 예산(Budget)내에서 생산된 자료를 이용하여 이용자가 의사결정시 오류를 최소화하기 위해 다음과 같은 사항들을 필수적으로 명시할 필요성이 대두되었다.

조사의 처음부터 마지막 단계까지에 대한 전반적인 내용을 명시하는 것이다.

① 조사대상에 대한 정의와 개념
② Sample Frame의 선택방법 및 배경
③ Sampling
④ Editing
⑤ Imputation 및 무응답(nonresponse)에 대한 처리방법 (appropriate adjustment)

공표된 수치만으로는 자료 전반에 대한 정도(Quality)나 자료의 수집 및 처리과정등에 대한 정보를 이용자는 전혀 알 수가 없다. 이에따라 조사결과자료의 생산자는 조사에 대한 주요한 특징을 명시할 의무가 있다. 이는 이용자가 자료를 평가하고 조사결과에 대한 해석을 하는데 있어 결정적인 오류를 최소화하기 위함이다. 또한 현대의 통계이용자는 자료분석 결과에 따른 의사결정시 오류를 줄이기 위해 상기정보 이외의 더 많은 정보(Bias, Variance등)를 얻으려 한다.

(1) 모집단의 정도와 유효한 domain에 대한 정보
(2) 조사항목에 대한 정의 및 내용
(3) Variance와 Bias에 대한 사항

분산(Variance)은 Total Variance를 말하며, Sampling error, Response error, nonresponse error, missing data에 의한 error, 자료처리과정에서 발생하는 editing, imputation error 등이 있다. Total Variance에 대한 좋은 추정치를 얻기 위해서는 많은 비용이 소요되어야 한다.

☞ MSE($\hat{\theta}$)= V($\hat{\theta}$)+ [B($\hat{\theta}$)] ^{2}, where $\hat{\theta}$ is an estimator the parameter θ

2) 평가대상 및 기준

주요통계작성 기관들은 통계자료 공표시 자료의 quality를 명시함과 동시에 명시되는 내용에 대해 중요성을 인식하고 있다. U.N 회원국들은 조사결과자료의 quality를 증진시키는데 앞장서 왔다.

이미 20년전인 지난 1982년 "Timeliness, Cost & Quality Attributes of Statistics"에 대한 비공식 모임이 the National statistical agencies in Canada, France, The Netherlands, Spain, Sweden, Switzerland, United Kingdom, International Labor Organization and the Economic Community 이 참여한 가운데 결성되었다. 이 모임의 목적은 "Guidelines for the presentation of the Quality of Statistics"에 대해서 토의하기 위해서이며, 그리고 영국, 캐나다, 스웨덴, 미국의 quality guideline을 상호 비교하여 개선하고자 함이었다.

다음과 같은 용어 "Error Profile documents"의 사용은 조사과정을 좀 더 구체적이고 체계적으로 언급하고 점검하고자 하는 것이었다. "The error profile documents"는 조사과정에서 발생한 오류의 내용으로서 만약 가능하다면 오류의 크기 또한 주어져야 한다. 그러나, 오류의 내용(효과, 영향)과 원인에 대한 정보를 이용할 수 없다면 이 사실을 알려주어야 한다. Error Profile documents는 이용자에게 통계의 한계에 대한 이해를 작성기관에게는 조사기획에 대한 재설계 및 보완을 통해 조사결과의 quality를 높일 수 있는 도구로 이용한다. U.S Department of Commerce 에서는 미국내 고용관련 통계에 중요한 자료를 제공하는 Current Population Survey의 경우 Error Profile document의 내용이 있다.

여기서 검토되어졌던 일종의 Check List는 다음과 같다.

<Error profile 작성을 위한 점검단계>

1. 표본설계 (Sampling design)

a. 표본틀 (Frame)
b. 표본추출 (Sample selection)
c. 중간점검 (품질관리,Quality Control of Sample Process)

2. 실사설계 (Observational design)

a. 자료수집과정 (Data collection procedure)
b. 조사표설계 (Questionnaire design)
c. 자료수집 staff (Data collection staff)
d. 조사원 교육 (Interviewer training)
e. 중간점검 (품질관리, Quality Control of field work)

3. 자료처리 설계 (Data preparation design)

a. 자료입력 (Data input operations)
b. 자료내 항목별 연관관계에 따른 오류 수정 및 Imputation (Cleaning, editing and imputation)
c. 중간점검 (품질관리, Quality control of data processing)

4. 추정 (Production of estimates)

a. 가중치 부여(Weighting procedure)
b. 추정 (Estimation procedure)
c. 중간점검 (품질관리, Quality control of estimation procedure)

5. 분석 및 공표 (Analysis and publication)

1982년 The Conference of European Statistics 모임에서는 이용자를 위한 자료의 quality를 명시하는 방법에 대해 논의했는데, 다음과 같은 항목은 최소한 언급되어야 한다고 결론지었다.

a. data에 대한 출처 및 개황 정보
b. data 포괄범위 (예, 틀(frame)의 타당 및 적합성 여부)
c. 추출방법과 추출단위 등에 대한 정보
d. 응답률
e. sampling error 와 standard error 의 계산방법과 해석에 대한 정보
f. data의 크기, 주요 오류의 경향(유형) 및 오류로 인해 결과자료에 미칠 영향과 상대적인 중요성
g. 결과자료 작성 과정상 중대한 변화(변경) 내용과 과거자료와의 비교시 유의할 사항(항목 및 요인들)
h. 다른 자료와 비교시 유의사항 및 비교에 관련된 참고정보
i. 이용자에게 유용한 기타 자세한 기술적인 설명 및 참고자료

즉, 다음과 같은 점에 해당되는 정보를 제공해야 한다는 것이다.

① Target Population and definition
② Frame에 대한 quality
③ 응답률(response rate)과 sampling error
④ 비교성 (지난자료와 비교가능여부 및 유사통계 결과자료를 이용 상호 비교가능 여부)

토론의 결실로 일부국가의 경우 통계작성기관들이 “자료 품질에 대한정보정책”을 발표했다. 스웨덴의 경우 이미 1983년 “올바른 공표통계에 대한 해석을 위한 이용자의 요구”에 대한 정책(방침)을 채택했으며, 주요 골자는 다음과 같다.

『통계자료 생산자는 이용자에게 공표통계에 대한 정확한 해석을 위한 중요하고 분명한 요인들을 알려주어야 한다. 그 정보는 매우 쉽게 이해 할 수 있고, 이용자의 요구를 충족시킬 수 있어야 한다.』

위에서 기술된 바와 같이 통계선진국들은 통계품질 평가 및 관리에 대한 중요성을 인식하여 정책적으로 실시하고 있다.

3) 요 약

통계품질과 관련하여 이를 다시 요약하면, 품질관리의 주요 요소는 첫째로, 이용자의 요구를 파악하여 이를 통계조사의 기획, 실사, 자료처리, 결과집계 등 해당되는 단계에 반영하는 것이다. 둘째는, 품질관리를 위한 기준설정, 점검 대상정의, 평가결과 분석을 위한 계량화이며 끝으로, 이용자에게 제공된 자료에 대한 이용자의 재평가 및 개선방안 등을 수집하여 지속적으로 개선해 나아가는 것이 중요하다.

통계의 품질 평가는 지난 수십년간 단순히 정확성만을 최고 평가기준으로 여겨오면서, 단 1%의 실수도 인정치 않았다. 그러나 이제는 통계도 하나의 제품으로 인식됨과 동시에 단순히 정확성측면 뿐만 아니라 이용자에게 얼마나 적합하게 작성되고 제공되었는지에 대한 개념이다.

여기에는 ①통계개념의 **관련성(Relevance)**으로 통계조사가 조사목적과 이용자의 요구에 부합되는 정도를 나타내며, ② **정확성 (Accuracy)**은 조사결과가 실제사실을 반영하는 정도, ③ **시의성 (Timeliness)** 통계의 신속성 및 공표의 사전예고와 준수정도, ④ 정보의 **접근성 (Accessibility)** 이용자의 자료이용 및 통계해석의 편의성, ⑤ **일치성 (Coherence)** 통계간 비교 가능성 및 일치되는 일관성 확보여부, ⑥ 결과의 객관성 (Objectivity), ⑦ 명확성 (Clarity)으로 quality에 대한 평가방향이 다양화되고 있다.

통계품질을 평가하고 그 결과를 이용자에게 사실대로 알려주는 것은 최종 수요자이자 평가자인 고객이 의사결정을 할 때 제공받은 자료의 품질을 고려하여 이용할 수 있도록 하고자 한다.

통계자료는 100% 정확성을 갖고 있어야 한다는 그릇된 고정관념에서 벗어나 품질을 보고 이용자가 선택하여 의사결정을 하는데 이용되어야 할 것이다.

Ⅲ. 통계조사의 품질향상을 위한 방안

가. Survey Quality

1) Survey Quality 향상과 사회여건 변화

미국에서는 여러가지 조사방법중 특히 표본크기가 작고, 조사범위가 방대한 경우에 전화를 이용한 『Telephone survey』를 조사방법으로 이용하여 지난 10년간 조사를 실시하였다. 그러나 항상 사회변화 형태에 따라 Survey quality를 control 하는 방법도 변화해 왔다.

최근 통계조사에서 Survey quality를
높일수 있는 요인은 어떤 것들이 있을까?

① 응답률
② 기술의 변화
(Internet보급, Mobile Phone이용확대, Handhold computer보급)
③ Communications (내부 직원간의 의사전달, 조사원과 응답자 등)

응답률 저하의 주요 요인으로는 응답자들의 생활여건 변화에 있다. 특히 낮에는 빈집이 늘고 저녁 늦은시간과 주말에만 방문이 가능한 가구의 증가, 맞벌이 가구의 증가 등 응답률과 응답자의 생활여건은 밀접한 관계가 있다. 또한 통신 부가서비스 확충에 따라 사생활이 더욱더 보호되면서 응답자와 연결되기도 전에 응답을 거부당하는 사례도 있으며, 휴대폰 보급의 급증도 응답률과 밀접한 관련이 있다. 이에따라 이런 생활여건을 잘 반영하여 조사를 실시할 필요가 있다. 다음은 응답률 저하와 관련된 여건 변화의 유형별 예시이다.

<응답률 저하와 관련된 사회변화>

① 성인가족 모두가 일한다. (맞벌이 가정과 같은 경우)
- 낮에 모두 일을 하기 때문에 가구원들이 집에 머무르는 시간이 적다.
- 가령 집에 가구원이 퇴근하여 머무르고 있어도 너무 가정일로 바빠서 조사표에 대해 질문이 어렵다.

② 통신업체와의 문제
- 전화번호, 기본가구정보 등 List에 대해 통신업체로부터 제공거절
- 장난전화 자동제어와 같은 Telezappers와 유사한 기술 보급
- Caller ID 서비스 보급에 따라 전화 건 사람을 미리 알고 전화를 받지않는 방법, 즉 screening에 따라 응답률이 낮아진다.

③ 전화번호 List에 등재되지 않은 가구의 처리문제

④ 가구의 대표전화가 유선전화에서 Mobile phone으로 변화하는 추세

2) 통계조사 환경 변화에 따른 대처방안 (예시)

응답률 저하와 관련된 사회변화와 관련하여 응답률을 높이기 위해 다음과 같은 노력을 시도했다.

① 응답자와 통화가 가능할 때까지 계속해서 접촉을 시도한다.
(심지어 7회이상 시도하는 것이 보통이 되어 버렸다.)

② 낮시간보다 저녁시간과 주말을 이용하여 응답자와 접촉을 시도한다.
이와같은 방법은 최초접촉 성공률을 높이기 위함이다.

③ 조사방법의 혼용을 통해 응답자의 특성에 따라 유연하게 대처함으로써 응답거부를 최소화하고자 함

- Telephone Survey & face-to-face
- face-to-face & mailing Survey
- Telephone Survey & mailing Survey

④ 숙련된 Staff를 투입하여 응답거부를 최소화한다.

⑤ 응답자에게 조사의 중요성 등을 다시한번 설명하여 조사에 적극 참여토록 유도한다.

⑥ 성실히 응답한 응답자에 대한 재정 및 기타 Incentives를 부여방안 검토

■ Survey quality에 영향을 미치는 요인에는 어떤 것들이 있는가?

① 조사를 위한 투입비용

② 조사과정에서 발생하는 무응답, 이를 조정(Imputation or adjustment) 하는 방법에 대한 연구

결국 Survey quality에 영향을 미치는 요인이나 사회여건 변화 등에 대한 분석은 어떻게 Survey Quality를 높일 것인가와 무엇으로 (어떤기준)으로 이를 측정할 것인가이다.

먼저, Survey Quality를 평가하는 기준으로는 Statistic Canada, EFQM, ABS에서 1999~2000년에 사용한 것으로 다음과 같다.

① Relevance ② Accuracy ③ Timeliness ④ Accessibility ⑤ Coherence ⑥ Objectivity ⑦ Clarity

다음은 quality를 높이기 위한 방법중 ① Staff의 시간도 조사비용에 포함되어야 하며, ② 의사교환시 간결성과 정확성 확보를 위해 노력해야 한다. 우선 내부 의사교환으로 Programmer & Statistician 그리고 Project staff 상호간에 정확한 의사교환과 외부 의사교환의 형태로 조사원과 응답자간의 의사교환, Staff와 조사원과의 의사교환 등이다. ③ 매 단계마다 quality Control을 실시하여 단계별 재 검토를 실시하여야 한다.

나. 조사표설계 (Questionnaire Design)

1) 조사(연구)대상의 정의와 조사표설계의 연관성

조사표설계에 대한 언급전에 조사의 기획단계에서 분석 및 공표단계까지 흐름을 통해 일반적인 사항을 살펴보고자 한다.

첫 번째는 조사(연구)개요에 대한 정의이다. 이는 조사항목과 밀접한 관계가 있음으로 다음과 같은 사항에 대해 특히 명확히 명시되어야 한다.

- 이 조사가 어떤 연구를 위해 실시되는가? (어떤 문제를 해결하기 위해 이 조사를 실시하는가?)
  - ☞ 이를 위해서는 조사의 주목적(Main goal) 과 기타의 목적들도 자세하게 정의되어져야 한다.
- 조사목적을 충족시키기 위해서(문제를 해결하기 위해서)는 어떤 새로운 정보들이 필요한가?

두 번째는 자료수집계획(Planning the Data Collection)을 수립하는 것인데, 다음사항에 대한 신중한 검토가 필요하다.

- 자료 수집방법은 무엇인가?
  - ☞ face-to-face, mail, computer-administered, telephone, Internet 중 조사목적과 조사 항목에 따라 결정한다. 각 조사방법별 장단점은 <표>를 참조한다.
- 수집된 자료의 자료처리 가능여부 점검 및 응답자가 응답시 혼돈하지 않도록 정확한 개념을 명시하는 것이다. 이는 조사표 설계시 선택형과 개방형 항목을 적절히 활용하기 위함이며, 컴퓨터를 이용 집계 및 분석이 불가능한 항목은 무의미해 짐에 따라 신중한 검토와 결정이 필요하다.
- Pre-test를 위한 표본추출

- Pre-testing
- 조사를 위한 표본크기 결정
- 본조사 실시

<조사방법간 장·단점(Mode of Administration)>

	Personal	Telephone	Mail
Sensory Channel	Auditive	Auditive	Visual
Time Pressure	Medium	High	Low
Additional explanations possible?	Yes	Yes	No
Interview length	Longest	Shortest	Medium
Establishing Report	Easiest	Medium	Difficult
Item-Nonresponse	Lowest	Medium	Highest

세 번째는 자료입력 및 처리과정(Data Entry & Data Processing)이다. 이 과정에서는 전산을 자료처리를 위해 coding과 오류 점검 등이 이루어진다.

- 개방형 항목(Open Question)에 대한 Coding
- 조사단위별 식별을 위한 ID점검
- 자료점검 (Data Cleaning)
    - Missing Values 점검
    - Checking Filters
    - Formal inconsistencies
    - Content inconsistencies

마지막으로 자료분석이다. 이 단계는 연구목적에 따라 SAS, SPSS 등과 같은 통계패키지를 이용하여 분석표를 생성시키고 이를 해석하여 이용자에게 (혹은 의뢰인)에게 제공하는 것이다.

위에서 언급된 바와 같이 연구대상에 대한 정의는 조사표설계 및 자료해석과 밀접히 관련되어 있다는 것이다.

2) 응답자의 응답과정

응답자가 조사시 응답하는 과정을 살펴보면 크게 4가지 정도로 나뉜다.
『질문의 의미이해』→ 『질문과 관련된 정보를 생각』→『응답에 대한 적절한 답을 찾음』→『응답한 내용을 다시 한번 생각 및 수정』

첫 번째로 질문내용을 이해하는 것이다. 여기에서 중요한 것은 응답자가 연구자(혹은 조사기획자)가 의도하는 방향과 일치되어야 한다. 응답자가 질문을 바라보는 측면은 두가지가 있는데,

- 질문의 문장상 의미
- 실질적인 연구자의 목적이 담겨있는 의미이다.

결국 질문의 문장상 의미보다는 연구자의 의도가 잘 드러나도록 조사표 설계시 고려되어야 한다.

두 번째는 응답에 필요한 정보를 지난 기억으로부터 가져와야 하는 경우, 관련된 모든 정보를 기억하여 응답한다는 것은 거의 어려운 해 낸다는 것은 어렵다. 이 경우 응답자는 기억속에서 접근이 가능한 정보안에서 응답에 필요한 판단을 내리기 위해 생각한다. 비록 응답을 위한 기억이 전부가 아닐지라도 응답에 필요한 충분한 정보라고 생각하는 범위내에서 응답을 위한 정보를 찾는다.

셋째는 두 번째 단계에서 결정한 응답내용을 조사표에 제공된 응답범위 또는 선택내용에 맞도록 기입(응답)을 한다.

마지막으로는 응답자는 이미 기입(응답)한 내용을 수정하려고 할지도 모른다. 특히, 개인적인 것과 사회적인 문제등과 같이 개인적인 판단(의견)을 묻는 질문은 최종제출(응답)에 앞서 판단을 변경할 수도 있을 것이다.

3) 어떤 질문이 좋은 질문인가?

좋은 질문이 되기 위해서는,

① 질문 내용이 의도한 대로 응답자에게 이해(전달)되어 질 수 있는가?

② 질문내용을 모든 응답자들이 동일하게 해석하고 이해하고 있는가?

③ 모든 응답자가 응답할 수 있는 질문내용인가?

④ 전후 질문에 의해 영향을 받는가?

상기와 같은 내용에서 명확성, 객관성, 보편성 등이 확보되어야 한다.

즉, 표본으로 추출된 모든 응답자가 동일하게 질문을 이해할 수 있어야 하며, 응답자의 학력, 성별, 사회적지위 및 경제력 등에 관계없이 쉬운 질문이어야 하며, 모든 질문들이 독립적으로 응답되어 질 수 있도록 설계되어야 한다. 좋은 질문만이 각각의 응답자로 부터 신뢰할 수 있는 정보를 얻을 수 있기 때문이다.

그러면 간단한 예를 통해 좋은 질문에 대해서 살펴보도록 한다.

> (예제) 당신은 얼마나 자주 미사에 참석합니까?
> ① 규칙적으로 ② 때때로 ③ 거의 가끔 ④ 전혀 참석치 않음

위 예제를 좋은 질문의 조건과 비교하여 보면,

첫 번째 문제점은 『①질문의 객관성이 없다는 것』이다. "***자주***" 라는 의미에서 ***얼마나***, ***어느 기간동안*** 이 구체적으로 명시되지 않았으며, "참석"이라는 의미에서 또한 응답자마다 판단과 이해를 달리할 가능성이 있다.

두 번째 문제점은 『②응답범위의 모호성』이다. 이는 응답자가 정한 빈도를 ①~④중 선택할 때 오차를 발생할 수 있으며, 자료 분석시 빈도를 수치화하는 것이 결과 해석시 더욱더 명확성을 부여할 것이다. 그러나 이 경우는 연구(조사)목적에 따라 표현해야 함에 유의하여야 한다.

그래서 이 질문을 다음과 같이 변경해 볼 수 있다.

(질문) 당신은 지난 6개월 동안 (2002.1월~6월) 미사에 몇 번이나 참석 했습니까?
(선택1) ① 24회 이상 ② 12~18회 ③ 6~12회 ④ 1~6회 ⑤ 0회
(선택2) ( ) 회

위 예제에서는 선택1과 같이 ①~⑤에서 선택하는 것이며, 선택2는 개방형으로 응답자가 직접 횟수를 기입하도록 하는 방법이다. 이는 연구(조사) 목적에 따라 응답자의 응답부담 경감이나 정확한 횟수를 산출해야 하는 경우 등을 고려하여 설계한다.

4) 이해와 해석이 용이한 질문이란?

질문이 내용이 다음과 같을 때 응답자로부터 더 좋은 응답을 얻을 수 있다. 질문내에 사용되는 단어는 모호하거나 익숙치 않은 단어는 사용하지 말아야 한다. 이는 단어를 통해 의미가 와전될 가능성이 있기 때문이다. 응답자가 이해한 질문의 의미는 연구자(기획자)의 의도와 일치되어야 한다. 이는 본조사에 앞서 Pre-Test를 통해 질문이 응답자에 따라 어떻게 받아들여지고 있는지 확인해야 한다.

응답자가 design된 질문을 이해하는 과정에서 발생되는 문제는 대체적으로 질문에서 ⅰ) 의미의 모호성 ⅱ) 문장의 복잡성과 모호한 표현 등으로부터 비롯된다.

A. 의미의 모호성(Semantic Ambiguity)

질문에 사용되는 단어는 지역적, 문화적, 교육정도, 성별, 나이 등에 따라 서로 다른 의미를 갖거나, 범주가 달라서는 안 된다는 것이다. 이와같은 문제점은 Pre-Test를 통해 의미상 모호성 여부를 시험조사 해야 한다.

예를 들어, “soda”라는 단어를 질문내에 사용한 경우, 지역에 따라 “coke”을 “soda”라고 하기도 하며, 탄산음료 전체를 “soda”라고 말하기도 한다.

☞ 그래서 “soda”와 같은 단어보다는 “coke”과 같이 명확한 단어를 사용해야 한다.

B. 문장의 복잡성 (Syntactic Complexity)

질문내용은 응답자에게 명확하게 본래의 의미가 전달될 수 있어야 한다. 그러므로 은유적인 표현(간접표현)이나, 이중부정과 같이 여러번 문장을 읽어야 하는 경우는 응답자에 따라 서로 다른 해석을 할 수 있음으로 문장은 간결하고 직설적인 표현이 문장의 간결성 확보 및 모호성 배제에 도움이 될 것이다.

5) 질문작성 규칙

조사표 항목 설계시 질문내용이 다음과 같은 규칙에 의해 작성되었는지 참고하여야 한다.

첫째, 단어는 가급적이면 자주 많이 평상시에 사용되는 것을 사용한다. 만약 익숙치 않은 단어를 꼭 사용해야 하는 경우 이에대한 정의를 포함시켜야 하며, 그 단어에 대한 정의는 질문전에 설명되어져야 한다.

둘째, 지역적, 문화적 다양성에 따른 질문의 의미와 나이 및 성별에 따라 의미가 동일하게 전달되어야 한다.

셋째, 모호한 표현을 사용하지 말아야 한다. 만약 필요하다면 그것에 대한 정의를 언급해야 한다.

넷째, 질문들의 마지막에 응답조건(response option)을 서술한다. 예를들어 “*매우좋다, 보통이다, 좋지않다* **중에서 내년중에 다른집으로**

이사하는 것에 대한 당신의 생각을 답해주세요" 이 질문은 문두에 선택해야 할 response option들이 나와있다. 무엇에 대한 response option인지는 문장을 다 읽은 후에 알 수 있다. 응답자는 다시 response option을 읽고 의견을 말해야 함으로, 좋은 형태의 질문은 아니다. 이 경우 response option을 맨 뒤에 서술하면 좀 더 명확한 표현이 된다.

☞ **"내년중에 당신이 다른집으로 이사하다면, 당신의 생각을** *매우좋다, 보통이다, 좋지않다* **중에서 답해주세요"**

다섯 번째, 연구(기획)자가 원하는 응답형태로 답해지도록 명확하게 응답자와 의사소통이 되도록 질문을 작성해야 한다.

예를들면 **"당신은 언제 대전광역시에 이사왔습니까?"**라고 질문한 경우 응답은 ① 13살 때 ② 결혼한 다음해 ③ 1998년도 등과 같이 다양한 형태로 이루어 질 수 있다. 그러므로 질문을 좀 더 응답자와 명확히 의사소통이 되도록 다음과 같이 수정되어야 한다.

☞ **"당신은 몇 년도에 대전광역시로 이사왔습니까?"** 이 경우의 응답은 위의 ③과 같은 형태로만 응답되어진다.

여섯 번째, 한 문장 안에는 한 개의 질문만 있어야 한다.

예를들면 **"당신은 여자와 아이들이 독감주사를 매년 맞아야 한다고 생각합니까?** 이경우도 간단한 Yes/No question이 아니다.

일곱 번째, 문장내 이중부정은 피한다.

강한 긍정을 나타내기 위해 이중부정을 사용하기도 하나, 말하는 사람의 어조 등에 따라 의미가 잘못 전달될 수도 있기 때문에 이중부정의 사용은 피한다. 좀더 직설적으로 명확히 표현해야 한다.

여덟 번째, 이미 앞에서 언급되었다하여 다음 질문에서 요약된 형태의 질문은 피한다. 응답자는 queue와 같이 이미 지나간 질문은 잊어버렸다고 간주하여 다음 질문도 항상 자세히 명료하게 설명되어질 필요가 있다.

아홉 번째, 질문이 긴 경우 가급적이면 간단해 지도록 partition한다.
끝으로, 연구(조사) 목적에 반드시 필요한 항목만 응답자에게 묻는다.

6) 조사표의 시각적 효과증대 및 이에따른 응답자의 태도

조사표는 모든 응답자가 동일한 과정(순서)에 의해 읽고, 모든 page의 질문에 응답할 수 있도록 설계되어야 한다. 그래서 지금부터는 항목의 배열, 글자의 크기 및 모양 등 시각적 효과를 높임으로서 응답자의 태도와 어떤 관련이 있는지 살펴보도록 한다.

항목배열 순서는 응답 결과에 영향을 준다. 이는 응답자가 해당 질문을 이해하고 → 답을 기억하고 → 답을 결정하는 과정에서, 이전에 질문된 내용에 따라 기억된 내용이 그 다음 답을 결정하는데 많은 영향을 줄 수 있다. 이에따라 pre-test를 통해 항목배열에 따라 조사결과가 어떻게 바뀌는지 반드시 확인한 후 항목배열을 결정해야 한다. 특히, 민감한 사항(sensitive topics)에 대한 조사에서는 좋은 결과를 얻기 위해서는 더욱더 신중해야 하며, 다음과 같은 사항을 참고하여 작성한다.

민감한 사항(sensitive topics)에 대한 조사는 먼저, 연구(조사)의 목적을 충분히 설명한 후 응답자의 비밀보호에 대한 사항을 충분히 주지시켜야 한다. 그리고 응답자와 응답의 정확성이 가져다 주는 효과를 설명하여 응답자로 하여금 성실히 응답하도록 유도한다. 다음은 항목배열시 민감한 질문은 후반부에 배치하고, open question에서 close question 등이다. 그러나 무엇보다 중요한 것은 아무리 응답자 비밀보호와 적절한 항목배치를 한다 하더라도 응답자의 원초적인 불신을 제거하기는 어렵다. 따라서 요즈음은 다음과 같은 조사방법을 이용하여 sensitive topics에 대한 조사를 실시하고 있다.

① 자계식 조사표 ② 우편조사 ③ CAPI ④ Walkman ⑤ Diaries

조사표의 layout은 다음과 두가지 측면에서 고려되어야 한다.

첫 번째는 응답자들이 기대하는 항목배열순서, 즉 항목배열의 연관성 및 일관성과 symbol 사용되는 특수문자 등의 정보를 파악하여 반영한다.

두 번째는 응답자들이 진행해 나아갈 수 있도록 조사표에 이정표와 같은 진행유도이다. 이는 시각적인 신호를 통해 진행순서를 알 수 있도록 하는 것이다.

어떤 것들이 응답자에게 응답여부에 대해 영향을 미치는지 알아보도록 한다.

a. 조사표 종이의 색깔은 응답률(response rate)과 관련이 적다.
b. 글자색과 밝기는 무응답(non-response)과 깊은 관련이 있다.
c. 항목과 관련된 설명을 항목번호 밖의 지면에 서술하면 응답자는 동 내용을 잘 읽지 않는 경향이 있다.
d. 세부적인 정의는 answer box뒤에 위치시키는 것이 더 효과적이다.
e. 가로세로로 설계된 형태 (matrices)는 이해하기 어려운 형태이다.

조사표설계에 있어 중요한 3가지 단계가 있다.

Step1은 조사표의 각 page에 명시된 모든 정보 특히 지시문 등을 읽기 위한 바람직한 진행순서를 정하는 것이며,

Step2는 Step1에서 정해진 순서에 따라 정보를 찾아 갈 수 있도록 시각적인 도구에 의해 유도하는 방법을 개발하는 것과,

Step3는 항목을 skip 하거나 전 항목의 응답결과에 따라서 더 이상 응답할 필요가 없는 경우 등 항목의 진행을 유도할 수 있는 시각적인 유도방법을 추가적으로 개발하는 것이다.

Step2 와 Step3는 다음과 같은 방법에 의해서 다루어 질 수 있다.

a. visual elements의 글자색깔, 밝기, 크기를 조정한다.
b. visual elements 위치와 공간의 일관성을 확보하여 응답자가 인지할 수 있도록 한다.
c. 특정부분의 강화를 위해서는 간결성(simplicity), 규칙성(regularity), 조화성(symmetry)을 강조하여 응답자로 하여금 강조하는 부분을 인지하고 예측 가능하도록 하여 시각적 효과를 높일 수 있다.

7) 조사표설계에 관한 기본원칙 (요약)

조사표설계는 응답자에게 시각적 효과와 응답부담을 최소화하여 최대의 정보를 얻을 수 있는 도구이다. 즉, 통계품질을 높일 수 있는 단계중의 하나라고 생각한다. 이에따라 조사표설계와 관련하여 체크해 보아야 할 사항을 아래와 같이 요약한다.

원칙1) 응답자가 질문에 답하기 위해서 각각의 질문을 여러번 반복해서 읽는 것을 최소화하도록 질문을 작성해야 한다
원칙2) 지시문은 어느 내용(항목)과 관련이 있는지 정확히 응답자에게 알려지도록 작성되어야 하며, 조사표 처음부분에 명시해서는 안 된다.
원칙3) 한번에 하나씩만 질문해야 한다.
원칙4) 가로, 세로를 혼용한 구조의 질문은 가급적 자제한다.
원칙5) 크고 밝은 symbol들을 이용하여 각각의 시작페이지를 알리는데 사용하는 것이 시각적 효과를 높일 수 있다.
원칙6) 항목번호는 처음부터 마지막까지 연속적으로 단순하게 부여되어야 한다.
원칙7) 응답자가 항목을 쉽게 식별할 수 있도록 각각을 동일한 방법(형태)로 일치시켜야 한다.

원칙8) 질문과 answer choice는 명암의 차이를 이용하여 구분되도록 한다.
(*dark print* for ***questions*** and *light print* for ***answer choices***)

원칙9) 만약, 특별한 지시문을 반드시 추가해야 하는 경우는 질문내용 부분에 함께 연결하여 강조할 수도 있다.

원칙10) 선택적으로 경우에 따라 사용되는 지시문의 경우 반드시 질문 내용부분에서 글자크기 및 모양과 symbol등으로 분리될 수 있도록 하여야 한다.

원칙11) 특별히 응답자에게 효과적인 지시나 안내가 필요한 지시문은 음영이 있는 밝은 색상을 이용하는 것도 효과적이다.

원칙12) Answer choices는 가로로 연결하여 나열하는 것보다 세로로 각각의 Answer choices를 열거하는 것이 효과적이다.

원칙13) Answer space는 모든 항목에서 일관성을 갖고 있어야 한다. 예를들면, 왼쪽에 1항의 답을 기입했으면 다른 항목들도 왼쪽에 답을 기입하는 answer space가 있어야 한다.

원칙14) 질문에 대한 답을 위해서 번호를 이용하거나 □(answer box) 등을 이용한다.

원칙15) Answer choice 간에 서로 중복되거나 내용이 중복되지 않도록 하여야 한다.

원칙16) 항목배열은 쉽고 일반적인 사실에 대한 질문부터 어렵고 까다로운(전문적인) 질문순서로 구성하며, 가장 주요한 항목은 중간 위치에 배치하는 것이 효과적이다.

원칙17) 항목은 유사성이 있는 것들끼리 모여있도록 grouping하는 것이 조사시 유리하다.

끝으로 연관관계 파악을 위한 항목등을 추가하는 것도 중요하나, 반드시 연구(조사) 목적에 필요한 항목만 선별하여 응답부담을 경감시키고 간결한 조사표를 설계하는 것이 무엇보다 중요하다.

【붙임1】

Research on Survey Data Quality

Robert M. Groves

A. Survey Research as a Methodology Without a Unifying Theory

Survey research is not itself an academic discipline, with a common language, a common set of principles for evaluating new ideas, and a well-organized professional reference group. Lacking such an organization, the field of survey research has evolved through the somewhat independent and uncoordinated contributions of researchers trained as statisticians, psychologists, political scientists, and sociologists. These brief encounters between the survey method and bodies of theory have produced what we know about survey quality today.

Such a *melange* of workers certainly breeds innovation, but it also spawns applications of the method for radically different purpose, suffers severe problems of communication, and produces disagreements about the importance of various components of quality. The status quo in survey research can be described as a set of role pairs, the members of each pair oppositional in their practices with regard to some survey design feature. There are data collectors, who implement surveys, and analysts, who study substantive issues using data. There are those who use surveys to describe populations(describers) and those who test causal theories using survey data(modelers). There are the ***measurers*** who try to build empirical estimates of survey error and the ***reducers*** who try to eliminate survey error. The survey methodology literature is filled with articles by representatives of these groups, which concentrate on their favored use of surveys or their "error of choice" and who ignore the concerns of the others. They rarely confront one another, because they can retreat to their individual disciplines for reinforcement of their viewpoints. This article examines how reducers and measurers and describers and modelers approach different error sources. Structuring the review of research on survey data quality in this way helps to understand why different areas of research are chosen by different investigators.

Because of fundamental discrepancies in views about the nature of the measurement process, the group employ competing language of survey quality and survey error. Survey statistics most commonly views total error as the expected squared

difference between a sample statistics (e.g., the mean value of a variable measured on sample respondents) and attribute in the entire target population). The ***mean square error*** is the label given to this statistical concept. In contrast, psychologists who use survey data tend to focus on errors in measures on individuals, using the notions of ***validity*** and ***reliability*** as key concepts of quality.

Many survey researchers borrow their concepts of quality from survey statistics and speak in terms of ***bias*** and ***variance***. Bias denotes a fixed (over replications) departure from some underlying true value for the statistics (e.g., we say a survey estimate of the mean number of years of education is biased if it is above or below the true target population mean value). Variance or variable error is used to refer to departures from the true that change direction or magnitude across different replications. One key term in both these definitions is replication. This word is important because different groups of survey researchers mentioned above use the term differently. To those interested in sampling error alone, replication means a different implementation of the survey using a different sample drawn in the same manner, and bias means that, on replication, the same departure from true value would occur. In contrast, for many analysts who focus their attention on measurement error, a bias in response (e.g., subtracting 5 years from the report of one's age) refers to a constant departure over repeated administration to the same sample. Further, to those focusing on interviewer variance, the errors of interest are the variations in results that might have been obtained if a different set of interviewers had done the work. In short, what constitutes a replication changes from one perspective to another.

Most of the conceptual differences among survey researchers stem from differences in which features of a survey are considered fixed and which are considered variable over replications. For example, most survey practice data as of they remain unchanged if different interviewers had administered the survey. That is, the analyst assumes that all possibles interviewers would obtain the same results from the chosen sample. When Kish(1978) describes the process of "bringing errors into the design,: he refers to explicit design features that permit measurement of variable or fixed errors. But if these errors are not conceptualized, they cannot be introduced into the design or measured. Further, much attention to errors fails to estimate their magnitude. For example, reducers often try to eliminate errors, not measure them. They study methods of training interviews to improve survey quality; they search for questioning protocols to improve recall; they construct methods of improving response rates. By using the single method they jidge is better than all others, they tend to focus on biases, errors that would be fixed over replications. In contrast, measures depend on variations internal to the design to estimate error

magnitudes. For that reason, they tend to focus on variable errors, variance terms, since they are easier to estimate with the survey data themselves. In contrast to the reducers they focus on differences among interviewers as sources of errors, error variance associated with different indicators of the same concept, and variation in compliance likelihood across sample persons.

From the perspective of those who think in terms of bias and variance to describe survey quality, there are three major sources of errors due to nonobservation : (1) coverage error (2) nonresponse error and (3) sampling error. There are four potential sources of errors of observation or measurement error: (1) the interviewer (2) the respondent (3) the questionnaire and (4) the mode of interview (e.g., face-to-face, telephone, or self-administered). Other source of measurement error have been identified but frequently studied (e.g., the effect of the presence of others, joint effects of mode question wording).

The major competitor to languages of quality using concept of bias and variance is that arising from psychometric theory. The labels for error in this tradition are ***validity*** and ***reliability.*** Although in causal conversation validity might be equated with unbiasedness and reliability with low variance, most careful definitions of the terms illustrate their different meanings. Validity is most often defined as a correlation between a measure and the true the value of the attribute, taken on a set of individuals (Lord and Norvick, 1968). It differs from bias in that it is defined in terms of individual measurements; it is not necessarily a leads to the possibility of a perfectly valid measure of some entity (i.e., correlation of 1.0 with the individual true values) having large bias (e.g., overestimating a mean on the entity). Auxiliary types of validity have been coined by many psychologists (e.g., predictive validity, Lord and Novick, 1968; statistical conclusion validity, Campbell and Stanley, 1963). Indeed, the adjectives applied to the term ***validity*** have so proliferated, that by itself it has little meaning. Reliability has been less often a target of innovation in usage. It typically refers to correlations over replications of measurements on a set of individuals and is often estimated by correlations between initial and second measurements. (i.e., test-retest correlation).

In addition to the first language's focus on errors in statistics (i.e., bias and variance) versus errors in data provided by individuals, it also tends to give more attention to error in descriptive measures (e.g., means and proportions) than to those in analytic statistics (e.g., regression coefficients in structural models). Most describers use surveys to estimate means, proportions and totals and they tends to use ***bias*** and ***variance*** to refer to errors. Most modelers use surveys to estimate regression coefficients, coefficients and parameters of causal models and they tend to

use error concepts related to reliability and validity.

B. Components of Survey Quality

Despite the lack of full understanding of survey quality or even agreement on a language of errors, there have been rich developments in survey methods over the last 50 years. These developments have affected errors of nonobservation as well as measurement errors.

Errors due to nonobservation(noncoverage, nonresponse, and sampling) are most salient to those for whom external inference to a clearly defined population is important. These tend to be describers, and hence, for example, researchers in government agencies often devote more effort to coverage and nonresponse error than do others.

Coverage error refers to the discrepancy between sample survey results and the results of a full enumeration of the population under study which arises because some members of the population are not covered by the sampling frame. ***Nonresponse errors*** include all discrepancies between the population characteristics and those estimated from a sample survey which arise because some members of the sample were not measured in the survey.

Sampling errors are discrepancies between population characteristics and those estimated from a sample survey which arise because some members of the population were deliberately excluded from the survey measurement through selection of a subset. All three of these errors arise because some part of the population was not measured(i.e., their data are missing from calculations using the survey data).

B-1. Coverage Error

Coverage error is the forgotten child among the family of errors to which surveys are subject. Indeed, many modelers ignore coverage error entirely. They place their emphasis on the specification of a multivariate relationship and not on the estimation of population characteristics. Their concerns with regard to error are whether the model they are investigating has the correct functional form and whether any causal variables have been omitted from the model. In a sense, they seek evidence that the model will hold in some group, regardless of composition. Psychologists do use the term ***external validity*** to reflect their desire to go beyond the set of subjects measured. However, ***external validity*** involves considerations of all three errors of nonobservation, coverage, nonresponse, and sampling errors.

What we know about household survey coverage errors comes from special

studies mounted to observe the kinds of persons missed in traditional operations. These include early studies (e.g., Kish and Hess, 1958) focusing on the kinds of blocks in which there were large discrepancies between census counts and survey counts of housing units, and more recent studies of the characteristics of persons not coveraged by telephone sampling frames (e.g., Thornberry and Massey, 1983). In addition, there have been special studies, most commissioned by the Census Bureau, that have used ethnographic methods as a tool to measure the kinds of person omitted from traditional household sampling frames (e.g., Valentine and Valentine, 1971). These have yielded the consistent finding that the poor, more socially isolated, more transient, younger , and male members of U.S. society tend to be subject to greater noncoverage errors in surveys attempting measurement of the household population.

However, coverage errors are mainly addressed by efforts to reduce them, not to measure them (i.e., they are the domain of reducers, not measurers). The typical adjustment procedures involve use of the latest estimates of the age, race and gender population distribution provided by the Census Bureau (see, for example, U.S. Department of Commerce, 1978). An alternative to adjustment attempts is restrictions made on the population of inference. For example, the large-scale movement to telephone surveys has essentially stripped away from inferential populations those persons who cannot be reached by telephone. Some careful researchers note their telephone surveys do not describe that population. In doing so, they have eliminated coverage error by a change of reference population. However, this merely trades error in inference for restrictions on the population described.

B-2. Nonresponse Error

In general, survey research has no more useful measures of nonresponse errors now than it did at its beginnings. Response rates have tended to be treated as proxy measures of nonresponse bias. In truth, they are only one component of such error. For household surveys the decline in response rates (Steeh, 1981) appears to continue, and there are more or less continuous efforts increase them. These take the form of incentives to respond (Chromy and Horvitz, 1978), learning the best times to call on sample persons (Weeks et al., 1980), etc.

A completely different approach to nonresponse error is found in work by survey statisticians, who build models of nonresponse likelihood and relationships with key survey variables and use them to adjust the data (Rubin, 1987) or imput for item missing data (Kalton, 1983). Weighting adjustments generally require the

assumption that the underrepresented (adjustment) groups are entirely homogeneous on the statistic of interest. Thus, by using weights to inflate those groups relative to others, the nonresponse bias will be removed. Given the assumptions of the model, the unadjusted statistics are biased, and the bias reduction can be estimated from comparison of adjusted and unadjusted statistics.

The development of selection bias in econometrics (Heckman, 1979) is relatively new acknowledgment by modelers that statistics of interest to them might be harmed by errors of nonobservation. These models require the specification of a predictive equation for the likelihood of not including some person in the survey. They thus relate to the traditional survey concerns of both coverage error and nonresponse error. Like the weighting adjustments of the survey statisticians, they yield measurements of the effect of nonresponse and noncoverage through comparisons of adjusted estimates. They are an example of a model-based adjustment for nonobservation error, since they require the analyst to specify a predictive model for nonresponse (or noncoverage) and use functions of the predicted likelihood of nonobservation to estimate adjusted parameters in a substantive model. Most analysts use selection bias models as an error reduction tool, not as a method of measuring error in estimates due to noncoverage or nonresponse.

B-3. Sampling Error

The basic developments in sampling theory were made early in this century and form the basis of practice today. Effects on the quality of simple statistics (e.g., means and totals) from changes in sample design are generally well understood, and reduction of sampling error to desired levels and measurement of resulting errors through probability sampling is common. With sampling error, as with other errors, however, modelers and describers take separate paths.

The simplest example of this debate concerns model-based and design-based inference in regression models (Holt et al., 1980; Dumouchel and Duncan, 1983). Design-based analysis of survey data reflects the sample design (e.g., weighting for unequal probabilities of selection) in statistical calculations. Model-based estimation does not, and treats the data as if they came from an unrestricted random sample. The debate between the two approaches generally focuses on whether the regression estimation is meant to describe a finite population or reflect an ongoing social process (of infinite duration), and on whether themodel is assumed to be well specified. For a more detailed discussion of sampling and sampling error, see Frankel and Frankel, this issue, pp. S127-S138.

C. Measurement Error

By far the most active field of research on survey quality concerns measurement error, the discrepancy between respondents' attributes and their survey response. There appear to be at least two reasons for the disproportionate attention to these errors : (1) statistical techniques have improved the capability of analysts to acknowledge some kinds of measurement errors (e.g., the development of confirmatory factor analytic techniques), and (2) in contrast to errors of nonobservation many measurement errors can be investigated using the available survey data themselves (without requiring outside sources). For our purposes, measurement error will be viewed as arising from influence of the interviewer, the weakness of the survey questions, failures of the respondent to give appropriate answers to the questions, and effects of the mode of data collection on survey answers. This section reviews the direct effects of these four sources of measurement error but omits mention of their combined effects.

C-1. Measurement Errors Arising From The Interviewer

For no other error source is the distinction between reducers and measurers clearer than for errors arising from the interviewer. Early work in survey methodology focused on how to "improve" interviewer performance, by selecting the correct kinds of persons to do the job (e.g., Sheatsley, 1951), by measuring the effect of experience (e.g., Booker and David, 1952). or by building rapport between interviewer and respondent(e.g., Kahn and Cannell, 1968). Evolving from that tradition are studies of the effect of demographic and characteristics of interviewers on respondent behavior, for example, race (e.g., Schuman and Converse, 1971) and gender (e.g., Groves and Fultz, 1985). Reducers have used this literature in a prescriptive manner - for example, the frequent practice of matching interviewer and respondent on race. The goal of such practice is to eliminate an error through a design change, not to measure the effects of race on responses.

Measurers attack the interviewer error problem in two different ways - one based on a statistical model of variance components, the other on observations of the failure of interviewers to follow training guidelines. Estimating what portion of the variance of a response distribution or survey statistics is attributable to interviews is prevalent in government survey agencies (e.g., Hansen et al., 1961). Since the estimation requires randomization of interviewer assignments, most empirical studies are special methodology experiments (e.g., Bailey et al., 1978) and not integrated with other analysis of ongoing surveys. The exception to this rule is found in centralized

telephone surveys, where randomization is more easily performed (e.g., Groves and Magilavy, 1986). The component of variance associated with interviewers is often found to be rather small in professional survey work but has been found to be large in responses to open questions dependent on interviewer probing behavior. The vast majority of studies measuring interviewer variance, however, are plagued by unstable estimates because of the small number of interviewers used in the study. Another measurement of interviewer effect is based on interaction coding techniques (Cannell et al., 1975). This techniques began with those making detailed records of the nature of small group interaction(Bales, 1950), has been implemented with audio tape recordings to measure compliance with interviewer training guidelines (Morton Williams, 1979), and has been adapted to simultaneous coding of interviewer behavior through monitoring of telephone interviews. This coding yields counts of alteration of question wording, inappropriate or inadequate probing, and skipped questions. It thus does not directly estimate errors in data but behaviors that are thought to produce them. These data could be used by modelers as indicators of the interview situation that might affect survey reponses, but little such work has been done. They are favored by reducers in attempts to identify problem questions.

Finally, there appears to be a recent sentiment to rethink the structure of the survey interview. For years most interviewing practice has focused on assuring consistency of questionnaire administration. The implied goal of these efforts is the standardization of the measurement instrument in hopes of achieving a consistent product from each respondent. Researchers of social interaction and discourse have noted that standardization of question wording does not necessarily imply constancy of meaning. Instead, using concepts from conversational analysis, they note that many of the normal mechanisms of assuring clear communication, of correcting misimpressions, of addressing the questions of the listener have been stripped away from the "standardized" interview. The effects of this may have been to minimize interviewer variance but to increase ***bias***, due to poor comprehension or minial memory search for relevant information (Jordan and Suchman, 1987). This perspective is inherent in the work of most ethnographers (e.g., Briggs, 1987) and is important to those who study interview discourse itself(e.g., Mishler, 1987).

C-2. Measurement Errors Arising From The Respondent.

Much methodological work on respondent error focuses on social psychological influences - motivation and social desirability effects. Flowing from the traditional view of the interview as a conversational with a purpose, efforts were made to increase positive effects of rapport on motivating the respondent to attend to

the task at hand. Similarly, investigations were mounted by reducers to find ways to limit the deleterious effects of social desirability, through open questions (Bradburn et al., 1979) and through randomized response techniques (Warner, 1965).

In contrast to social psychological influences on the respondent, less attention was paid to the cognitive demands of the interview. While there are many examples of early research on problems of recall for health events (e.g., Cannell and Fowler, 1963), for expenditure data (e.g., Biderman and Lynch, 1981), only recently have these problems been linked to theories from cognitive psychology (Jabine et al., 1984). This has taken the form of methods of cuing the retrieval of memories of past behavior (Loftus and Marburger, 1983), which have found temporal landmarks useful for reminding people of the time of occurrence of events. It has also focused on the effect of present mental states on measures of past mental states, on the mechanism by which context effects come about, and on the nature of reconstructing past events for reporting in the interview. The aim of this line of research is to reduce measurement error through the use of questions that prompt memory retrieval more efficiently. It is clearly in the reducer camp.

The work of measurers on respondent error mainly takes the form of estimating reliabilities of respondents in panel studies or in reinterview studies.

C-3. Measurement Errors Arising From The Questionnaire

Between Payne's work(1951) in the 1950s and Schuman and Presser's(1981) in the late 1970s, there was something of a hiatus in methodological work on survey errors arising from the questionnaire. Most current research is examining the effects of question error, structure, and wording and does not purport to investigate the measurement of error properties of questions. Instead, researchers note changes in response distributions associated with the alterations. They have explored why the difference exist (e.g., Schuman and Ludwig, 1983), what types of respondents are most subject to question wording effects (e.g., Kalton et al., 1978), and what topics seem most sensitives to question effects. Because much of the work in this field offers evidence that one question is preferable to another for certain purpose , it takes a perspective more akin to that reducers than to measures.

Measurers attack the problem of survey error arising from the questionnaire by asking multiple questions measuring the same concept of each respondent and estimating the amount of error variance associated with each (e.g., Andrews, 1984). This work relies on the multitrait-multimethod approach first suggested by Campbell

and Stanley(1963), but made statistically attractive with covariance modeling techniques (e.g., Lisrel). The perspective is inherent in much psychometric measurement theory, and is used most by those studying attitudes and abilities. The work provides model-based estimates of measurement error variance associated with question wording, but theoretical guidance is needed for identifying a set of equivalent measures of the same concept. If the questions do indeed measure the same concept then the estimates of error variance are accurate. If not, they are not. Given the popularity of multiple equation casual modeling in the social sciences, such estimation of error variances seems likely to proliferate.

C-4. Measurement Errors Arising From The Mode Of Data Collection

The mode of data collection as a source in survey data became a popular topic as costs and response associates with personal interviews became a source of concern. These have been many studies comparing mail, telephone, and face-to-face surveys (e.g., Hochstim, 1967) and telephone and personal visit surveys (e.g., Groves and Kahn, 1979). However, the unique contribution of mode to survey error is difficult to measure apart from its effects on nonresponse and coverage error, and most past studies provide insight only into the combined effects of these errors in two or three modes.

The nature of the effects of mode on measurement error is generally believed to focus on the comprehension of survey questions (hence, experiments on the cognitive demands of questions, like that of Miller (1984)) and on response delivery behavior(hence, the observation of truncated answers on open question (Groves and Kahn, 1979)). In addition, however, the mode of data collection is seen as a variable that interacts with interviewer effects: the Schuman et al.(1985) finding of more conservative racial attitudes reported by white Southerners in personal interviews (with largely white interviewers) than in telephone interviews (with interviewers of unknown race calling from Ann Arbor, Michigan). Similarly, the argument that interviewer variance is lower in centralized telephone surveys (Groves and Magliavy, 1985) than personal interviews is an example of how mode alters other measurement errors.

D. Underemphasized Properties Of Survey Quality

There are two contrasting reactions to the survey methodology literature among students encountering it for the first time: (1) why after so many years do

we know so little about how to improve survey quality, and, based on very different reactions to the same material, (2) why isn't all the knowledge about survey error used in practice? The answers to these questions. I think, lie in the nature of methodological research and its relationship to practical design problems.

Despite more than 50 years of research aimed at improving survey quality, there remain two issues that rarely receive serious attention:
(1) cost implications of error reduction, and (2) interrelationships of error sources.

Anyone who is a data collector or works with quickly learns that one of the dominant influences on design decisions is the available financial resources for the survey. Contrasting with the emphasis on quality and reduction of error in survey research methods texts is the frequent disregard of those prescriptions under the constraints of cost. The reduction of errors requires the expenditure of scarce resources. The easiest example is the reduction of sampling error with increasing sample size, requiring the measurement of more members of the population. Similarly, however, from a psychometric perspective, construct validity is increased by increasing the number of indicators measuring the same underlying concept (see lord and Novick, 1968). This means more questions in the survey; more questions imply more time to complete the questionnaire. In surveys as in business, time is money. Another example is that most models of interviewer variance imply that increasing the number of interviewers can reduce the impact of interviewer variability on survey statistics (e.g., Hansen et al., 1961). But hiring and training more interviewers inflates supervisory costs. Finally, nonresponse bias on sample statistics such as sample means can be presented as a function of the proportion of nonrespondents and the difference between the statistic for nonrespondents and respondents. Decreasing nonresponse rates often demands more followups on reluctant cases or use of more expensive interviewers (or more expensive measurement techniques). In all of these examples, reducing error costs money.

The inextricable link between costs and errors rarely is formally acknowledged in methods articles in POQ, or I any other scholarly jonurnal for that matter. That state of affairs has two detrimental effects: (1) methodologists invent methods to reduce an error, fail to measure the cost impact of the new idea, and (2) practitioners reject new ideas until it becomes clear that they result in reduced costs. Given the link between errors and costs, many new ideas require spending money to reduce an error. Given fixed budgets, the reduction of one error often leads to the increase in another.

A solution to the divergence between the results of survey methodology and

survey practice requires acknowledgment that surveys are inherent compromises. To become perfect measuring devices they must stop being surveys (as we know them). For example, any researcher who has constructed a questionnaire knows that each single question could usefully be expanded to produce an entire survey of its own. No single question minimize measurement error for most purposes, but (and this is the constraining factor) no survey is designed to measure just one attribute of the respondents. As Kish(1987) notes, surveys are multipurpose, both in the measures implemented and the subpopulations or subclasses of interest (e.g., regions, age groups, etc.). The clear implication of such observations is that both survey errors and survey costs ought to be assessed when survey quality is at issue. There are subfields of survey research design which sometimes incorporate this perspective. Survey sampling, for example, offers a set of classic design problem, in which the cost per unit sampling error is minimized. This requires the construction of a cost and error impacts (e.g., the number of cases chosen for the sample). In stratified design in which separate samples are taken from each stratum and there are different costs or levels of variation on the sample to strata and the best overall sample size, given the fixed resources available for the study. Similar models could be built on the optimal number of indicator for an underlying concept, given the cost constraint of a maximum amount of time to complete the questionnaire.

The second problem ignored in most methodological investigations is the existence of relationship among different error sources. Decreasing the error from one source often changes the value from another. For example, the Valentine and Valentine study(1971) of low income black communities shows that reluctance to report the existence of a household member was related to the frequency of his presence there. That is, within-household noncoverage is no doubt related to nonresponse due to noncontact. Groves and Magilavy (1985) show that questions presenting the respondent with more difficulty are those most affected by interviewer variation. That is, response error due to the question or the respondent is related to response effects due to the interviewer. In addition, there is a strong folk belief among survey researchers that reluctant respondents (requiring vigorous persuasion) tend to provide response of low quality. That is, nonresponse error and measurement error are correlated.

Despite this evidence there is little work examining the relationships between different error sources. The absence of this work implies that overall progress inimproving survey quality may be related because of negative correlations between individuals errors. For example, improving response rates through heavy persuasion may lead only to a more comprehensive data with larger measurement error. Which error is preferable? Which error is more deadly? Obviously, answers to such questions can be obtained only as a result of the joint investigation of both errors.

E. Ingredients Of a Theory Of Surveys

Many of the problems listed above could be ameliorated if a theory of surveys existed. The word theory can be so inflammatory when used in conjunction with surveys that it is necessary to offer careful definitions of what is and is not meant by the term. By a "theory of surveys" I mean a set of linked concepts and propositions that can be used to guide a particular survey design to achieve maximum cost efficiency. "Cost efficiency" in turn means that survey quality is maximized given the cost available for the survey.

From the discussion above, it should be clear that a theory of surveys must include concepts that relate to each the error sources identified above and to their interrelationships. Those espousing :total survey error" considerations at the design phase support this notion (see Andersen et al., 1979). Further, it must relate design features to costs, because they form the constraint on any increase in survey quality. Unfortunately, this point has been underemphasized in past work.

A theory of surveys would unite social science concepts with the statistical properties of survey estimates. It would consist of statements explaining, for example, how properties of the survey interaction (e.g., similarity of characteristics of respondent and interviewer) act to affect attitudes of each actor and, though them, compliance and response behavior on the past of the respondent. In addition, it would speak to the costs of altering the nature of the interaction, because these costs directly affect other errors. It would inform the researcher of links between the causes of absence of persons from the sampling frames (noncoverage), motivation for compliance with the survey request, and response performance. It would identify design alternatives that would manipulate these casual factors and provide assessments of their cost and of their indirect effects on other errors.

Ingredients for a theory of surveys must come from cognitive psychology, which concentrates on the processing of questions by the respondent, from the study of social interaction, focusing on how two speakers affect each other in the communication process, and from sociological perspectives on intergroup relations, which provide insights into the role of social measurement in societal processes. Psycholinguistic insights into comprehension have important implications for survey interviews and social-psychological concepts of compliance and influence can help us understand the respondent's decisions to participate in surveys. Indeed, a theory of surveys must draw on all these finds and unite them with a focus on survey quality and cost.

F. Summary

With this call for integration, this short article comes full circle. It began by observing that the various disciplines that use survey methods employ diverse and conflicting sets of concepts to guide the measurement of survey quality. It notes that these differences led to subset of researchers focusing on different components of the survey as particularly problematic and worthy of study. The different approaches tend to use different word to describe survey quality. None of them offers conceptual structures that encompass all the relevant issues of survey data quality.

The article ends with a call for the exploitation of that loose confederation of disciplines to develope a more formal theory of surveys. Such a development would require a rejection of the disciplinary blinders that limit researchers' focus to single error sources. To overcome these constraints will require the work of cross-disciplinary teams and the training of new survey methodologists whose vision is unimpeded by blinders.

♠ References ♠

Andersen, R.,J. Kasper, kasper, M.R. Frankel, and Asocistes (1979)
Total Survey Error. San Francisco: Jossey-Bass

Andrews, Frank (1984)
"Construct validity and error components of survey measures: A structural modeling approach." Public Opinion Quarterly 48:409-422

Bailey, L., T.F. Moore, and B.Bailar (1978)
"An interview variance study for the eight impact cities of the National Crime Survey cities sample." journal of the American Statistical Association 73:16-23.

Bales, R. F. (1950)
Interaction: A methodology of the study of small Groups. Reading, MA: Addison-Wesley.

Biderman, Albert D, and James P. Lynch (1981)
"Recency bias in data in self-reported victimization." Proceedings of the American Statistical Association.

Booker, H. S., and S.T. David (1952)
"Differences in results obtained by experienced and inexperienced interviewers." Journal of the Royal Stasistical Society, Series A, Part II : 232-257.

Bradburn, N.M., S. Sudman and Associates (1979)
Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass.

Briggs, Charles (1987)
Learning How to Ask . New York: Cambridge University Press.
Campbell, Donald T., and Julian C. Stanley (1963)
Experimental and Quasi-Experimental Designs for Research. Chicago: Rand-Mclay.
Cannell, C. F. and F.J.Fowler (1963)
A study of reporting of visits to doctors in the National Health Survey. Ann Arbor: Survey Research Center, The University of Michigan.
Cannell, C. F., S. Lawson, and D. Hausser (1975)
A Technique for Evaluating Interviewer Performance. Ann Arbor : Survey Research Center, The University of Michigan.
Chromy, James, and Daniel Horvitz (1978)
"The use of monetary incentives in the National Assessment Household Surveys." Journal of the American Statistical Association 73:473-4478.
Groves, Robert M., and Nancy Fultz (1985)
"Gender effects among telephone interviewers in a survey of economic attitudes." Sociological Methods and Research 14:31-52.

【붙임2】

Improving Quality of Surveys

David A. Marker
Westat

Keynote Speech
International Conference on Improving Surveys
Copenhagen, Denmark
August 26, 2002

DavidMarker@Westat.com

Where are the Greatest Opportunities for Quality Improvement in Surveys?

- Response Rates
- Technological Changes
  - *Internet*
  - *Mobile telephones*
  - *Handheld computers*
- Communications

Regardless of the scarcity of hard data and the difficulties of making precise comparisons there are a significant number of reports of completion rates declining, or where achieving a satisfactory completion rate is becoming increasingly more difficult.

American Statistical Association
Conference on Surveys of Human Populations, 1973
Reported in the February 1974 American Statistician

Regardless of the scarcity of hard data and the difficulties of making precise comparisons...

■ We have made progress on collecting data and making comparisons.

De Leeuw and De Heer (2002) and others have documented rates internationally.

Not all have dropped, but there is a general trend.

Contact rates more affected than refusal rates.

Drop in contact rates independent of type of survey.

■ CASRO (1982) and more recently AAPOR (2002) have established standard formulae for computing response rates.

achieving a satisfactory completion rate is becoming increasingly more difficult.

- This is definitely true
- More families have all adults working
  - *Home less hours*
  - *When they are home they are too busy to talk*
- Competition from telemarketers
  - *Don t call lists*
  - *TeleZappers and similar technologies*
  - *Caller ID*
  - *Call screening*
- More common to have unlisted phone numbers
- Increased number of households have only mobile telephones

What have survey organizations done to overcome this?

- Worked harder
  - *7 attempts not uncommon on in person surveys*
  - *Double the number of attempts on telephone surveys*
    - *30 attempts to convert max. calls on NSAF*
    - *Double the effort for same response rates in 1996 as 1979 (Curtin, Presser, and Singer, 2000).*
  - *More of the contacts in evenings and on weekends*
  - *More sophisticated advance letters*
  - *Mixed mode surveys*
- Longer data collection periods
- Research on process of gaining cooperation (Groves and Couper, 1998)

What have survey organizations done to overcome this?

- Refusal conversion
  - *Expert staff*
  - *Written letters verifying importance of survey*
  - *Wait before contacting again*
- Financial and other incentives
  - *Can reduce overall costs*
  - *Can reduce biases*
  - *Can introduce other biases*

What is the impact of this on survey quality?

- Costs more
- Delays timeliness of results
- Increases potential biases
  - *Depends on topic*
  - *Respondents have higher SES (Goyder et al., 2002) and are healthier (Cohen and Duffy, 2002)*
  - *AAPOR (2002) examples with no apparent bias*
- Need more research on improving *processes* that can minimize nonresponse and/or help adjust for it

Technological Change

Ronald Snee (16 September, 1991 Washington Statistical Society talk) :

【 3 sure ways to ruin yourself 】

- Gambling → Quickest
- Sex → Most Enjoyable
- Overemphasis on Technology → Most sure to lead to ruin!

Internet :

■ **Data collection**

Inexpensive, timely, full-color graphics
How representative is it, now and in the future?
Tailor questionnaires to each respondent
Allows low cost, low quality, competition
Avoiding multiple responses by interested participants

■ **Data dissemination**

- *Web meta-data*
- *Hyperlinks to methodology sections and more detailed results*
- *How to provide measures of accuracy for on-line analyses?*
- *How to maintain confidentiality for respondents?*

Mobile Phones:

- Fact of life in Europe, growing in the United States
- Replacing land lines in U.S. requires changing current cost structure (maybe 15-20% talk primarily on cell phone?)
- There will be talks this week on how European NSOs are dealing with this
- Will be required within 10 years for representative telephone surveys in U.S.
- Especially for studies of young people
- Does allow for collection of data via telephone from non-telephone households
- Non-contacts will be reduced, but refusals will probably go up

Handheld Computers:

- Allow for easier data collection in more extreme conditions

  Factories, homes, beaches, street corners

  Need rugged design
- Tablet use for NHANES will be described in talk by Berman et al. this afternoon (Binzer and Hill, 2002)
- Currently Palm pilots cannot access all the look-up tables and other mega-memory requirements

How Do We Improve Quality?

EFQM, ABS, (Trewin, 2001) and Statistics Canada (Brackstone 1999) measure quality by

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Interpretability
- Coherence
- Objectivity
- Clarity

How do we improve quality (cont.)?

- Can t *improve* quality without including cost (staff hours)
- Communications
- Continuous quality improvement

Communications - Internal

Two-way communication (NCES STD 3-1-02)

Staff responsible for NCES data collections that are used as sampling frames should maintain two-way communications with survey staff who use their collection as a frame. Procedures such as sharing preliminary data files with survey staff in order to develop frames may be instituted.

NCES survey staff that use NCES data collections as a frame should share any coverage or usage issues with the NCES data collection staff so that the coverage can be improved for future uses.

Communications - Internal (cont.)

- Process Maps
  - *Customer-supplier relationships*
  - *Staff don t work by themselves*
- Nothing as depressing as doing the wrong thing well!
- Wasa - no one wanted to tell the king

Current Best Methods

- Statistician - programmer communications
- Statistician - project staff
- Project Staff - field staff

  Revisions - numbers and changes paragraph

  Inputs and output files

  Check outputs (tables, record counts)

Communications - With Customers

- Meta-data

  Documentation on web and hard copy

  Methodology

  - *Effect of response rates on accuracy*
  - *How to compute accuracy*

  How to interpret results

- Press releases

  Develop templates

  Senior management agreement on content

  Train media on using statistics

Continuous Quality Improvement

- Management must be responsive to staff improvements
  - *Provide resources (time, skills)*
  - *Recognize process improvements are key*
- Staff must view this as part of their job
  - *Work together in teams*
  - *Measure their processes*
- Need stable systems
  - *To predict timeliness, accuracy, and costs*
  - *To identify real improvements*

Summary

- Response rates
- Technological improvements
- ***Improve*** quality, don t just measure it
- Communications - internal and external
- Continuous quality improvement

참 고 문 헌

Robert F. Belli , William L. Shay and Frank P. Stafford,
"Event History Calendars and Question List Surveys"

Ba"rtel Kna"uper ,
"Introduction to Questionnaire Design"

Dr. Paul P. Biemer and Lars E. Lyberg ,
"Introduction to Survey Quality"

M. GHOSH and J.N.K. Rao ,
"Small Area Estimation"

Robert M. Groves,
"Research on Survey Data Quality"

David A. Marker
"Improving Quality of Surveys"