Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 591~603

Implementation of Quantitative Unrelated Question Model for Obtaining Sensitive Information at On-Line Survey[1)]

Hee-Chang Park[2)] · Jee-Hyun Ryu[3)] · Sung-Yong Lee[4)]

Abstract

This paper is planned to use randomized response technique which is an indirect response technique on internet as a way of obtaining much more precise information, not revealing secrets of responsors, considering that respondents are generally reluctant to answer in a survey to get sensitive information targeting employees, customers, etc.

Keywords : 확률화응답기법, 양적무관기법, E-R diagram

1. 서론

최근 들어 기업 등에서 온라인 회원, 즉 고객을 대상으로 혹은 자회사의 사원을 대상으로 기업의 경영과 관련된 각종 정보를 얻기 위해 온라인 설문조사가 현장에서 많이 행해지고 있으나, 온라인 설문조사 역시 기존의 설문조사와 마찬가지로 조사자가 응답자들의 프라이버시나 사생활과 관련된 민감한 정보를 얻고자 할 경우에는 정확한 정보를 얻는 것이 쉽지 않다. 응답자들은 기존의 설문조사와 마찬가지로 온라인 상에서 민감한 질문을 직접적으로 받게 되면 자신의 비밀이나 사생활이 노출될까 의심하여 혹은 응답 후의 부당한 대우를 받게될까 두려워 정직한 응답을 꺼리게 된다. 따라서 기업 의사결정에 대한 찬·반 여부, 성적 질문 등과 같은 민감한 정보를 얻고자 할 경우 간접응답기법인 확률화응답기법(randomized response technique ; RRT)을 온라인 설문조사에 적용해 볼 수 있다. 이 기법을 사용하게 되면 응답자들의 응답이

1) This research is financially supported by Changwon National University in 2002.
2) Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea
E-mail : hcpark@sarim.changwon.ac.kr
3) Graduate Student, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea
4) Associate Professor, Department of Industrial & Systems Engineering, Changwon National University, Changwon, Kyungnam, 641-773, Korea

확률장치를 이용하여 간접적으로 이루어지게 되므로 응답자가 자신의 신상에 대한 불안이나 개인 정보의 유출을 이유로 정확하지 않은 응답을 할 가능성을 줄일 수 있다.

본 연구에서는 기업체 등에서 사원이나 고객을 대상으로 민감한 정보를 얻기 위한 조사에서 응답자가 정직하게 응답하기를 꺼리는 질문에 대하여 응답자의 비밀을 노출시키지 않고서 보다 정확한 정보를 얻을 수 있는 간접응답기법인 RRT 중에서 양적무관질문기법에 대해 인터넷상에서 사용할 수 있는 시스템을 구현하고자 한다. 이러한 RRT는 Warner(1965)에 의해 처음으로 제안되었는데, 그는 응답자들에게 민감한 질문과 민감한 질문에 배반되는 질문으로 구성된 확률장치를 사용하여 민감한 속성에 대한 질적 정보를 얻고자 하였다. Greenberg 등(1969)은 민감한 질문과 배반되는 질문 대신에 민감한 질문과 전혀 무관한 질문을 사용하는 무관질문기법(unrelated question technique)을 제안하였다. 또한 Greenberg 등(1971)은 이를 민감한 변수에 대한 양적 정보를 얻기 위해 양적속성기법으로 발전시켰다. Loynes(1976)는 Warner기법의 민감한 질문과 배반이 되는 질문 대신에 “예”라고 응답하도록 강요하는 강요질문기법(forced answer technique)을 제안하였다. Fox와 Tracy(1986), Chaudhuri와 Mukerjee(1988)는 RRT를 정리, 요약하여 체계화하였으며, 최근에는 이들 이론들의 실제적 활용에 많은 관심이 집중되고 있다.

국내에서는 이기성(1992)이 2단계 확률화응답모형에 관한 연구를 수행한 바 있으며, 류제복 등(1993, 1995)은 RRT가 적용된 사례들을 비교 분석하여 실용화를 위한 방안을 제시하였다. 박희창 등(2001)은 질적 자료의 관련질문기법에 대한 온라인 설문조사 시스템을 구현한 바 있으며, Park과 Myung(2002)은 질적 자료의 무관질문기법에 대한 인터넷시스템을 개발하였다.

본 논문에서는 RRT 중에서 양적 무관질문기법을 인터넷상에서 사용할 수 있는 시스템을 구현하고자 한다. 2절에서 양적 무관질문기법에 대해 전반적으로 살펴본 후, 3절에서는 시스템의 개발환경, 시스템 흐름, 그리고 시스템의 구성 등에 대하여 살펴보고, 4절에서는 실제로 적용한 예제를 보여주고, 5절에서 결론을 맺고자 한다.

2. 양적 무관질문모형

본 절에서는 RRT 중에서 양적 무관질문기법에 관하여 기술하고자 한다. 양적 무관질문모형은 Greenberg 등(1971)이 연구한 질적 무관질문모형을 양적 속성으로 확장한 것으로, 술 소비량, 낙태 횟수, 미성년자의 성교 횟수 등 민감한 속성이 양적인 속성을 가지는 경우에 응답자의 신분이나 비밀을 노출시키지 않고서도 민감한 질문에 대한 정보를 이끌어 낼 수 있는 기법이다. 민감한 속성을 가진 질문과 민감한 질문과는 전혀 관련을 가지지 않는(민감한 그룹과 전혀 관련이 없는 그룹 Y) 질문을 사용함으로써 응답자에 대해 더 많은 신뢰를 얻을 수 있다.

양적 무관질문모형은 “민감한 그룹과 전혀 관련이 없는 그룹 Y”의 모집단평균 μ_y를 알 경우와 모를 경우의 2가지로 나누어진다. 먼저 μ_y를 알 경우 Greenberg 등이 사용한 양적 무관질문기법의 확률장치는 다음과 같은 2개의 설문으로 구성된다.

설문 1 : 당신의 민감한 변수 X에 대한 값은 얼마입니까?

설문 2 : 당신은 무관한 변수 Y에 대한 값은 얼마입니까?

이 때, 단순임의복원으로 추출된 n명의 응답자들은 확률장치에 의해서 선택된 질문에 대하여 양적으로 응답한다. 설문 1이 선택될 확률을 p, 설문 2가 선택될 확률을 $1-p$ 라고 하면 응답자들이 "Z"라고 응답할 확률은 다음과 같다.

$$\mu_z = p\mu_x + (1-p)\mu_y \tag{2.1}$$

여기서 μ_x는 민감한 변수 X에 대한 모평균이며, μ_y는 무관한 변수 Y에 대한 모평균으로 알고 있다고 가정하며, 위의 두 설문 중 선택된 설문에 대하여 $z_i(i=1,2,\ldots n)$라고 응답을 하도록 한다. 이 때, 선택된 질문은 응답자만이 알고 있을 뿐 조사자는 알지 못한다.

민감한 변수 X에 대한 모평균 μ_x의 추정량과 $\hat{\mu}_x$의 분산 추정량은 다음과 같다.

$$\hat{\mu}_x = \frac{\bar{z} - (1-p)\mu_y}{p} \tag{2.2}$$

$$\widehat{Var}(\hat{\mu}_x) = \frac{s_z^2}{np^2} \tag{2.3}$$

여기서 $\bar{z} = \sum_{i=1}^{n} z_i/n$이고, $s_z^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2/(n-1)$이다.

한편 Y의 모평균 μ_y를 알고 있다는 가정은 일반적으로 타당하지 않으므로 μ_y를 추정하여야 한다. 구하고자 하는 미지의 모수가 μ_x 하나일 때는 단지 하나의 표본이면 충분하나, 두 개의 모수 μ_x와 μ_y를 모두 모를 때는 최소한 두 개의 표본이 필요하게 된다. 따라서 모집단으로부터 단순임의복원으로 크기가 n_1과 n_2인 두 개의 독립표본을 추출하여 μ_x와 μ_y의 추정량을 구한다. 두 개의 표본을 사용해야 되므로 $i\ (i=1,2)$ 번째 표본에서 민감한 질문이 선택될 확률이 p_i가 되는 두 개의 확률장치가 필요하게 된다. i번째 표본에서 응답자들이 "Z"라고 응답할 확률은 다음과 같다.

$$\mu_{zi} = p_i\mu_x + (1-p_i)\mu_y \tag{2.4}$$

민감한 변수 X에 대한 모평균 μ_x의 추정량과 $\hat{\mu}_x$의 분산추정량은 다음과 같다.

$$\hat{\mu}_x = \frac{(1-p_2)\bar{z_1} - (1-p_1)\bar{z_2}}{p_1 - p_2} \tag{2.5}$$

$$\widehat{Var}(\hat{\mu}_x) = \frac{\frac{(1-p_2)^2 s_1^{\,2}}{n_1} + \frac{(1-p_1)^2 s_2^{\,2}}{n_2}}{(p_1-p_2)^2} \tag{2.6}$$

여기서 $p_1 \neq p_2$, $\overline{z_1} = \sum_{j=1}^{i} z_{1j}/n_1$, $\overline{z_2} = \sum_{j=1}^{i} z_{2j}/n_2$, $s_1^{\,2} = \sum_{j=1}^{n_1} (z_{1j}-\overline{z_1})^2/(n_1-1)$, 그리고 $s_2^{\,2} = \sum_{j=1}^{n_2} (z_{2j}-\overline{z_2})^2/(n_2-1)$이다.

3. 양적 확률화응답시스템의 구현

3.1. 시스템 개발 환경 및 시스템 흐름

구현된 시스템의 개발 환경에서 개발 언어는 gnu c compiler, Java, html등이며, 운영 체제는 Linux를 사용하였으며, 데이타베이스는 MySQL을 사용하였다. 본 연구에서 개발된 시스템은 데이터베이스에 바탕을 두어 기존의 온라인 설문조사 시스템과 더불어 사용할 수 있을 뿐만 아니라 독립된 스팟 서베이(spot survey)가 가능하도록 개발되었다. 그리고 본 시스템은 동일한 응답자가 여러 번 답하는 것을 막기 위해 로그인(log in)을 하는 사이트에서는 동일 아이디에 대하여 중복 응답을 하는 것을 막을 수 있도록 구현하였다.

양적 확률화응답시스템은 관리자(조사자) 모드와 응답자 모드의 두 가지 모드로 구성되어 있으며, 관리자 모드에서는 설문을 작성하는 에디터와 확률을 입력하는 부분으로 이루어져 있고, 모집단의 평균을 알 때와 모를 때를 선택할 수 있다. 응답자 모드는 실제 응답자가 응답을 할 수 있도록 이루어져 있으며, 모평균을 알 경우에는 모든 응답자가 확률 장치로 넘어 가며, 모평균을 모를 경우에는 접속 순서에 따라 모집단 1과 모집단 2로 나누어져 응답에 참여하도록 하였다. 응답의 결과는 데이터베이스로 저장되며, 모평균의 추정값과 분산추정값을 계산한 후에 응답자에게는 모평균의 추정값만을 보여주고, 조사자(관리자)에게는 모평균의 추정값과 분산추정값에 대한 결과를 모두 보여주도록 구성하였다.

본 시스템은 자료의 입력에서 처리, 결과를 모두 데이타베이스를 바탕으로 이루어져 있다. 이로 인하여 동일 응답자 등의 반복 측정에서도 기존의 설문응답시스템과 쉽게 합쳐서 사용할 수 있다. 처리과정에서도 데이타베이스를 사용함으로 인하여 쿼리(query)를 사용하여 수행 속도면에서 파일시스템보다 빠르게 진행할 수 있다. 또한 기본적인 결과를 데이타베이스에 저장함으로 계속적인 조사에서 추세 분석이 가능하다. 본 시스템의 구현을 위해 설계한 테이블은 <표 3.1> 부터 <표 3.4>와 같다. <표 3.1>의 메인 테이블은 고유한 테이블이다. 여기에 모든 설문 문항들에 대한 정보를 보관하고 있으며 아래에 나오는 설문지라는 테이블과 연계하여 분석을 수행한다. 여기에는 설문 작성일시, 주제, 문항수, 소개 등이 기록되며 고유한 인텍스키로 설문지 테이블을 참조한다.

<표 3.1> 메인 테이블

메 인			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
설문 작성일	day	date	
설문 작성시간	time	time	
설문의 주제	subject	varchar(79)	
설문 문항수	number	tinyint	
설문 작성 종료 검사	check	char(2)	
설문의 소개 및 인사말	title	blob	

<표 3.2> 설문 테이블

설 문			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
설문 1	q1	varchar(79)	
설문 2	q2	varchar(79)	
확률화응답기법종류	mode	varchar(2)	
확률장치종류	type	varchar(2)	
표본 1에서 설문1이 선택될 확률	p1	float	
표본 2에서 설문1이 선택될 확률	p2	float	
알고있는 모비율	p3	float	
표본 1참여자	n1	smallint	
표본 2참여자	n2	smallint	

<표 3.2>에서 제시하는 설문 테이블은 실제 설문 문항의 정보를 보유하는 테이블이다. 설문 문항의 내용, 확률장치, RRT 종류, 설문 1이 선택될 확률 등 상위 테이블인 메인에서 인덱스키를 참조하여 설문 생성시 추가로 생성된다.

<표 3.3> 표본 1 테이블

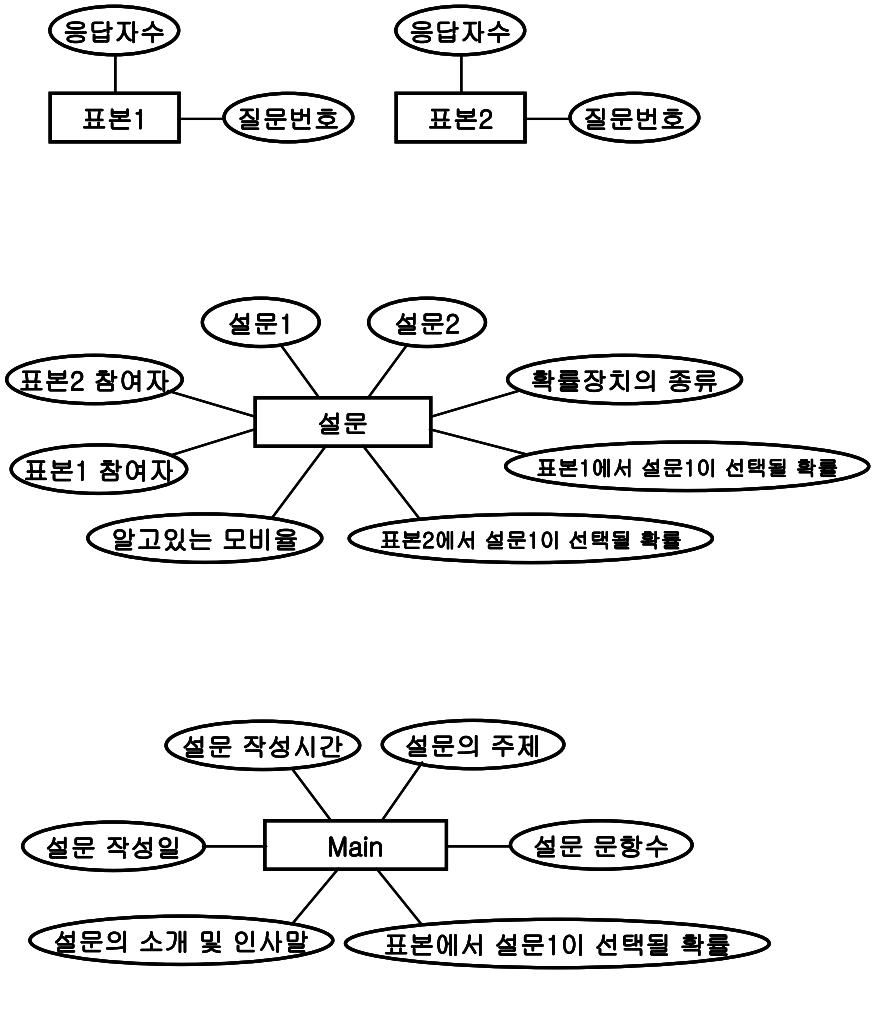
표본 1			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
양적 수치	s1	mediumint	

<표 3.4> 표본 2 테이블

표본 2			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
양적 수치	s1	mediumint	

<표 3.3>과 <표 3.4>의 테이블은 응답자 집계용 테이블이다. 설문지 테이블에서 넘어온 문항을 응답한 수치가 기록되는 테이블이다. 표본 2테이블은 양적 이표본 무관 질문기법 - 무관한 속성의 모평균을 모르는 경우에는 최소한 두 개의 표본이 필요하게 된다. 따라서 모집단으로부터 단순임의복원추출법으로 크기가 n_1, n_2인 두 개의 독립 표본을 추출할 경우 사용된다.

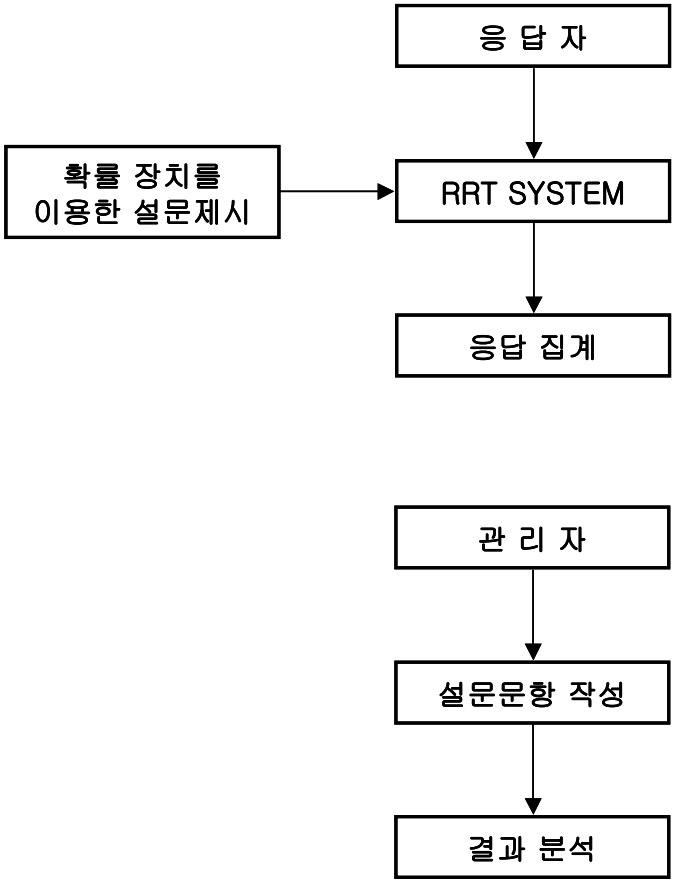
본 시스템에 사용된 데이터베이스 구성은 <그림 3.1>과 같은 구조를 가지고 있다. 메인 테이블은 설문이 선택될 확률 및 주제, 작성일자 등의 자료를 가지고 있으며, 표본 테이블은 설문 테이블에서 응답자들이 응답한 결과를 집계하여 결과를 보여주는 구조를 하고 있다.



<그림 3.1> E-R Diagram

3.2 시스템의 구성

본 시스템의 구성은 응답자 모드와 관리자 모드 2개로 구성되며 구성도는 <그림 3.2>와 같다.



<그림 3.2> 시스템 구성도

본 시스템은 관리자 모드와 응답자 모드의 두 부분으로 구성되어 있다. 관리자 모드는 설문의 생성 및 삭제를 담당하는 부분이며, 응답자 모드는 실제응답자들이 설문에 응답을 하는 부분이다.

설문을 생성하기 위해서 <그림 3.3>에서 <그림 3.5>와 같이 관리자 초기화면에 접속하여 설문의 소개, 설문의 주제, RRT의 종류, 문항의 수 등을 입력한다.

<그림 3.3> 관리자 초기화면

<그림 3.4> 양적 무관질문기법 - 무관한 속성의 모평균을 아는 경우

<그림 3.5> 양적 이표본 무관질문기법 - 무관한 속성의 모평균을 모르는 경우

본 시스템은 4종류의 확률장치 중 하나를 선택하도록 되어 있으며, 선택될 확률에 따라 동전의 앞면 혹은 뒷면, 연필의 오른쪽 혹은 왼쪽과 같이 두 사건 중 하나의 형태로 나타난다.

응답자 화면은 <그림 3.6>과 같으며, 여기서 응답자들은 설문 문항에 대하여 응답할 수 있다.

<그림 3.6> 응답자 화면

관리자는 응답자가 설문 1에 대하여 응답을 하였는지 설문 2에 대하여 응답을 하였는지 알 수 없도록 설계되어 응답자가 진실 되게 응답할 수 있도록 하였다.

4. 예제

이 장에서는 확률화응답시스템을 적용하여 실시한 설문 조사로부터 얻어진 결과에 대해 논의 하고자 한다.

기간 : 2002. 1.4 ~ 1.31

대상 : (주)Esab일반 사무직 사원(150명)

문항 : 귀하는 한달에 평균 몇 번 무단 결근을 합니까?

응답결과는 응답자용과 관리자용으로 구분되어 있으며, 응답자용은 민감한 속성 A의 모평균의 추정값만을 <그림 4.1>과 같이 보여준다.

<그림 4.1> 응답자용 결과 화면

<그림 4.1>은 “귀하의 상사로부터 인사고과에서 불이익을 당한 적이 있습니까“라는 민감한 질문에 대하여 82명이 참여 하였고 모평균 π의 추정량 $\hat{\pi}$은 0.2995임을 보여주고 있다. 관리자는 <그림 4.2>과 같다. 여기서는 응답자 화면에는 볼 수 없는 설문 1이 선택될 확률, 표본 1의 크기, 표본 2의 크기, 표본 1이 선택될 확률 p_2, 표본 1에 응답한 평균, 표본 2에 응답한 평균, 모평균 μ의 추정량 $\hat{\mu}_x$와 $\hat{\mu}_x$의 분산 추정치 등을 볼 수 있도록 구현하였다.

<그림 4.2> 관리자용 결과 화면

<표 4.1> 관리자용 결과

민감한 설문의 선택확률	0.7
표본1의 크기, 표본 2의 크기	55, 27
표본 1에서 응답자 평균, 표본 2에서 응답자 평균	0.47, 0.7
모평균 μ의 추정량 $\hat{\mu}_w$	0.2995
$\hat{\mu}_w$의 분산추정치	0.1165

"귀하는 한달에 평균 몇 번 무단 결근을 합니까?"에 대한 직접질문에 대한 결과는 <그림 4.3>과 같다.

<그림 4.3> 직접 질문 기법

<표 4.2> 직접 질문 기법

표본의 크기	40
모평균 π의 추정량 $\hat{\pi}$	0.023
$\hat{\pi}$의 분산추정치	0.000562

RRT를 이용한 온라인 설문조사에 참여한 사람의 수는 82명이었고, 응답자 평균은 0.47, 0.7이었으며, 이로부터 월 평균 무단 결근을 한 사람들의 모평균을 추정해 본 결과 0.2995로 나타났으며, 분산추정치는 0.1165이었다. 그리고, 직접질문을 이용한 온라인 설문 조사에 참여한 사람의 수는 40명이었고, 이로부터 월 평균 무단 결근을 한 사람들의 모평균을 추정해 본 결과 0.023으로 나타났으며, 분산추정치는 0.000562이었다. 이와 같은 결과에서 알 수 있듯이 RRT를 이용한 온라인 설문조사에서 구한 모평균의 추정치가 직접질문을 이용한 온라인 설문조사에서 구한 모평균의 추정치보다 높게 나타났다.

5. 결론 및 향후과제

본 논문에서는 민감한 정보를 얻기 위한 조사에서 응답자들이 정직하게 응답하기를 꺼리는 질문들에 대하여 직접응답 대신에 간접응답을 통해 응답자의 비밀을 노출시키지 않고서 보다 정확한 정보를 얻을 수 있는 간접응답기법인 RRT를 인터넷 상에서 사용할 수 있도록 구현하였다. 본 시스템은 기존의 설문조사 시스템과 연계하여 민감한 질문에만 확률장치를 이용할 수 있도록 하여 다른 속성에 따라 민감한 질문에 대한 차이도 볼 수 있을 뿐만 아니라 독립된 단일문항 질문으로도 사용이 가능하도록 하였다. 기업내의 인트라넷이나 공공기관의 경우 민감한 질문에 대한 응답자들의 좀 더 정확한 응답을 기대할 수 있을 것으로 생각된다.

향후의 과제로는 다양한 표본추출방법을 적용하여 회원이나 구성원들을 대상으로 목적에 맞는 응답자를 추출하는 기능을 부과하여 기업 등에서 사원이나 고객을 대상으로 한 조사에서 실용성을 보다 높이는 방법이 필요하다고 생각된다. 그리고 강요질문기법, 질적 무관질문기법 모형 등 또 다른 모형들을 고려해야 하겠으며, 또한 다양한 질문에 대한 대처 방법과 응답자에게 흥미를 유발할 수 있는 멀티미디어적인 요소를 도입한 확률장치의 고안이 필요하다.

참고문헌

1. 김정기 · 김희재 · 남기성 · 박희창 · 이성철 · 정정현 (1999). 사회조사분석론, 창원대학교출판부.
2. 류제복 · 홍기학 · 이기성 (1993). 확률화응답모형, 자유아카데미.
3. 류제복 · 이계오 · 이기성 (1994). 확률화응답기법의 실용화 방안,

응용통계연구 8(1) : 9-26.

4. 박희창 · 이기성 · 김희재 · 남기성(2001), 인터넷조사와 설문조사시스템, 자유아카데미.
5. 이기성 (1992). 2단계 확률화응답모형에 관한 연구, 박사학위논문, 동국대학교 통계학과.
6. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*(1st ed.). New York : Marcel Dekker, Inc..
7. Coomber, R. (1997). "Using the Internet for Survey Research." *Sociological Research Online*, 2(2). <http://www.socresonline.org.uk/socresonline/2/2/2. html>
8. Fox, J. A. and Tracy, P. E. (1986). *Randomized Response : A Method for Sensitive Survey,* Sage Publications.
9. Greenberg, et. al. (1969). The Unrelated Question Randomized Response Model : Theoretical Framework, Journal of the American Statistical Association, 64, 520-539.
10. Schwarz, C. J. (1997). "StatVillage : An On-line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling." *Journal of Statistics Education* 5(2). <http://www.amstat.org/ publications/jse>
11. Warner, S. L. (1965). "Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60 : 63-69.

[2003년 7월 접수, 2003년 8월 채택]