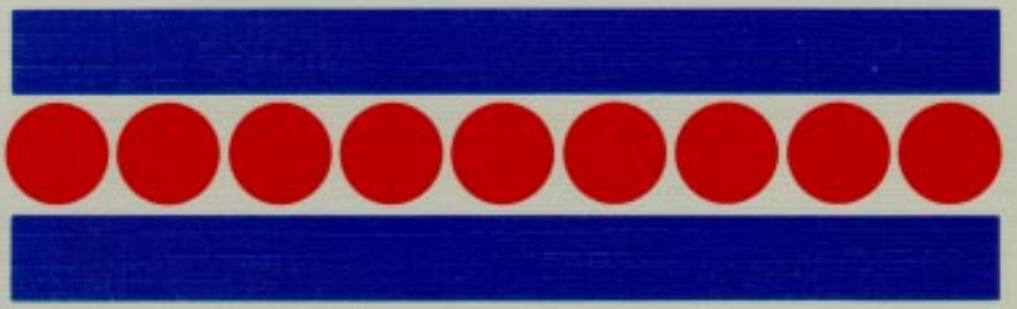
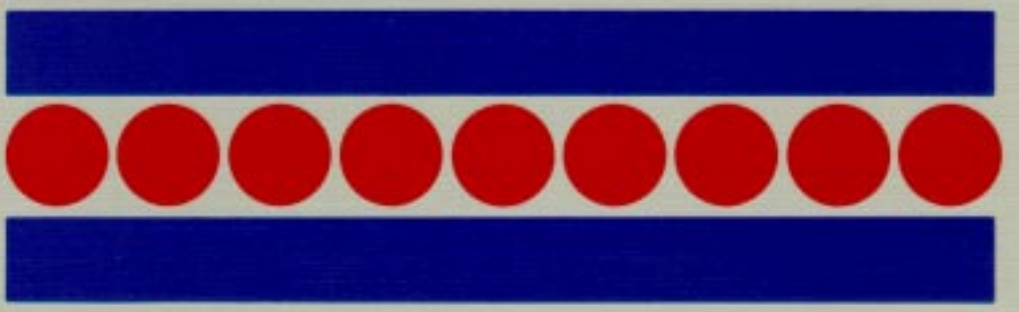
NAZIOARTEKO ESTATISTIKA MINTEGIA EUSKADIN 1990 SEMINARIO INTERNACIONAL DE ESTADISTICA EN EUSKADI



EVALUATION OF QUESTIONNAIRE DESIGN EFFECTS

GAD NATHAN



EUSKAL ESTATISTIKA - ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

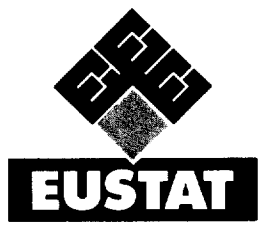
EUSTAT

NAZIOARTEKO ESTATISTIKA MINTEGIA EUSKADIN **1990** SEMINARIO INTERNACIONAL DE ESTADISTICA EN EUSKADI

EVALUATION OF QUESTIONNAIRE DESIGN EFFECTS

GAD NATHAN

KOADERNOA 19 CUADERNO



***Lanketa** / Elaboración:*
Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística

***Argitalpena** / Edición:*
Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística
C/ Dato 14-16 - 01005 Vitoria-Gasteiz

***Botaldia** / Tirada:*
1.000 ejemplares
IX-1990

Inprimaketa eta Koadernaketa:
Impresión y Encuadernación:
ITXAROPENA, S.A.
Araba kalea, 45 - Zarautz (Gipuzkoa)

***Lege-gordailua** / Depósito legal:* S.S. 729/90

ISBN: 84-7542-127-10 Obra completa
ISBN: 84-7749-073-2

BIOGRAPHICAL SKETCH OF PROF. DR. GAD NATHAN

Gad Nathan is Professor of Statistics and Chairman of the Department of Statistics at Hebrew University, Jerusalem. Previous positions include Director, Research and Development, Central Bureau of Statistics, Israel, and visiting positions at various universities and at government statistical offices around the world. Dr. Nathan received his M. Sc. in Statistics at Hebrew University in 1960 and his Ph. D. in Operations Research at Case Institute of Technology in 1964.

His research interests include: sampling methodology and nonsampling errors; inference from complex samples; extensions of response error models; surveys with multiplicities; and cognitive aspects of survey methodology. He has participated in several research projects with groups in Australia, Canada, England, France and the U.S. Dr. Nathan has some 50 publications in scientific journals, books and proceedings.

Dr. Nathan has served as Vice-President of the International Statistical Institute and is on the Council of the International Association of Survey Statisticiens. He has recently been appointed as Chairman of the Israel National Public Council of Statistics.

CONTENTS

0. INTRODUCTION

It is well known that the survey process, just as any other measurement process, is associated with a variety of errors. For sample surveys the component of sampling error has been the subject of intensive investigation and research over the past five decades. A variety of sampling methods have been devised in order to minimize sampling variance for given costs and special methods have been developed to allow the extensive use of auxiliary information. These developments have led to the use of sample designs which attain a high degree of efficiency. Sophisticated methods, such as multiple frame sampling, multiplicity sampling, double sampling, controlled sampling and a whole range of PPS selection methods, have led to further gains in efficiency in particular situations. In addition, a large number of estimation methods - various types of ratio and regression estimators, post-stratification, model-based estimation and others - have been developed and are widely used in order to achieve additional gains in efficiency.

Together with the development of methods of sampling and of estimation, to ensure the most efficient (or more efficient) utilization of resources, sample survey methodologists have taken great pains in order to ensure that not only are sampling errors reduced as far as possible, but also that the results of this reduction can be evaluated. This has been done primarily by developing methods of variance estimation, which ensure that the sample design and estimation method used provide the possibility to estimate the sampling error with sufficient accuracy. Thus many statistical agencies use stratification only to the degree that at least two sample units are guaranteed to be selected in each stratum (rather than the more efficient stratification with only a single sample unit per stratum). This is done to ensure that unbiased estimates of the sampling variance can be obtained. Several sophisticated methods have been devised to ensure efficient estimation of sampling errors, even for the most complex sampling and estimation procedures. These include methods of linearization, balanced repeated replication and Jackknife. Most statistical agencies, both governmental and private, have some facility for estimating sampling errors on a current basis and many of them publish these on a routine basis. While most of the effort in reducing and in evaluating sampling errors relates to variances (and in some case also to the covariances between estimators of different parameters), attention has also been given to sampling bias. Most sampling methods and methods of estimation are unbiased (when non-sampling errors, such as those due to non-response, are disregarded). However, even if methods with inherent bias are used, such as ratio-estimation, care is taken to ensure that the bias is small, and where possible, can be evaluated.

The subject of the complementary survey errors - non-sampling errors, although well-recognized as extremely important, has not received the same kind of systematic and thorough attention, that the subject of sampling errors has received. This lacuna can be explained by a number of factors. Due to their very nature and definition as complementary to sampling error, non-sampling errors cannot be considered as clearly defined as sampling errors. In fact the term covers a variety of different types of errors due to various sources - such as: questionnaire design, interview effects, non-response, misclassification, reporting error, coding error etc. Furthermore, no uniform theoretical framework for non-sampling errors has been widely accepted by the statistical or survey methods research communities. Thus, while classical sampling theory is based on well-established and uniquely defined sampling distributions (basically relating to all possible samples which could be selected from a given population in a given way), non-sampling error theories are, of necessity, model-based. Since there is an infinity of possible models to choose from, this fact in itself hinders the development of a single theory. Furthermore measurement errors in general, and non-sampling errors, in particular, are notoriously elusive in character, difficult to define and to model and extremely expensive to evaluate. In general no simple modification to survey design will ensure the possibility of evaluating even the simplest types of non-sampling errors and special evaluation studies (based on record checks, re-interview, split ballot designs etc.) have to be carried out in order to assess them.

Despite the difficulties, it must be emphasized that over the last four decades, increasing attention has been given by statisticians and by sample survey methodologists to the problems of non-sampling errors. At least the awareness of their existence and of their major contribution to the overall survey error can no longer be denied. In fact, vigorous attempts to model and to evaluate non-sampling errors have become major activities of advanced statistical agencies and of survey organizations. The theoretical developments of survey sampling in recent years and, in particular, the continuous development of model-based estimation and inference - either together with design-based considerations or in contrast to them - can be related to non-sampling error. In fact one interpretation of the models employed in model-based inference is that they represent errors of measurement or non-sampling errors.

As pointed out already, the theory and practice of non-sampling errors relate to a variety of aspects of the survey process. In a similar way to the practice with respect to sampling errors, we shall be most interested in those aspects of the survey process which can affect non-sampling errors and which can be manipulated. Although there could be great interest in the ways in which response errors arise and how respondents' characteristics might effect them, there is usually very little that can be done about these characteristics, in order to reduce the non-sampling errors. On the other hand, other aspects of the survey process, such as the method of data collection, questionnaire design, interviewer characteristics and their training can be controlled to a certain extent and are known to have in many cases a considerable effect on non-response errors.

For sampling errors it is well known that they are determined by a relative few sample design features, such as sample size, stratification level and degree of clustering, in a simple functional relationship. On the other hand, the number of design factors which effect non-sampling errors is considerable and the relationships between these factors and the errors is very complex and difficult to determine. We shall only be able to deal with a selected few of these - primarily those relating to the questionnaire or other methods of data collection. It must be borne in mind that whatever factors we choose to deal with, there will always remain a large number of additional factors - either uncontrollable or which are too expensive to deal with - which are not explored.

At this seminar we shall be dealing exclusively with non-sampling errors and within this topic with those errors which can be reduced by the manipulation of survey design aspects - primarily, but not only, the questionnaire. However, before leaving the topic of the relationships and comparisons between sampling and non-sampling errors, it must be remembered and emphasized that in the final analysis the survey designer has to worry about total error. Its breakdown into components of error and the source of each component are the concern of the survey designer only insofar as he can manipulate them and thereby effect the various components of error. Usually, the factors effecting sampling errors and non-sampling errors will differ and the feasibility of using different combinations of levels of these will results in different levels of expenditures. The integral design of surveys to reduce as far as possible total survey error has been treated only to a limited degree. It is usually dealt with by the judicious allocation of resources between the various design factors effecting the components of errors in a way which achieves an efficient balance between sampling and non-sampling errors. Some simple examples of this can be found in Nathan (1973), where length of recall period was determined in order to balance between the reduction in sampling error resulting from increasing the recall period and the accompanying increase in response error, due to recall. Further examples of work on total survey design, in the context of health surveys, is found in Andersen, Kasper, Frankel et al. (1979).

In the following, the theories of response error and its relationship to other sources of error will be reviewed and we shall consider the balancing of sampling errors and non-sampling errors in total survey design. In order to attempt to systemize questionnaire design and other aspects of the survey design, a theory of questionnaire design effects is proposed. The theory considers a universe of interchangeable design options, from which, in general, a single design option is selected. The relationships between questionnaire design effects and recent research in cognitive aspects of survey methodology are also considered. Specific methods of response error evaluation will be reviewed, with examples. These include macro-methods, such as: comparisons with data from administrative sources and other surveys, internal consistency checks, misclassification models, and split-ballot experiments; and micro-methods, such as: re-interview, administrative record checks and dual system and multiplicity methods. We shall consider examples of the application of some of these methods to the improvement of the design of questionnaires and other survey instruments.

1. RESPONSE ERROR THEORY AND ITS RELATIONSHIP TO OTHER SOURCES OF ERROR.

As mentioned above, the theory of non-sampling error and of response error, in particular, did not develop in the same way as the general theory of sampling. The difference is primarily in that a variety of different theories and models have been proposed for describing non-sampling errors, whereas the treatment of sampling error has been far more unified. In the following, we shall consider the more central and well-accepted theories, but his does not imply that alternative theories are unacceptable. However they are often suited to special situations or to specific design factors. An idea of the broad range of work in theory and practice of non-response errors can be found by a perusal of the bibliography on non-sampling error, prepared by Dalenius (1977). It includes some 1500 entries up to about 1975 - mostly from 1950 and onward.

Indeed, the major statistical development of response error theory can be traced to the early fifties in the pioneering work of a small group of statisticians working at the U.S. Bureau of Census. They included: Morris Hansen, William Hurwitz, Max Bershad, William Pritzker, Eli Marks and others. They were involved primarily with the decennial censuses of population and housing. The U.S. censuses had evolved, mostly in the thirties and forties, from basic counts of persons, households and dwellings to a fundamental detailed data base on a variety of substantive issues. These including detailed demographic data (such as household composition and fertility data), educational attainment, labour force characteristics, housing conditions and ownership of consumer durables. In addition to serious problems of undercoverage in the census - a type of non-sampling error which we shall not deal with here - the more detailed and difficult questions of the census were thought to suffer from serious response errors.

The theory developed at the Bureau of Census was aimed at describing the response error mechanism, primarily as it relates to the influence of the interviewer (or enumerator). Until the 1950 census all census information was obtained by face-to-face interviews with at least one member of each household and it was thought that the influence of the enumerator might be considerable. The theory was closely connected with the practical aspects of census taking and was proposed as a basis for the design and analysis of the first large-scale evaluation operation ever taken - the Post-Enumeration Survey programme of the 1950 Census.

1.1 The basic theory of response error.

The fundamental idea of the theory was set out in Hansen, Hurwitz and Bershad (1961). It considers the decomposition of the response provided by a respondent for a given question under a given set of conditions into a bias term and into a response deviation. Thus the total mean square error of the survey estimate is decomposed into a squared bias term and into a variance term, which is in turn separated into sampling variance and response variance. In order to clarify the basic concepts, we shall require some notation. At this point it should be mentioned that the theory of non-sampling errors, in general, and of response errors, in particular, are notorious in their lack of uniform notations and definitions. We shall use here the notation used by Hansen, Hurwitz and Bershad (1961), which relates to a dichotomous qualitative variable, i.e. to the estimate of a proportion. The extension to quantitative variables is straightforward.

Let U_j (j=1,...,N) be the "true" values (or "desired measures) of the dichotomous variables so that:

$$\bar{U} = \frac{1}{N} \sum_{j=1}^{N} U_j \qquad (1.1)$$

is the the "true" or "desired" proportion to be measured. In practice U_j will be unknown, but it is important to conceptualize these unknown "true" values, even if they cannot be exactly defined.

Under a given set of survey conditions, we consider a conceptual series of trials or repetitions of the survey process. In practice only a single trial (or survey) will be carried out, but we are interested in the behaviour of the survey estimates as determined by their distribution over this series of trials. In concept this is similar to the situation with respect to sampling distributions, where we observe a single sample but are interested in the hypothetical distribution over all possible samples. For the sake of simplicity, we assume that the data are collected by a an equal probability sample of size n (or alternatively by a census, i.e. with n=N). Let x_{jt} be the observed value for the j-th unit at the t-th trial. The survey estimate at the t-th trial is then:

$$p_t = \frac{1}{n} \sum_{j=1}^{n} x_{jt} \qquad (1.2)$$

Taking the expectation of p over all samples and all trials, we obtain the expected value of the survey estimate:

$$P = E(p_t) \ , \qquad (1.3)$$

which is not necessarily equal to the true value, $\bar{U}$. Thus the bias of p_t as an estimator of $\bar{U}$ is:

$$B = P - \bar{U} \ , \qquad (1.4)$$

its total variance is:

$$\sigma^2_{p_t} = E(p_t - P)^2 \qquad (1.5)$$

and the mean square error is:

$$MSE = E(p_t - \bar{U})^2 = \sigma^2_{p_t} + B^2 \ . \qquad (1.6)$$

In order to separate the total variance into a sampling and response error component, we consider the conditional expectations of the observations, over the probability space of the trials, for a fixed unit. We denote the conditional expectation of the observed value over all trials (i.e. over all measurements for that unit under the same conditions) for the j-th unit as:

$$E_j x_{jt} = P_j \qquad (1.7)$$

and its average over the sample of units as:

$$p = \frac{1}{n} \sum_{j=1}^{n} P_j \ . \qquad (1.8)$$

The total variance can then be decomposed as follows:

$$\begin{aligned} \sigma^2_{p_t} &= E(p_t - P)^2 \\ &= E(p_t - p)^2 + 2E(p_t - p)(p - P) + E(p - P)^2 \ . \end{aligned} \qquad (1.9)$$

We define the first term in (1.9) as the response variance of p_t and note that it can be written as the variance of the sample average of the response deviations, i.e:

$$\sigma^2_{d_t} = E(p_t - p)^2 = E(\bar{d}_t)^2 \ , \qquad (1.10)$$

where:

$$\bar{d}_t = \frac{1}{n} \sum_{j=1}^{n} \bar{d}_{jt} \ . \qquad (1.11)$$

The second term of (1.9) is the covariance between sampling and response deviations and vanishes in the case of a complete census. In the case of sample surveys it will not necessarily vanish, but its effect will, in general, be trivial and we shall ignore it in the following.

The final term of (1.9) is the classical sampling variance of p_t, defined as:

$$\sigma^2_p = E(p - P)^2 = E\left[\frac{1}{n} \sum_{j=1}^{n} (P_j - P)^2 \right] \ . \qquad (1.12)$$

Note that in the case of a census this component vanishes, since p=P, but that in general it will not depend only on n and on P, as when response error is not taken into account, since the values of P_j will, in general, lie between zero and one.

The response variance (1.10) can be further decomposed into a simple response variance and a correlated component, as follows:

$$\sigma^2_{\bar{d}_t} = E(\bar{d}_t)^2 = \frac{1}{n}\sigma^2_d + \frac{n-1}{n}\rho\sigma^2_d$$
$$= \frac{1}{n}\sigma^2_d [1 + \rho(n-1)] , \qquad (1.13)$$

where the variance of individual response deviations over all trials:

$$\sigma^2_d = E(d^2_{jt}) = \frac{1}{N}\sum_{j=1}^{N} P_j(1-P_j) \qquad (1.14)$$

is defined as the simple response variance and

$$\rho = \frac{E(d_{jt}d_{kt})}{\sigma^2_d} \quad (\text{for } j \neq k) \qquad (1.15)$$

as the intraclass response correlation between response deviations within a trial.

It is easily seen from (1.13) that even a small intraclass correlation may have a substantial impact on the response variance. Thus if $\rho = .01$ and the sample size is 3,000, the contribution of the correlated component - $\sigma^2_d \rho(n-1)/n$ is thirty times as large as that of the simple response variance - σ^2_d/n.

The relationship between the simple response variance and the sampling variance can be seen if we denote the population variance of the expected unit proportions, P_j, similarly to (1.14), by:

$$\sigma^2_S = \frac{1}{N}\sum_{j=1}^{N} (P_j-P)^2 . \qquad (1.16)$$

Then we have:

$$PQ = \sigma^2_d + \sigma^2_S , \qquad (1.17)$$

where $Q = 1-P$.

Note that for simple random sampling without replacement the sampling variance, defined by (1.12) is:

$$\sigma^2_p = \frac{N-n}{N-1}\frac{\sigma^2_S}{n} \qquad (1.18)$$

and for simple random sampling with replacement by:

$$\sigma^2_p = \frac{\sigma^2_S}{n} . \qquad (1.19)$$

Thus, in the latter case, the total uncorrelated variance - the sum of sampling variance and of simple response variance - is simply:

$$\frac{\sigma^2_S}{n} + \frac{\sigma^2_d}{n} = \frac{PQ}{n} , \qquad (1.20)$$

This implies that if response deviations are uncorrelated, then the total variance - sampling and response variance - does not exceed PQ/n. The ratio of simple response variance to the total variance, PQ is frequently used as a measure of the relative influence of simple response variance. It is often termed the "Index of Inconsistency", , originally proposed by Pritzker and Hanson (1962), and is defined as:

$$I = \frac{\sigma_d^2}{\sigma_d^2+\sigma_s^2} = \frac{\sigma_d^2}{PQ} . \qquad (1.21)$$

It indicates the relationship between response error and sampling error and in the U.S. 1950 Census of population its value was found to be of the order of 3-10% for most items which could be measured reliably, such as age, while it rose to about 50% for difficult items, such as condition of housing and unemployment.

However, in practice, at least for a survey carried out by interviewers, the correlated response error component to (1.13) may well be the dominant part of response error. This contribution depends not only on the inherent error structure of the population, but also on the survey design. Thus if this component is primarily due to the correlation between response deviations for the same interviewer (rather than within higher-level operators, such as crew-leaders), then the size of the interviewer assignment determines the contribution, rather than the total sample size, in (1.13) and thereby the influence of correlated response errors is reduced considerably.

1.2 Methods to estimate response errors.

The above theory indicates several methods which can be used to estimate the response error and its components. The most commonly used is the method of replication. In its simplest form this assumes that the whole survey process can be replicated under exactly the same conditions, but in as far as possible, by a different set of enumerators, crew-leaders and processors. If p_1 and p_2 are the estimates of the same proportion obtained by the two replications then the following statistic can be used as an estimator of response variance:

$$C = \frac{(p_1 - p_2)^2}{2} . \qquad (1.22)$$

It is easily seen that if the expected proportions, P_1 and P_2, are equal, then approximately:

$$E(C) = \frac{\sigma_{d_1}^2 + \sigma_{d_2}^2 - 2\mathrm{cov}(\bar{d}_1,\bar{d}_2)}{2} . \qquad (1.23)$$

Thus if $\sigma_{d_1}^2 = \sigma_{d_2}^2$ and the covariance term in (1.23) is close to zero, then C is an unbiased estimator of the response variance (1.14). The difficulty with this estimator is that it is based on a single degree of freedom and also that the assumption that the correlation vanishes or is small may be a serious limitation. In particular if the replication is close in time to the original interview, the respondent may remember his answer and his response deviations will be correlated. On the other hand if there is a long time between replications, the assumption of equal expected values, P_1 and P_2, may no longer hold.

The method of interpenetration, first used by Mahalonobis (1964) in India, overcomes most of these obstacles. Under this method, the sample selected from two contiguous strata is randomly divided into two sub-groups, each of which is assigned at random to one of a pair of interviewers. Similarly crew leaders and processors are randomly assigned enumerator loads within each of the enlarged strata. Let:

$$D = \frac{\bar{n} \sum_{i=1}^{M} p_i q_i}{(\bar{n} - 1) M\bar{n}} , \qquad (1.24)$$

where p_i is the estimated proportion in the i-th interviewer assignment (1=1,...,M) and $q_i = (1-p_i)$; $\bar{n}$ is the number of units in an interviewer's assignment and $\bar{n}$ is the number of units in a stratum. Let C be defined by (1.22), where now p_1 is the estimate obtained from one set of enumerators (one per enlarged strata) and p_2 is that obtained from the other. Then E(C-D) is an approximately unbiased estimator of $\rho\sigma_d^2$, which is the dominant term of the total response variance (1.13).

1.3 Some results of evaluation studies.

These methods have been widely used, primarily in the evaluation of censuses. An estimator of the index of inconsistency (1.21), which is often used for the case of replication, is given by:

$$\hat{I} = \frac{G}{p_1(1-p_1)+p_2(1-p_2)} , \qquad (1.25)$$

where G is the "Gross Difference Rate", defined as the proportion of cases classified differently in the two trials, i.e.:

$$G = \frac{1}{n} \sum_{j=1}^{n} (x_{j1} - x_{j2})^2 . \qquad (1.25)$$

The high values of this measure obtained from the evaluation studies of the 1950 census in the U.S. - see U.S. Bureau of Census (1960) - were the major reason for introducing sampling in the census for the more complicated questions, since the 1960 census. The reasoning was that with such high response errors, a more efficient allocation of effort would be to reduce response error by a smaller but better trained enumerator staff, even at the expense of the increase in sampling error. For example, in exhibit 1.1 - from Hansen, Hurwitz and Bershad (1961) - estimated response variances for proportions are compared to sampling variances for a 25% sample (as used in the 1960 Census) with respect to an area of 6,500 population (1625 sample persons. From column (4) it is seen that response variances (which in this case represent primarily interviewer variability) are as much as four times as large as sampling variances). An additional recommendation based on these evaluation studies was to use self-enumeration as far as possible in order to reduce the very large effect of the correlated component of response error.

EXHIBIT 1.1

Table I *Effect of Interviewer Variability for a Census of an Area of 6,500 Population, Based on an Experimental Study in the 1950 Census*

Characteristic	Proportion of population having the characteristic P	σ_d^2 $\times 10^{-4}$	$\frac{PQ}{1625}$ $\times 10^{-4}$	$\sigma_d^2 / \frac{PQ}{1625}$ [1]
	(1)	(2)	(3)	(4)
Total population	1.000			
Nativity				
Native white	.905	.31	.53	.6
Foreign born white	.036	.15	.21	.7
Residence 1 year earlier [2]				
Same house	.803	1.88	.97	1.9
Different house, same county	.114	1.48	.62	2.3
Different county or abroad	.047	.28	.28	1.0
Age, Males				
Under 5 years	.057	.03	.33	.1
15 and older	.353	.13	1.41	.1
35 and older	.207	.21	1.01	.2
55 and older	.084	.03	.47	.1
Highest grade of school attended[3]				
Grade 5 or over	.540	.04	1.53	.03
Grade 9 or over	.321	.73	1.34	.5
Grade 13 or over	.068	1.16	39	2.9
Income [4]				
Wage and Salary				
None	.348	1.58	1.40	1.1
Under $2,500	.184	1.36	.92	1.4
$2,500 and over	.165	.27	.85	.3
From own business				
None	.636	2.51	1.42	1.7
Under $2,500	.040	.29	.24	1.2
$2,500 and over	.020	.07	.12	.6
Other				
None	.576	4.48	1.50	2.9
Under $2,500	.114	2.66	.62	4.2
$2,500 and over	.005	[6]	.03	
Major occupation group [5]				
Craftsman, foremen, etc., males	.061	.15	.35	.4
Farmers & farm managers, males	.022	.15	.13	1.1
Farm, unpaid family workers, male	.003	.01	.02	.8
Farm laborers, paid, male	.005	.01	.03	.2
Industry group [5]				
Manufacturing	.152	.41	.79	.5

[1] Computed from columns (2) and (3) before rounding.
[2] Persons 1 year of age and over.
[3] Persons 25 years of age and over.
[4] Persons 14 years of age and over.
[5] Employed workers 14 years of age and over.
[6] Estimate of variance negative.

Source: Hansen, Hurwitz and Bershad (1961).

Similar studies were carried out in several other countries in the 1960's with similar results. Thus a combination of replication and interpenetration methods was used by Statistics Canada in its evaluation of the 1961 Census together with sophisticated methods of estimation developed by Fellegi (1964). Studies based primarily on replication methods were carried out in Israel for the evaluation of the 1961 Census - see Kantorowitz (1969). In exhibits 1.2-1.3 examples of the results are shown for a census-register comparison and for a re-interview survey from the Israeli Census post-enumeration studies. Both the gross difference rates and the estimates of indices of inconsistency are seen to be very large for some characteristics (aggregate measures for polytomous variables, based on averaging over the categories, are used here).

EXHIBIT 1.2

TABLE D/2.- CENSUS-REGISTER COMPARISON - OVERALL AND AVERAGE INDICES FOR THE DISTRIBUTIONS

The Distribution	Average Indices (Percentages)				Overall Indices (Percentages)	
	Gross Difference Rate $\bar{g}$	Index of Inconsistency $\bar{I}$	Net Difference Rate (Absolute) $\overline{\lvert\beta\rvert}$	Percent Identically Distributed (Relative to Register) $\bar{r}$	Gross Difference Rate g^*	Net Difference Rate $\lvert\beta^*\rvert$
Population group	0.1	0.5	0.0	99.8	0.1	0.0
Sex	1.2	2.4	0.3	98.8	1.2	0.3
Marital status						
TOTAL	2.7	6.4	0.5	96.2	3.9	1.4
Males	2.9	6.6	1.3	96.2	4.0	1.6
Females	2.5	6.1	0.5	96.3	3.7	1.2
Country of birth						
TOTAL JEWS	0.9	5.7	0.3	94.9	4.8	1.4
Males	1.0	5.8	0.3	94.9	5.0	1.5
Females	0.8	5.5	0.2	94.9	4.5	1.3
Year of immigration						
TOTAL BORN ABROAD	1.9	6.8	0.8	95.0	5.1	1.2
Born in Asia-Africa	2.4	7.6	0.6	94.6	5.5	1.0
Born in Europe-America	1.8	6.2	0.9	95.1	4.8	1.3
Year of birth						
TOTAL	1.6	11.9	0.2	89.9	11.1	..
Jews	1.5	11.9	0.1	88.9	..	..
Non-Jews	2.2	12.0	0.3	88.7	..	..
Month of birth	10.4	21.0	0.1	88.6	10.4	0.1

Source: Kantorowitz (1969).

EXHIBIT 1.3

TABLE D/3.- RE-INTERVIEW SURVEY - OVERALL AND AVERAGE INDICES FOR THE DISTRIBUTIONS OF "NUMBER OF ROOMS" AND OF "HOUSING DENSITY" (Percentages)

The Distribution	Average Indices					Overall Indices			
	Gross Difference Rate	Index of Inconsistency	Net Difference Rate	Net Shift Relative to Survey	Percent Identically Distributed Relative to Survey	Unweighted Gross Difference Rate	Unweighted Net Difference Rate	Weighted Gross Difference Rate	Weighted Net Difference Rate
	$\bar{\varepsilon}$	$\bar{I}$	$\overline{\lvert\beta\rvert}$	$\overline{\lvert\beta\rvert}/p_{t'}$	$\bar{r}$	ε^*	β^*	ε_w^*	β^*_w
Number of Rooms									
ALL HOUSEHOLDS	15.1	36.5	2.1	6.9	75.5	24.9	- 0.3	26.2	- 0.4
Urban	14.6	35.6	1.8	5.7	75.9	24.3	+ 0.2	25.2	- 0.2
Rural	20.0	46.8	4.8	16.4	70.8	30.4	- 4.0	34.3	- 2.6
Self-enumeration	15.4	26.4	1.0	4.1	75.5	25.8	- 2.5	28.0	- 2.8
By enumerator	14.8	35.7	3.7	11.6	75.9	23.8	+ 2.6	23.8	+ 2.6
Housing Density									
ALL HOUSEHOLDS	7.0	23.3	0.8	3.7	81.1	19.2	+ 0.8	53.0	+ 2.5
Urban	6.9	22.8	0.8	4.1	81.4	18.8	- 0.6	54.6	+ 1.3
Rural	7.9	28.4	1.9	15.4	77.7	22.2	+13.9	38.9	+13.9
Household heads: Israel born	7.4	27.7	1.1	13.9	78.7	20.3	- 2.9	82.6	- 4.3
Born in Asia-Africa	8.9	28.6	1.2	7.7	77.6	22.0	+ 1.8	70.9	+ 4.0
Born in Europe-America	7.5	21.7	0.9	3.9	83.5	17.4	+ 1.0	38.2	+ 2.9

TABLE D/4.- RE-INTERVIEW SURVEY - OVERALL AND AVERAGE INDICES FOR THE DISTRIBUTIONS OF "OWNERSHIP OF DWELLING" (Percentages)

The Distribution	Average Indices					Overall Indices			
	Gross Difference Rate	Index of Inconsistency	Net Difference Rate	Net Shift Relative to Survey	Percent Identically Distributed Relative to Survey	Unweighted Gross Difference Rate	Unweighted Net Difference Rate	Weighted Gross Difference Rate	Weighted Net Difference Rate
	$\bar{\varepsilon}$	$\bar{I}$	$\overline{\lvert\beta\rvert}$	$\overline{\lvert\beta\rvert}/p_{t'}$	$\bar{r}$	ε^*	β^*	ε_w^*	β^*_w
ALL HOUSEHOLDS	12.6	29.5	3.1	7.6	84.5	14.6	+3.0	16.4	+2.7
Urban	11.8	27.5	3.6	8.1	85.4	13.9	+3.5	15.4	+3.3
Rural	21.2	44.4	1.0	1.5	78.5	21.7	-1.4	26.1	-2.9
Household heads: Israel born	6.5	18.3	2.5	6.6	89.2	9.7	+6.5	9.7	+6.5
Born in Asia-Africa	17.8	40.1	5.5	25.1	77.3	21.3	+1.9	26.9	+0.9
Born in Europe-America	10.8	25.9	2.6	6.1	86.2	11.8	+3.1	11.8	+3.1

Source: Kantorowitz (1969).

2. SYSTEMIZATION OF QUESTIONNAIRE DESIGN AND THE THEORY OF QUESTIONNAIRE DESIGN EFFECTS.

The theories of response error, proposed in the previous section, and their application to the evaluation of censuses and surveys have continued to develop over the past three decades. This can be seen from the detailed bibliography of Dalenius (1977). In parallel, but practically divorced from the theory of response error, an enormous amount of work has been going on in the survey research area on the improvement of survey design in general, and of questionnaire design, in particular. Here we shall concentrate on questionnaire design, but in this field too a vast body of research is available in articles and books - see for instance Belson (1981), Bradburn, Sudman et al. (1979) and Schuman and Presser (1981). For the most part of it is based on well-designed and analyzed empirical studies. However, it hardly relates to any theory and is to a large degree ad-hoc. This does not imply that the results are not extremely useful and there is no doubt that the practice of survey research has improved considerably, as a result of these studies.

We shall discuss, in the following, an attempt to reach decisions about questionnaire design on a theoretical base, in conjunction with experimentation. The method of evaluation considered is the frequently used split-ballot experiment, in which alternative questionnaire designs (or other variants of the survey procedure) are assigned at random to sub-samples by an experimental design. We consider the case were the experiment is embedded in the sample survey itself. A variety of paradigms have been used to draw inferences about the questionnaire or the survey procedure from such experiments - see for example - Feinberg and Tanur (1989).

2.1 Alternative approaches to the analysis of survey design effects.

In this section we discuss two alternative approaches which can be used to analyze questionnaire design (or other survey design) effects from split-ballot experiment data. The formulation is in terms of effects of questionnaire versions, but is just as applicable to other survey design features, such as mode of data collection (see e.g. section 4.1). We term the two alternative approaches: the **descriptive analysis approach**, which views each questionnaire version as eliciting its unique response distribution; and the **response variance approach**, which assumes that the questionnaire versions are randomly selected from a universe of exchangeable versions. After describing each approach, I shall discuss the main advantages and disadvantages of each approach and suggest a compromise which could be used to enhance the use of embedded experimentation in surveys. A detailed model developed for a multi-factor case is described. Finally an example from a split-ballot experiment, carried out in connection with the U.S. National Center for Health Statistics Health Interview Survey, and of the results of the analysis, under each of the two approaches, will be given.

Overall, we are considering the usual survey situation, where, potentially at least, the survey designer could choose from among a variety of alternative questionnaire versions to obtain the information required. The **descriptive analysis approach** views each questionnaire version as eliciting its own unique response distribution. In the conventional survey design, just one questionnaire version is selected (not always by an explicitly defined decision process) and the sample reflects the distribution of responses specific to this version. In a split-ballot experiment, several versions are used and the sample responses obtained for each of the versions represent the different distributions. This approach can be formalized by a simple fixed effects model, as follows:

$$Y_{ij} = \mu_i + e_{ij} = \mu + \alpha_i + e_{ij} \ , \qquad (2.1)$$

where μ_i is the mean response of the i-th questionnaire version,
e_{ij} is the random response deviation, with:

$$E(e_{ij}) = 0 \ , \text{ and: } V(e_{ij}) = \sigma_e^2 .$$

This model assumes that the response to each questionnaire version can be expressed as a random variable, Y_{ij}, which is a linear function of the expected response over all population units, μ_i, specific to the questionnaire version, plus a random error component due to sampling population units. Survey researchers, for example, use this approach to investigate psychological theories of questionnaire design error effects. The mean, μ_i, can be further broken up into an overall mean (over all versions), μ, and into a fixed effect of the questionnaire version, α_i, which measures its deviation from the overall mean.

Note that we do not need to assume that the overall mean, μ, is the "true" value. Basically the descriptive analysis just describes the differences between questionnaire versions.

The **response variance paradigm** assumes that the questionnaire versions are random selections from a universe of exchangeable versions. This implies that the survey designer can identify a set of potential questionnaire designs which he regards as exchangeable or equally acceptable. Thus the selection of a version at random for use in the survey will result, on average, in the same distribution of response. This can be expressed as a random effects model,as follows:

$$Y_{ij} = \mu + a_i + e_{ij} \ , \qquad (2.2)$$

where a_i is a random variable associated with the i-th questionnaire version,
e_{ij} is the residual random response deviation, with:

$$E(a_i) = E(e_{ij}) = 0 \ ; \ V(a_i) = \sigma_a^2 \ ; \ V(e_{ij}) = \sigma_e^2 \ .$$

This model assumes that the survey response can be expressed as a linear function of the expected responses over all questionnaire versions and population units, μ, plus two independent random error components - one due to the sampling of questionnaire versions, a_i, and the other, e_{ij}, due to the sampling of population units. The proposed paradigm is in the tradition of the classic test theory of Lord and Novick (1968) and the response error models developed by Hansen, Hurwitz and Bershad (1961).

In comparing the two approaches, we note that the descriptive analysis paradigm assumes only that the responses to the same version come from some distribution, which may vary from one version to another. It can thus measure differences in response bias by considering each of these distributions to have its own mean, none of which are necessarily the "true" value. However, in some cases, supplementary information, such as a validation study or administrative records, will allow us to decide which is the version which entails less bias. This paradigm is thus, most suitable for analysis at the design stage, when a choice of a single design may have to be made between alternative questionnaire designs or if we wish to identify a sub-universe of "best" designs. However it can also be useful at the analysis stage in providing insight into the underlying cognitive aspects of the different questionnaire designs.

The response variance paradigm, on the other hand, concentrates on the variation between questionnaire versions as a source of error, disregarding the bias. By using split-ballot experiments this approach allows the routine evaluation of this source of response variation as an integral part of the survey process, just as is generally done for sampling error. In this the response variance paradigm is most useful at the analysis stage. However, this evaluation of variance components leads to the possibility of improving the survey estimates, at the design stage, by the efficient allocation of resources to balance the sample size and the number of versions used in the survey, taking into account the increased cost of using more than a single version. It should be pointed out that the estimate of questionnaire-design variance obtained under this paradigm can also be useful, under the descriptive analysis approach, if it is considered as a standardized measure of variation between questionnaire versions.

Each of the approaches has its drawbacks. Thus the descriptive analysis paradigm not only does not evaluate the response variation between versions, but it will also usually lead to the choice of a single questionnaire version, under the illusion that it is the "best" one. The main difficulty of the response variance paradigm lies in the difficulty encountered when trying to define the exchangeable universe of versions. Furthermore this paradigm ignores the problems of bias involved in the choice of a single questionnaire design, or even of a sample of them, from this universe.

The solution may lie in some sort of synthesis between the two approaches. Thus we might be able to divide the overall universe of questionnaire versions into sub-sets, each of which could be considered as a sub-universe of exchangeable versions. The descriptive analysis could then be applied to the differences between the means of these sub-universes, as evaluated in an exploratory study. If external information were available, the sub-universe with the smallest bias could be selected. Then the response variance paradigm would be used to analyze the variation within this sub-universe, on the basis of a split-ballot experiment embedded in the survey itself. To take an extreme example the sub-universes may be defined in terms of different question formulations, the versions within each sub-universe differing only with respect to the colour of the questionnaire.

2.2 A multi-factor model for the response variance approach.

In fact both basic questionnaire design error models formulated above relate to a single factor. Thus under the response variance paradigm only a single questionnaire design variance is used for measuring the variation between questionnaires - see for instance Nathan and Sirken (1986) for an application. In most cases, the universe of questionnaire designs will have a more complex structure and several design factors may be involved in determining differences between questionnaires. For instance, question layout, typography or colour may each be considered as a different factor and the separate contribution of each factor to the questionnaire design variance may be required. A simple multi-factor extension of the basic model (2.2) was considered in Nathan and Sirken (1987). For the application to the example in section 2.3, a somewhat different model to deal with more than a single factor is proposed in Sirken, Nathan and Thornberry (1989). It is formulated in terms of two factors, but can easily be extended to more than two factors. Basically it is a univariate finite-population balanced random-effects two factor crossed linear model - without interaction.

The assumptions required are as follows:

1. Simple random sampling of population units.

2. Simple random selection of questionnaire design versions.

3. Balanced random allocation of units to the levels of each of the two factors.

4. No interaction effect between factors.

The values of the quantitative variable of interest, Y, are assumed to be potentially measurable for each of the units in a sample of size n, by means of each of a set of rq possible questionnaire designs. These represent all possible combinations of r levels of a random effect, A, and of q levels of a random effect, B.

The n units are randomly allocated to these rq combinations of levels, so that m=n/rq units are allocated to each combination (m is assumed to be integral). Thus the model relates to a structured set of measurements $\{Y_{ijk}: 1=1,\ldots,r;\ j=1,\ldots,q;\ k=1,\ldots,m\}$, where i denotes the level of factor A; j denotes the level of factor B; and k denotes the unit. We consider the following additive model (without interaction), which is a simple extension of the model (2.2):

$$Y_{ijk} = \mu + a_i + b_j + e_{ijk} , \qquad (2.3)$$
$$(i=1,\ldots r;\ j=1,\ldots,q;\ k=1,\ldots,m)$$

Here $\{a_i\}$, $\{b_j\}$ and $\{e_{ijk}\}$ are independent random variables with:

$$E(a_i) = E(b_j) = E(e_{ijk}) = 0;\ V(e_{ijk}) = \sigma_e^2 . \qquad (2.4)$$

The model implies that the response obtained for the unit (i,j,k) is a random variable whose mean over all units and over all questionnaire designs is μ and that the residual can be decomposed into three independent random components: one, a_i, specific only to the level of the random factor A; the second, b_j, specific only to the level of the random factor B; and the error (residual) term, e_{ijk}, which depends also on the unit's index, k.

The r levels at which the factor A is measured are assumed to be selected at random from a larger set of R possible levels, so that the random variables $\{a_i: i=1,\ldots,r\}$ are considered as a simple random sample without replacement from a finite population of R fixed values $\{\alpha_1,\ldots,\alpha_R\}$, with $\Sigma_i\alpha_i = 0$. Thus:

$$V(a_i) = \sigma_a^2 = \frac{1}{R}\sum_{i=1}^{R} \alpha_i^2 . \qquad (2.5)$$

Similarly, the random variables $\{b_j: i=1,\ldots,q\}$ are assumed to be selected as a simple random sample without replacement from the finite population of Q fixed values $\{\beta_j,\ldots,\beta_Q\}$, with $\Sigma_j\beta_j = 0$. This implies:

$$V(b_j) = \sigma_b^2 = \frac{1}{Q}\sum_{i=1}^{Q} \beta_j^2 . \qquad (2.6)$$

It should be noted that the sets of random variables $\{a_i\}$, $\{b_j\}$ and $\{e_{ijk}\}$ are mutually independent, but that the random variables of the first two sets are not independent within the sets, due to the fact that sampling of the factor levels is assumed to be without replacement.

The overall mean is then:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{q}\sum_{k=1}^{m} Y_{ijk} , \qquad (2.7)$$

with expected value: $E(\bar{Y}) = \mu$ and variance:

$$V(\bar{Y}) = \frac{R-r}{R-1}\frac{\sigma_a^2}{r} + \frac{Q-q}{Q-1}\frac{\sigma_b^2}{q} + \frac{\sigma_e^2}{n} \quad . \tag{2.8}$$

it should be noted that the above can cover a variety of situations as special cases. Thus, if the number of levels of one of the factors, say B, is very large (e.g. $Q \rightarrow \infty$), the appropriate finite population factor $(Q-q)/(Q-1)$ in (2.8) can be replaced by 1. In this case the model reverts to the standard random effects model (with respect to that factor), similar to (2.2). Thus if the factor B is that of question order which has a very large number of possible levels, the factor may be considered as a standard random factor, for which we wish to infer about a virtually infinite universe of levels on the basis of a small number of observed levels, q, of the factor.

Another special case is where the number of levels of a factor is so small that all of them can be measured, i.e. $R = r$. This basically transforms the the random effect into a fixed effect, similar to model (2.1). In this case the finite population factor, $(R-r)/(R-1)$ is zero, so that the term involving the variance of the factor A in (2.8) vanishes (since the universe of levels is exhausted).

Note that, although the variance term due to a fixed effect factor does not appear in the variance of the overall mean (2.8), the variance due to the factor in the general model, (2.3), can still be estimated. This is useful since it allows the application of the estimates of the model parameters to the important special case of the standard survey situation where only a single questionnaire design is used (i.e. $r = q = 1$). Note that in this case the variance of the mean in (2.8) becomes:

$$V(\bar{Y}) = \sigma_a^2 + \sigma_b^2 + \frac{\sigma_e^2}{n} \quad . \tag{2.9}$$

The estimation of the model parameters requires an experiment in which at least two levels of each factor are measured (i.e. $r,q \geq 2$). The estimates are obtained from standard analysis of variance methods for random effects models, e.g. by the method of moments. Thus we define the usual mean sums of squares, as follows:

$$MSA = \frac{qm}{r-1}\sum_i (\bar{Y}_{i..} - \bar{Y})^2$$

$$MSB = \frac{rm}{q-1}\sum_j (\bar{Y}_{.j.} - \bar{Y})^2$$

$$MSR = \frac{\sum_{ijk} (\bar{Y}_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y})^2}{rqm - r - q + 1} \quad , \tag{2.10}$$

where $\bar{Y}_{i..}$, $\bar{Y}_{.j.}$ and $\bar{Y}$ are the usual marginal and overall means. Then it is easy to show that unbiased estimators of the model parameters are obtained as:

$$\hat{\sigma}_a^2 = \frac{R-1}{qmR}(MSA - MSR)$$

$$\hat{\sigma}_b^2 = \frac{Q-1}{rmQ}(MSB - MSR)$$

$$\hat{\sigma}_e^2 = MSR \ . \qquad (2.11)$$

It should be noted that these estimators can take negative values. this would, in general, indicate that the factor does not significantly contribute to the questionnaire design variance.

For the case of categorical variables, the above model can easily be modified, by considering Y_{ijk} as an indicator variable, taking the values 1 or 0. However, then the set of variables $\{e_{ijk}\}$ can no longer be independent of the sets of variables $\{a_i\}$ and $\{b_j\}$. To see this set:

$$P_{ij} = \mu + a_i + b_j$$

Then conditional on P_{ij}, $\{Y_{ijk}\}$ are independent Bernoulli variables with expectation P_{ij} and:

$$E(e_{ijk}|P_{ij}) = 0; \ V(e_{ijk}|P_{ij}) = P_{ij}(1- P_{ij}) \ ,$$

where $E(P_{ij}) = \mu$ and $V(P_{ij}) = \sigma_b^2$.

However, unconditionally, $\{e_{ijk}\}$ still have expectation zero and variance σ_e^2.

2.3 Application to the analysis of response effects of ordering.

In order to analyze the response effects of question order and of response category order, a split ballot experiment was embedded in the design of the National Health Interview Survey, during the 3-month period, April-June 1988. The Health Interview Survey is a national survey designed to monitor the nation's health. It is based on a stratified multi-stage sample design in which face-to-face interviews are conducted continuously in weekly samples of 1,000 households.

The experiment investigated the effects of question order and response category order on responses to a battery of questions on knowledge and attitudes about AIDS. Four variants of the AIDS questionnaires - combining two orderings of the questions and two orderings of the response categories - were assigned for administration to randomly selected adults in the National Health Interview Survey households. The treatments were assigned by a balanced complete factorial design so that the sample design for each of the four sub-samples was essentially the same as that for the entire survey. This was attained by assigning households at random (systematically) to each of the four versions, without blocking on the interview assignments.

The AIDS knowledge and attitudes supplement contained a battery of 62 questions, but the split ballot experiment was limited to the subset of eight questions listed in exhibit 2.1. An example of the two question orderings with respect to question 24 - basically a reordering of the sub-questions within the main question on sources and effects of AIDS.

Initially, the descriptive analysis paradigm was applied to the data. The analysis was carried out only with respect to the modal category (which with only one exception, 23d, was the "correct" answer to the AIDS knowledge questions 23, 24, and 45). Each of the two factors - question order and response order - was treated separately, as if it were the only factor involved. Thus with respect to question order, the data from each set of two versions with the same question order (but different response orders) were pooled.

The differences between the proportions answering the modal category for the two pairs of versions were standardized. This was done by dividing each difference by its estimated standard error, based on the overall mean proportion and assuming simple random sampling (a t-statistic). For technical reasons, the design effects could not be incorporated into the analysis at this stage, but an initial evaluation of their effect on these differences shows that they are very small. Thus the standardized differences can be regarded as approximately standard normal deviates. The results are shown in tables 1a-1c and 2a-2c in exhibits 2.2-2.3.

As an example of this analysis for question order, note the difference for question 24c, on the likelihood of getting AIDS from eating in a restaurant where the cook has AIDS. From a high response of 35.4% answering "very unlikely", when this sub-question came immediately after the sub-question on sharing needles (24h) in version 2, the proportion is reduced to 27.9% in version 1, when asked following sub-questions on living or working near AIDS patients. The standardized difference here was 7.79 (see table 1) - the most significant of all differences found - indicating a strong context effect resulting from the ordering of sub-questions.

Thus it seems that respondents might be making relative assessments of likelihood. This would imply that they regard the possibility of getting AIDS from eating where the cook has AIDS as "very unlikely" more frequently, relative to getting AIDS from sharing needles and less frequently, relative to living or working near people with AIDS. Other significant differences for this question were found for 24a (living near a home or hospital for AIDS patients (lower in version 1, when appearing as the first sub-question than when appearing seventh after being asked about a range of potential risk factors). Similarly for question 24f (on sharing eating utensils with AIDS-infected persons) a significantly higher proportion answered "very unlikely" if this question came after the question on sharing needles then when it came before.

EXHIBIT 2.1

APPENDIX

SELECTED QUESTIONS FROM THE AIDS KNOWLEDGE AND ATTITUDE SUPPLEMENT: 1988 NATIONAL HEALTH INTERVIEW SURVEY[1]

21. How much would you say you know about AIDS --- a lot, some, a little, or nothing?

23. After I read each statement, tell me whether you think the statement is definitely true, probably true, probably false, definitely false, or you don't know if it is true or false.

a(h) AIDS can reduce the body's natural protection against disease.
b(i) AIDS is especially common in older people.
c(j) AIDS can damage the brain.
d(k) AIDS usually leads to heart disease.
e(l) AIDS is an infectious disease caused by a virus.
f(m) Teenagers cannot get AIDS.
g(n) AIDS leads to death.
h(a) A person can be infected with the AIDS virus and not have the disease AIDS.
i(b) Looking at a person is enough to tell if he or she has the AIDS virus.
j(c) ANY person with the AIDS virus can pass it on to someone else through sexual intercourse.
k(d) A person who has the AIDS virus can look and feel well and healthy.
l(e) A pregnant women who has the AIDS virus can give the AIDS virus to her baby.
m(f) There is a vaccine available to the public that protects a person from getting the AIDS virus.
n(g) There is no cure for AIDS at present.

24. After I read each statement, tell me if you think it is very likely, somewhat likely, somewhat unlikely, very unlikely, definitely not possible, or if you don't know how likely it is that a person will get AIDS or the AIDS virus infection that way.

How likely do you think it is that a person will get AIDS or the AIDS virus infection from --

a(g) living near a home or hospital for AIDS patients.
b(h) working near someone with the AIDS virus.
c(b) eating in a restaurant where the cook has the AIDS virus.
d(c) kissing - with exchange of saliva - a person who has the AIDS virus.
e(d) shaking hands, touching, or kissing on the cheek someone who has the AIDS virus.

---

[1]Question order for original NHIS version. Letter in parenthesis denotes order of question in alternate version. For example 23a(h) implies that the 23a in the original version was 23h in the alternate version.

Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.1 (cont.)

f(e) sharing plates, forks, or glasses with someone who has the AIDS virus.
g(f) using public toilets.
h(g) sharing needles for drug use with someone who has the AIDS virus.
i(i) being coughed on or sneezed on by someone who has the AIDS virus.
j(j) attending school with a child who has the AIDS virus.
k(k) mosquitoes or other insects.

45. Here are some methods people use to keep from getting the AIDS virus through sexual activity.

After I read each one, tell me whether you think it is very effective, somewhat effective, not at all effective, or if you don't know how effective it is in preventing getting the AIDS virus through sexual activity. How effective is ----

a(b) Using a diaphragm?
b(c) Using a condom?
c(d) Using a spermicidal jelly, foam or cream?
d(e) Having a vasectomy?
e(a) Two people who do not have the AIDS virus having sex only with each other?

46. What are your chances of having the AIDS virus; would you say high, medium, low or none?

47. What are your chances of getting the AIDS virus; would you say high, medium, low or none?

61. When Federal Public Health officials give information about AIDS, do you believe what they say or are you doubtful about the information they give? [believe them; doubtful; DK]

62. When they give advice about how to help keep from getting AIDS, do you believe their advice or are you doubtful about what they say? [believe them; doubtful, DK]

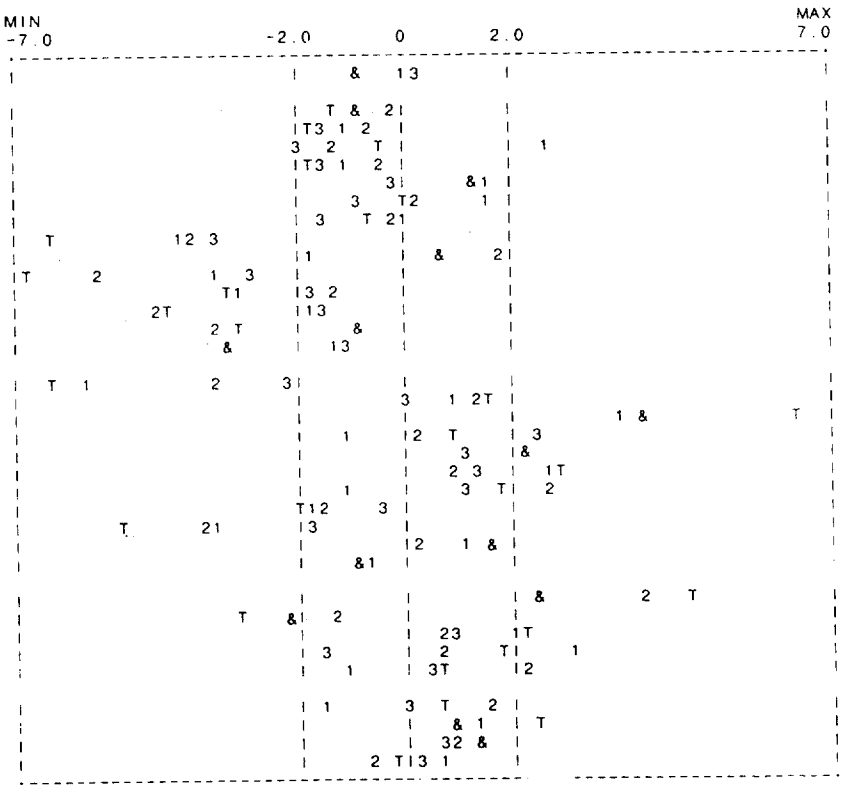
Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.2

Table 1a Standardized differences: question order by age group – modal frequency

Notes: Sign is positive if frequency is higher for sub-question appearing earlier
* Letter in parentheses denoted order in second version
& Denotes two (or more) values plotted in same position

Question No.*	Total "T"	18-29 "1"	30-49 "2"	50+ "3"
21	-0.89	-0.03	-0.86	0.15
23A(H)	-1.18	-0.75	-0.25	-0.38
23B(I)	-1.76	-1.09	-0.54	-1.79
23C(J)	-0.40	2.69	-1.38	-1.32
23D(K)	-1.74	-1.17	-0.50	-1.53
23E(L)	1.21	1.27	1.17	-0.20
23F(M)	0.09	1.56	0.18	-0.28
23G(N)	-0.72	0.14	-0.08	-1.75
23H(A)	-6.60	-4.14	-3.93	-3.71
23I(B)	0.78	-1.72	1.76	0.31
23J(C)	-7.05	-3.55	-5.77	-2.39
23K(D)	-3.25	-3.17	-1.25	-1.34
23L(E)	-4.44	-1.71	-4.51	-1.32
23M(F)	-3.09	-0.89	-3.56	-0.37
23N(G)	-3.23	-1.30	-3.26	-0.39
24A(G)	-6.63	-5.90	-3.45	-2.10
24B(H)	1.52	1.01	1.41	0.34
24C(B)	7.79	4.22	4.64	4.59
24D(C)	0.94	-0.98	0.36	2.51
24E(D)	2.25	0.34	2.27	1.12
24F(E)	2.91	2.62	1.01	1.34
24G(F)	1.87	-1.01	2.72	1.20
24H(A)	-2.05	-1.81	-1.54	-0.33
24I(I)	-5.34	-3.41	-3.72	-1.33
24J(J)	1.65	1.15	0.32	1.39
24K(K)	-0.75	-0.73	-0.78	0.29
45A(B)	5.43	2.56	4.46	2.51
45B(C)	-3.11	-2.15	-1.36	-2.11
45C(D)	2.19	2.08	0.78	0.75
45D(E)	1.78	3.18	0.65	-1.50
45E(A)	0.67	-1.17	2.19	0.52
46	0.61	-1.45	1.52	0.30
47	2.49	1.41	0.86	1.33
61	1.26	1.28	0.86	0.53
62	-0.18	0.59	-0.68	0.26



Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.2 (cont.)

Table 1b Standardized differences: question order by education – modal frequency

Notes: Sign is positive if frequency is higher for sub-question appearing earlier
* Letter in parentheses denotes order in second version
& Denotes two (or more) values plotted in same position

Question No.	Total "T"	0-11 "1"	12 "2"	13+ "3"	MIN -7.05 … -2.0 … 0 … 2.0 … MAX 7.8
21	-0.89	0.25	-1.81	0.51	
23A(H)	-1.18	-1.14	-0.99	-0.47	
23B(I)	-1.76	-0.74	-2.54	0.22	
23C(J)	-0.40	-0.31	0.41	-0.70	
23D(K)	-1.74	-0.16	-1.35	-1.22	
23E(L)	1.21	-1.61	1.64	1.54	
23F(M)	0.09	-1.11	-0.56	2.26	
23G(N)	-0.72	-1.72	1.19	-0.74	
23H(A)	-6.60	-3.58	-4.46	-3.23	
23I(B)	0.78	0.18	0.06	1.17	
23J(C)	-7.05	-1.97	-4.63	-5.07	
23K(D)	-3.25	-2.66	-1.34	-1.96	
23L(E)	-4.44	-1.03	-3.69	-2.76	
23M(F)	-3.09	-1.61	-2.18	-1.13	
23N(G)	-3.23	-1.01	-1.82	-2.69	
24A(G)	-6.63	-2.80	-3.81	-5.02	
24B(H)	1.52	-1.06	1.67	1.19	
24C(B)	7.79	2.50	5.21	5.70	
24D(C)	0.94	1.68	1.13	-0.77	
24E(D)	2.25	1.35	0.65	1.98	
24F(E)	2.91	2.25	1.42	2.26	
24G(F)	1.87	2.10	0.94	0.86	
24H(A)	-2.05	0.53	-1.77	-2.30	
[illegible]	-5.34	-2.92	-2.57	-4.38	
[illegible]	1.65	0.31	1.43	0.54	
24K(K)	-0.75	-0.53	-0.86	0.13	
45A(B)	5.43	2.24	3.73	3.27	
45B(C)	-3.11	-0.31	-2.64	-1.87	
45C(D)	2.19	1.94	1.86	0.60	
45D(E)	1.78	-0.15	2.12	0.97	
45E(A)	0.67	-0.30	0.06	1.63	
46	0.61	0.39	-0.48	1.32	
47	2.49	1.61	2.36	0.75	
61	1.26	-0.12	1.00	0.84	
62	-0.18	-0.74	0.33	-0.60	

Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.2 (cont.)

Table 1c Standardized differences: question order by sex – modal frequency

Notes: Sign is positive if frequency is higher for sub-question appearing earlier
*** Letter in parentheses denotes order in second version**
& Denotes two (or more) values plotted in same position

Question No.*	Total "T"	Fem. "F"	Male "M"	MIN -7.05 … -2.0 … 0 … 2.0 … MAX 7.8
21	–0.89	–1.09	–0.17	
23A(H)	–1.18	–1.36	–0.34	
23B(I)	–1.76	–2.46	–0.05	
23C(J)	–0.40	0.72	–1.28	
23D(K)	–1.74	–0.49	–2.08	
23E(L)	–1.21	–0.05	1.73	
23F(M)	0.09	0.57	–0.46	
23G(N)	–0.72	–2.46	1.17	
23H(A)	–6.60	–6.23	–3.14	
23I(B)	0.78	0.08	0.94	
23J(C)	–7.05	–6.24	–3.84	
23K(D)	–3.25	–3.35	–1.35	
23L(E)	–4.44	–4.74	–1.68	
23M(F)	–3.09	–2.17	–2.20	
23N(G)	–3.23	–3.33	–1.19	
24A(G)	–6.63	–4.56	–4.75	
24B(H)	1.52	2.39	–0.23	
24C(B)	7.79	4.05	6.89	
24D(C)	0.94	3.06	–1.63	
24E(D)	2.25	0.73	2.43	
24F(E)	2.91	1.80	2.30	
24G(F)	1.87	1.99	0.74	
24H(A)	–2.05	–1.84	–1.04	
24I(I)	–5.34	–4.43	–3.15	
24J(J)	1.65	1.09	1.22	
24K(K)	–0.75	–0.18	–0.87	
45A(B)	5.43	3.95	3.69	
45B(C)	–3.11	–1.88	–2.51	
45C(D)	2.19	2.73	0.41	
45D(E)	1.78	0.36	1.97	
45E(A)	0.67	–0.38	1.32	
46	0.61	1.99	–0.91	
47	2.49	2.18	1.36	
61	1.26	2.58	–0.66	
62	–0.18	0.08	–0.29	

Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.3

Table 2a Standardized differences: response order by age group – modal frequency

Notes: Sign is positive if frequency is higher when modal category appears earlier
***Letter in parentheses denotes order in second version**
& Denotes two (or more) values plotted in same position

Question No.*	Total "T"	18-29 "1"	30-49 "2"	50+ "3"	MIN -4.21 -2.0 0 2.0 MAX 5.7
21	0.09	−0.26	1.27	−1.32	
23A(H)	0.46	−0.70	1.12	0.49	
23B(I)	−1.66	−1.53	−0.33	−1.04	
23C(J)	0.30	1.00	−0.12	−0.29	
23D(K)	1.01	1.39	0.25	0.22	
23E(L)	0.98	1.83	−0.35	0.46	
23F(M)	−1.06	−1.83	0.74	−0.98	
23G(N)	2.48	2.31	1.08	0.91	
23H(A)	3.04	2.05	1.32	2.02	
23I(B)	−2.97	−3.48	−0.37	−1.69	
23J(C)	1.78	2.39	1.16	−0.34	
23K(D)	2.86	2.68	1.44	1.09	
23L(E)	1.84	3.78	0.29	−0.64	
23M(F)	−2.71	−2.19	−0.39	−2.22	
23N(G)	1.30	2.60	−0.85	0.63	
24A(G)	2.10	−0.73	2.04	2.11	
24B(H)	1.32	1.17	0.56	0.64	
24C(B)	1.84	1.29	1.46	0.47	
24D(C)	3.91	1.85	3.10	1.50	
24E(D)	−0.66	2.32	−1.17	−2.09	
24F(E)	0.48	0.53	1.14	−1.00	
24G(F)	1.53	2.28	2.07	−2.04	
24H(A)	5.71	2.56	3.50	3.78	
24I(I)	1.36	0.90	1.43	−0.18	
24J(J)	1.00	2.13	−0.53	0.51	
24K(K)	−0.73	0.79	−0.82	−1.05	
45A(B)	−2.86	−1.91	−0.92	−2.17	
45B(C)	1.82	2.05	1.49	−0.64	
45C(D)	−2.25	−0.55	−1.92	−1.49	
45D(E)	−1.67	−1.95	0.61	−1.23	
45E(A)	2.53	1.34	1.37	1.61	
46	1.25	0.66	1.64	−0.37	
47	−0.42	−1.78	0.33	1.13	
61	−4.21	−4.20	−1.53	−1.68	
62	−3.06	−1.99	−0.94	−2.37	

Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.3 (cont.)

Table 2b Standardized differences: response order by education – modal frequency

Notes: Sign is positive if frequency is higher when modal category appears earlier
***Letter in parentheses denotes order in second version**
& Denotes two (or more) values plotted in same position

Question No.*	Total "T"	18-29 "1"	30-49 "2"	50+ "3"	MIN -4.21 … -2.0 … 0 … 2.0 … MAX 5.7
21	0.09	-0.48	0.70	-0.49	
23A(H)	0.46	-0.96	0.83	1.24	
23B(I)	-1.66	-0.80	-0.29	-1.84	
23C(J)	0.30	-0.14	1.72	-1.38	
23D(K)	1.01	0.19	1.64	-0.55	
23E(L)	0.98	1.16	1.02	-0.04	
23F(M)	-1.06	-1.30	-2.10	2.15	
23G(N)	2.48	1.86	1.41	1.20	
23H(A)	3.04	0.80	2.64	1.77	
23I(B)	-2.97	-2.66	-0.87	-2.42	
23J(C)	1.78	0.89	2.43	-0.21	
23K(D)	2.86	0.53	3.12	1.28	
23L(E)	1.84	1.22	1.21	0.87	
23M(F)	-2.71	0.18	-3.02	-1.69	
23N(G)	1.30	-0.62	1.86	0.62	
24A(G)	2.10	0.52	1.90	1.03	
24B(H)	1.32	1.40	0.82	0.59	
24C(B)	1.84	0.92	1.11	1.24	
24D(C)	3.91	1.12	3.02	2.38	
24E(D)	-0.66	-0.20	-1.08	-0.04	
24F(E)	0.48	-0.57	0.22	0.69	
24G(F)	1.53	0.22	0.99	1.38	
24H(A)	5.71	2.45	3.71	3.74	
24I(I)	1.36	1.24	1.34	0.13	
24J(J)	1.00	0.89	1.63	-0.38	
24K(K)	-0.73	-0.34	0.92	-0.08	
45A(B)	-2.86	0.13	-2.43	-2.25	
45B(C)	1.82	1.25	0.33	2.04	
45C(D)	-2.25	-0.80	-0.17	-2.61	
45D(E)	-1.67	-0.63	-0.82	-1.53	
45E(A)	2.53	1.92	1.66	0.73	
46	1.25	-1.09	1.83	0.97	
47	-0.42	1.16	0.16	-1.15	
61	-4.21	-3.20	-2.90	-1.55	
62	-3.06	-3.38	-1.56	-0.86	

Source: Sirken, Nathan and Thornberry (1989)

EXHIBIT 2.3 (cont.)

Table 2c Standardized differences: response order by sex – modal frequency

Notes: Sign is positive if frequency is higher when modal category appears earlier
***Letter in parentheses denotes order in second version**
& Denotes two (or more) values plotted in same position

Question No.*	Total "T"	Fem. "F"	Male "M"	MIN -4.21 … -2.0 … 0 … 2.0 … MAX 5.7
21	0.09	0.80	-0.63	
23A(H)	0.46	0.40	0.21	
23B(I)	-1.66	-2.54	0.11	
23C(J)	0.30	-0.73	1.16	
23D(K)	1.01	-0.56	2.04	
23E(L)	0.98	1.26	0.12	
23F(M)	-1.06	-0.96	-0.55	
23G(N)	2.48	2.60	1.02	
23H(A)	3.04	3.35	1.10	
23I(B)	-2.97	-2.74	-1.54	
23J(C)	1.78	2.55	0.15	
23K(D)	2.86	3.46	0.64	
23L(E)	1.84	1.75	1.02	
23M(F)	-2.71	-3.28	-0.66	
23N(G)	1.30	0.13	1.71	
24A(G)	2.10	0.58	2.23	
24B(H)	1.30	2.41	-0.49	
24C(B)	1.84	1.88	0.69	
24D(C)	3.91	2.48	2.94	
24E(D)	-0.64	0.29	-1.19	
24F(E)	0.48	0.81	-0.12	
24G(F)	1.53	2.24	0.02	
24H(A)	5.71	4.61	3.49	
24I(I)	1.34	1.18	0.78	
24J(J)	1.04	1.42	0.08	
24K(K)	-0.73	-0.40	-0.63	
45A(B)	-2.84	-1.84	-2.19	
45B(C)	1.82	1.18	1.42	
45C(D)	-2.25	-2.23	-1.04	
45D(E)	-1.67	-1.02	-1.27	
45E(A)	2.53	2.07	1.51	
46	1.25	0.92	0.68	
47	-0.42	0.34	-1.01	
61	-4.25	-2.90	-2.92	
62	-3.00	-2.42	-1.89	

Source: Sirken, Nathan and Thornberry (1989)

The effects of question order were assessed not only for the total population but also for sub-domains, defined by socio-demographic characteristics - age group, education (number of years attended school) and sex. The results are shown in tables 1a-1c (exhibit 2.2), together with graphic displays, for age, education and sex. They show the standardized differences for the total population, as compared to the standardized differences for the sub-domains. The letter in parentheses next to the question number denotes its order in version 2 and the sign of the difference is positive if the frequency is higher for the sub-question appearing earlier. Thus, for example, in table 1a, the differences for questions 24a and 24c are significant both for the total and for all three age groups (though less so than for the total). On the other hand, for 24d (on the effect of deep kissing), there is a significant difference only for the older age group (aged 50+) and also for females (but not for the total). Once more significantly higher frequencies of correct response ("somewhat likely") are observed, for these groups only, when the question is preceded by the question on sharing needles.

Similar tables for standard differences between response orders are given in tables 2a-2c (exhibit 2.3). As an example of the analysis of the effects of response category order (which were generally found to be smaller than the effects of question order), consider the difference found for the question on sharing drug needles (the most significant). When the modal (and correct) answer - "very likely" - was placed first a significantly higher percentage gave that answer (92.7%) than when it was placed last (90.2%). For questions 23, 24 and 45 (relating to knowledge about AIDS), significantly higher response was found for the modal category, when the positive category ("definitely true", "very likely", or "very effective") came earlier. This was true in all cases when the modal category was positive and in all but one of the cases when the modal category was negative ("definitely false", "definitely not possible" or "not at all effective") and placed later. It might be that this ordering (version A) is the more natural one and enhances correct response.

As with question order, the analysis was also carried out for sub-domains - by age, sex and education. The resulting graphical displays are given in tables 2a-2c (exhibit 2.3). For instance for question 24 it can be seen that the sign of the difference is positive when the modal category comes earlier. Thus the significant differences are found for sub-question 24d (on kissing with exchange of saliva) - for the total and for the intermediate age group only - and for question 24h (on sharing needles) - for all age groups. The differences are positive, indicating that response to the modal category "very likely" was higher when it came first. On the other hand for question 24e (on shaking hands etc.) response to the modal category "very unlikely" was significantly higher (for the older age group) when it came last (and the category "very likely" came first), while the reverse holds for the youngest age group.

The response variance analysis is based on the multi-factor extension of the basic questionnaire design error model detailed above. Unbiased estimates of the components of questionnaire design variance for each of the two factors involved - question order and response category order - were obtainable due to the fact that two levels for each factor were measured in the experiment. However, just as most of the differences, both for response order and for question order, were not significant, so in many cases the standard "variance components estimator" provided negative estimates of the factor variances. While alternative methods of estimation could ensure positive estimates of variance, they are obviously of limited interest, in this case.

In order to consider meaningful cases, the response variance model was finally applied only to the seven sub-questions for which all components of variance were estimated as positive. The results are shown in table 3 (exhibit 2.4) for the total population and for selected sub-domains. The contributions of each of the factors - response order, question order and sampling error - are given, as proportions of the estimate of the overall variance, for the standard case where only a single questionnaire version is used (assumed to be selected at random from the universe of possible versions with respect to each of the two factors). The results clearly indicate the predominant contribution of question order variance for most of the questions. The only case where the response order variance has a larger contribution than that of question order is for sub-question 24h which relates to the likelihood of getting AIDS by sharing needles for drug use and the high effect of response order for this topic has been noted. The small contribution of sampling error should be noted and is due to the large sample size.

The last column gives the estimated decrease in total variance due to the use of four questionnaire versions in the experiment (two question orders and two response orders) rather than a single version. The gains are obviously larger (up to 89%) for cases where the contribution of the response order variance is predominant, since this component is eliminated when both of the possible response orders are included in the survey. However even when the question order component of variance is predominant, gains of at least 50% are obtained in almost all cases, since the contribution of this component is reduced by a factor of 2. The possibility of attaining further reductions in variance by further increasing the number of questionnaire versions used is considered in Nathan and Sirken (1987).

The AIDS split ballot experiment was embedded in the National Health Interview Survey primarily to field test the response variance paradigm, although the analysis by the descriptive paradigm was also in mind. What are the implications of the AIDS split ballot experiment for the survey? Does it suggest a viable and useful survey design and estimation strategy to measure and control the combined effects of sampling errors and response errors due to questionnaire design on the reliability of the National Health Interview Survey statistics? The survey routinely administers a single questionnaire version, and applies classical sampling theory to draw inferences about the civilian non institutional population from the responses of sample households. This approach disregards the questionnaire design effects on the reliability of the survey statistics. In assessing the

EXHIBIT 2.4

Table 3: Components of variances for modal category: selected questions and sub-groups (Percentages)

Total

Question No.	Modal category response	Coefficient of variation	Contributions of components: Response order	Question order	Sampling error	Relative decrease in variance
23H	51.3	8.35	8.80	90.20	1.10	53.80
23J	78.4	4.63	2.20	96.80	1.00	50.60
23K	44.9	5.34	26.30	70.10	3.60	61.30
23L	76.7	3.05	5.90	91.70	2.40	51.70
24C	31.7	7.69	1.20	97.50	1.30	50.00
24H	92.0	1.42	81.10	16.40	2.60	89.30
45A	53.9	1.93	0.70	96.60	2.70	49.00

Age 18-29

Question No.	Modal category response	Coefficient of variation	Contributions of components: Response order	Question order	Sampling error	Relative decrease in variance
23H	53.5	10.58	8.80	88.50	2.70	53.10
23J	82.0	4.83	16.40	80.20	3.40	56.50
23K	49.3	9.91	24.40	71.70	3.90	60.20
23L	79.9	4.01	73.30	21.30	5.50	83.90
24C	32.8	17.14	1.90	95.30	2.80	49.60
24H	93.6	1.54	49.90	41.10	9.00	70.50
45A	57.9	4.96	18.10	75.10	6.80	55.70

Females

Question No.	Modal category response	Coefficient of variation	Contributions of components: Response order	Question order	Sampling error	Relative decrease in variance
23H	51.8	10.44	11.80	87.10	1.10	55.30
23J	79.3	5.35	6.70	92.10	1.20	52.70
23K	43.0	8.02	33.80	63.10	3.10	65.40
23L	78.7	4.03	4.50	93.30	2.20	51.20
24C	30.5	11.14	7.40	89.70	2.90	52.30
24H	91.7	1.50	77.80	18.40	3.80	87.00
45A	54.9	4.87	7.30	89.60	3.10	52.10

Ed 12 yrs

Question No.	Modal category response	Coefficient of variation	Contributions of components: Response order	Question order	Sampling error	Relative decrease in variance
23H	50.3	9.79	13.40	84.40	2.20	55.60
23J	80.5	4.94	10.50	87.40	2.10	54.20
23K	43.6	5.84	77.10	14.10	8.80	84.20
23L	77.9	3.93	1.80	94.50	3.70	49.00
24C	29.7	17.67	0.40	97.70	1.90	49.30
24H	92.7	1.49	70.80	23.70	5.50	82.60
45A	54.6	5.91	15.50	81.40	3.10	56.20

Source: Sirken, Nathan and Thornberry (1989)

AIDS split ballot experiment, it is relevant to compare the results that would have been inferred by classical sampling theory with those that were inferred by the response variance paradigm. For example, increasing the number of questionnaires administered in the National Health Interview Survey from one to four versions, could reduce the coefficient of variation by 50 percent.

However, the results should be interpreted very cautiously. For example, they do not yet reflect the effects of the the survey's complex sampling design. Also, they are imprecise because the AIDS experiment administered only four questionnaire versions and they were not selected strictly at random. Despite these limitations, the findings imply four important potential benefits of embedding split panel experiments in a national survey: (1) to provide estimates of the response error effects of questionnaire designs that otherwise would not be measurable; (2) to test cognitive theories useful in improving survey design; (3) to reduce and control the response variance due to questionnaire designs by varying the number of administered questionnaire versions; and (4) to improve the total survey design by optimally allocating resources between sample size and the number of questionnaire versions. Embedding split ballot experiments in sample surveys creates conceptual and operational problems. Though often exaggerated, these problems are serious and add complexity to an already complex survey measurement process. Nevertheless, the AIDS split ballot experiment demonstrates that these problems are surmountable and that the response variance paradigm is a useful tool for drawing inferences about the response error effects of questionnaire designs on the reliability of sample survey statistics.

3. QUESTIONNAIRE DESIGN EFFECTS AND RESEARCH IN COGNITIVE ASPECTS OF SURVEY METHODOLOGY.

3.1 Survey methodology and cognitive research.

Practically all survey research and research in questionnaire design effects in particular has concentrated on making changes or improvements in questionnaire or in the survey method, measuring the effects of these changes on survey error and thereby reaching decisions on which are the optimal, or, at least, better procedures. However in order to understand the effects better, it seems that we need to find out more about the cognitive aspects involved in the survey process. We can view the process of eliciting answers from respondents in a survey schematically as follows:

Most survey methods research until recently has concentrated on the first and last stages of this process - manipulating the stimulus (methods of asking questions, mode of collection etc.) and then measuring and evaluating the response, in order to reach decisions on optimal or better survey methods. Indeed, as already mentioned, a surprising amount of progress has been made. However, recently it has been recognized that in order to advance further in this area, we must investigate the black box of the cognitive processes involved in getting answers from respondents.

For this purpose, the first questionnaire design laboratory was set up in 1985 at the U.S. National Center for Health Statistics and, since then, other laboratories have been set up in a number of institutions, including the Bureau of Labor Statistics and the Bureau of Census in the U.S. and, more recently, at Statistics Sweden. The work is carried out jointly by survey researchers and cognitive psychologists. For example, at the U.S. National Center for Health Statistics, where much of the pioneering work in Cognitive Aspects of Survey Methodology has been going on, it is carried out through the National Laboratory for Collaborative Research in Cognition and Survey Measurement, whose mission is "to improve the quality of health statistics by applying the methods of cognitive science to the design of data collection instruments". The laboratory is funded primarily by National Science Foundation grants which total over one and a half million dollars over the past five years.

The main methods used in these laboratories are:

1. Concurrent "think-aloud" protocols.

2. Focus interviews (unstructured group discussions).

3. Paraphrasing by respondents.

4 Measurement of response latency (time till response).

They all involve the use of trained interviewers, mostly cognitive psychologists, who try to understand the cognitive processes involved in response. Cognitive psychologists recognize the following cognitive stages in responding to survey questions:

1. Comprehension - interpretation of the question by the respondent

2. Retrieval - of relevant information from memory.

3. Estimation/ judgments - evaluation or estimation of response.

4. Response.

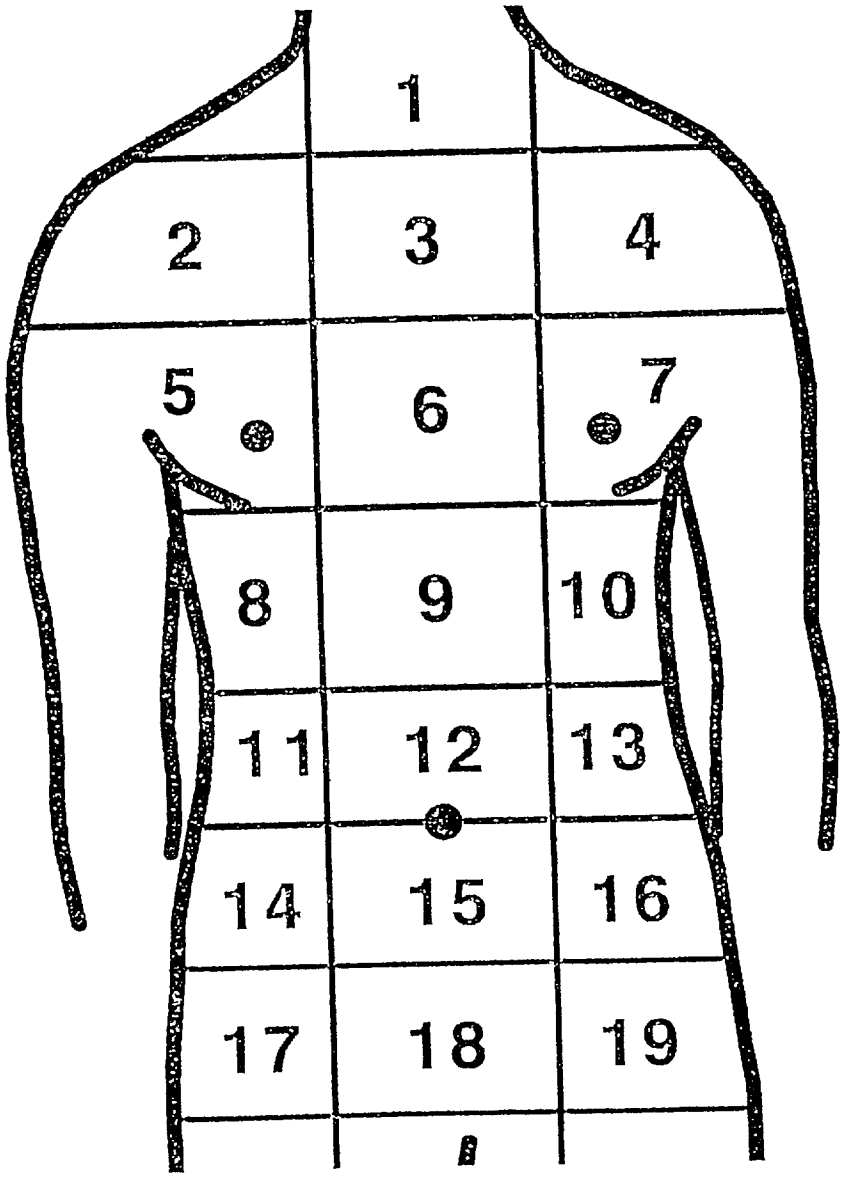
An example of the use of cognitive research in improving questionnaire design is in a comprehension problem discovered during the development of a questionnaire for the U.S. National Center for Health Statistics Health Examination Survey (HANES III). It is a good illustration of the difficulty caused by vague and unfamiliar terms. The original question was phrased: "During the past year have you been bothered by pain in your abdomen". Respondents were asked this question in the laboratory without indicating, on the basis of a probe, any special difficulty. None admitted that they did not understand the word "abdomen". They then were asked to shade in a diagrammatic representation of the human body (see exhibit 3.1) the areas containing the "abdomen". The findings were most striking. No two of the 12 subjects who participated in this experiment shaded the same area. This was a serious matter indeed, since the question on abdominal pain served as a screening question for a series of subsequent questions about chronic digestive diseases. On the basis of the laboratory findings, the question wording was revised. The term "abdomen" was dropped in favour of a graphic illustration, such as in exhibit 3.1, on which the areas referred to were designated.

3.2 Cognitive aspects of surveys with sensitive questions.

Most of the research in cognitive aspects of survey methodology to date has concentrated on the first three stages above - see, for instance, Lessler and Sirken (1985), Sirken, Mingay et al. (1988), Tourangeau (1984) and Willis, Royston and Bercini (1989). In the following we shall concentrate on the last stage of the cognitive process, where the respondent makes his final decision on whether and how to respond. Although this stage is a factor in all survey work, it is of particular importance with respect to sensitive issues and here

EXHIBIT 3.1

Diagrammatic representation for question on abdominal pain



we shall treat especially the problems of answering sensitive questions.

The issue of sensitive survey topics has become particularly acute recently due to the necessity to obtain detailed information relating to the AIDS epidemic and to the drug problem. The data needed obviously require detailed surveys relating to very intimate and sensitive behaviour, such as sexual practices, drug use etc. To be successful, the large scale population surveys currently being undertaken or planned to collect these statistics will have to overcome the well-known reluctance of survey respondents to respond honestly to sensitive questions.

Several survey techniques for enhancing the truthfulness of responses to sensitive survey questions have been proposed, primarily by protecting the anonymity of respondents. Perhaps, the best known of these techniques is that of randomized response, but alternative methods for assuring anonymity have been proposed. An example of how the randomized response method is used is in asking a question on whether the respondent ever used drugs intravenously (see example in exhibit 3.2). The interviewer asks the respondent to toss a coin, without the interviewer seeing the result, and to answer the sensitive question only if the coin fell on "heads". If it fell on "tails", he is to answer an innocuous question on his mother's month of birth. This is supposed to ensure anonymity in that the interviewer does not know if the respondent answered the sensitive question (on drugs) or the innocuous question on his mother's month of birth.

Since the idea of randomized response was first introduced by Warner (1965), a variety of procedures have been proposed and a great amount of empirical studies have been carried out. These have been summarized in several books and in over 250 articles - see the bibliography of Nathan (1988). The notion that "anonymity breeds honesty of response" is not, however, always supported by the empirical evidence. The quality of response in randomized response surveys has often been suspect and has varied unpredictably and inexplicably from one randomized response survey to another - see for example Brewer (1981), Goodstadt and Gruson (1975), Greenberg et al. (1977) and Soeken and Macready (1982). Compared to conventional surveys that ask sensitive questions directly (i.e. without response anonymity), randomized response estimates invariably have larger sampling errors and sometimes have as large or even larger response error. Other methods of ensuring anonymity suffer from similar problems and several empirical studies have been undertaken to compare different methods - Brown and Harding (1973); Krotki and Fox (1974); Locander, Sudman, and Bradburn (1976); Marquis, Marquis and Polich (1986); Rolnick et al. (1989).

Ongoing work at the National Center of Health Statistics attempts to consider this problem in the context of cognitive aspects of survey methodology. The research is aimed at improving our understanding of the response error effects of data collection instruments and strategies, designed to address the problems of response to sensitive questions. In order to do this a theory of the factors affecting the cognitive strategies adopted by respondents in deciding whether to

EXHIBIT 3.2

Examples of randomized response questions

[heads]

In the past 12 months, did you use a needle to inject or "shoot up" street drugs (for example, heroin, cocaine, amphetamines)?

[ANSWER: "YES" or "NO".]

[tails]

Was your mother born in January, February, March, April, May, or June?

[ANSWER: "YES" or "NO".]

[heads]

Have you ever seen a psychologist or a psychiatrist for professional help?

[ANSWER: "YES" or "NO".]

tails]

Was your mother born in January?

[ANSWER: "YES" or "NO".]

answer sensitive survey questions and whether to answer them truthfully has been proposed by Nathan and Sirken (1988). The cognitive model of survey responses to sensitive questions is based on classical utility theory.

The model relates respondents' decisions on whether to respond and to respond truthfully to their perceptions of the **risks** of response disclosure and the **losses** that would follow such disclosure, as compared to the benefits of responding truthfully. The model proposed follows classical utility theory - such as proposed by von Neumann and Morgenstern (1944) and by Pratt, Raiffa and Schlaifer (1964) - in considering decision-making as a function which defines for each possible state of nature an outcome which will result if a given course of action is taken. The decision is made on the basis of the decision-maker's personal evaluations of risk and of loss (or of gain). However, we do not necessarily assume, as is done in classical utility theory, that the decision taken is that which minimizes expected net loss and are willing to accept a more complex relationship between probability assessments, losses and decisions. Nevertheless, we still consider that the respondent's decisions are a function of his assessments of risk and of the net loss involved.

The respondent's perceived **risk** relates to the degree of belief that the respondent assigns to the event that he is identified as having a certain characteristic, via the survey process. The perceived risk of disclosure will, in general, differ according to the agent to whom information is revealed - the interviewer, the collecting agency or government agency - and on the characteristics of the respondent. The following components of risk will be considered:

1. Risk of disclosure to the interviewer (the **embarrassment** factor).

2. Risk of breach of **anonymity** (disclosure to the collecting agency only).

3. Risk of breach of **confidentiality** (disclosure to family, friends, employer or other government agency).

The risk should theoretically, be independent of the subject-matter, the sensitivity of the question, or the phrasing of the question itself, all of which relate to the respondent's perceived loss due to having information about him divulged.

In parallel to the components of risk, we can consider these components of **loss**:

1. **Benefit** of data to society (negative loss).

2. Loss due to **embarrassment** towards the interviewer.

3. Loss due to breach of **anonymity**.

4. Loss due to breach of **confidentiality** (data transfer to family, employer or government).

The additional component is that of benefit. The benefits (or negative loss) of responding honestly could relate to the respondent's view on the general benefits of the data to society (his "altruism"), to the benefit of positive social contact with the interviewer or to his fear of discovery if he does not answer truthfully. The remaining three components measure the loss (or negative utility) that the respondent assigns to the event that he is identified as having a certain characteristic, via the survey process. They parallel the components of risk and relate to the various potential recipients of the individual identifiable information.

Similarly to perceived risk, the loss would depend on the true status of the respondent, with respect to the sensitive group, and, possibly, on other characteristics of the respondent. It would, generally, differ according to the agent or agency to whom the information is given - the interviewer, collecting agency or other government agency. Thus, with respect to the interviewer, it might also depend on her/his characteristics (e.g. his/her social status relative to that of the respondent) or on the survey process. The loss components would, in general, depend on the subject-matter, on the sensitivity of the question, and, possibly, on its phrasing. However they should be independent of the method of collection and of ways used to ensure anonymity or confidentiality.

Finally, the respondent's decision function defines the way in which the two elements defined above - the risk and the loss - combine, together with other factors, to determine the respondent's decision with respect to two aspects of the survey process - whether to respond at all and, if he responds, whether to respond truthfully or not (we assume that the respondent knows the truth, or else we consider his perceived truth as true). The decision may depend only on the expected net loss (i.e. expected loss less gain), or on some other function of loss and of gain. It may be a deterministic one (e.g. not to participate if the expected loss exceeds the benefit), or it may be a probabilistic decision (e.g. the probability of answering truthfully is a decreasing function of the expected net loss).

3.3 Experiments for sensitive question research.

In order to get some idea on whether this model framework is at all relevant, the U.S. National Center for Health Statistics has embarked on a series of experiments. These experiments are primarily exploratory in nature and are aimed at finding out whether the elements of the proposed model are at all measurable - that is if respondents can assess the components of risk and of loss - and whether their perceptions relate in any way to their decisions on whether and how to participate. They are to be regarded as a first developmental stage out of three stages. The other two stages are full-scale laboratory testing and field testing.

The experiments are mostly laboratory-based in the sense that volunteers are recruited for interviewing in a laboratory setting. The interviewing is done by trained cognitive psychologists, mostly in the Questionnaire Design Research Laboratory of the U.S. National Center

for Health Statistics. The interviews include not only answering closed questions on assessments of the model components, but also open-ended questions and think-alouds, which are fully recorded and analyzed by the cognitive psychologists. A set of four initial experiments has just been completed. They are:

SQR1 - Evaluation of perceptions of **risk**.

SQR2 - Evaluation of perceptions of **loss**.

SQR3 - Assessment of respondents' **decisions**.

SQR4 - **Synthesis** of decision process and of its components.

A brief description of the design of each of these experiments together with its main results and conclusions follows. In the first experiment - the assessment of risk - the factors which were varied (each at two levels) were administration procedure, protection level, sensitivity of the question and order of administration. In addition, socio-demographic factors (sex, age and educational level) were controlled - see details in exhibit 3.3.

The two administration procedures used here were a randomized response procedure and an envelope procedure (where the respondent was asked to answer without the interviewer seeing his answer and then sealing it in an envelope). Each respondent was asked for assessments of risk for both procedures - with order of administration randomly assigned. Each procedure was administered at two levels of protection. For randomized response these were attained by using different non-sensitive questions (see exhibit 3.2).

Two different sensitive questions were asked - one about serious drug use and one about seeking psychological help. Each question was randomly assigned to one half of the subjects but they were also asked if their assessments would change if asked the other question. Finally an attempt was made to balance on socio-demographic characteristics of the subjects and with respect to the two interviewers. With only 24 subjects in total this was quite a feat and not completely attained. To illustrate how the interviews were conducted, some parts of the protocol are shown in exhibits 3.4-3.8.

After some opening remarks in which the basic set-up was explained, the interviewer proceeded in a different way depending on the procedure which was allocated as administered first. This was followed by detailed explanations of the randomized response technique and verification that they were understood (see exhibit 3.5). Examples of the randomized response questions are shown in exhibit 3.2. The upper panel is the drug question and the lower shows the question on psychological help. The upper panel provides high level protection, since by including six months there is a 50% chance that the correct answer is yes if the coin falls on tails rather than a 1 in 12 chance for the lower panel. This implies that a "yes" answer is less jeopardizing as revealing the true status of belonging to a sensitive group for the upper panel than for the lower panel.

EXHIBIT 3.3

SQR1 - RISK ASSESSMENT
Experimental design

FACTOR (levels)	Betw/Within subjects
ADMINISTRATION PROCEDURE (RR/envelope)	Within
PROTECTION LEVEL (high/low)	Between
SENSITIVITY OF QUESTION (high/low)	Between
ORDER OF ADMINISTRATION (RR/envelope first)	Between
SOCIO-DEMOGRAPHIC (sex, age, education)	Between

EXHIBIT 3.4

SQR1 PROTOCOL - OPENING REMARKS

Each year, the Federal government asks people a variety of health-related questions when it conducts its door-to-door surveys. Some of the questions are rather personal; so the government is currently exploring ways of asking these questions that protect the individual's privacy. There are a number of different procedures that could be used. I would like to try some of those procedures with you today. .....

****

I would like you to imagine the following: Suppose a woman in her early 40s came to your door and asked you to participate in an important survey for the U.S. Public Health Service. .....

EXHIBIT 3.5

[FOR RANDOMIZED RESPONSE ADMINISTRATION CONDITION]

Now, let's imagine that I am the interviewer and I go on to say:

> The last question that I want to ask is different. Survey organizations have found an interesting way to ask questions that some people feel are too private to talk about. It allows the person to answer honestly, yet secretly.
>
> The procedure ensures that the information provided is anonymous; in other words, it will not be possible to match a person's answer with the question that he or she is answering.

EXHIBIT 3.6

16

[FOR ENVELOPE CONDITION - HIGH PROTECTION]

Do NOT write your name on the questionnaire. Read the question and circle the appropriate response--either "yes" or "no". Do not let me see your answer. When you are finished, so that no one will see your answer, please tape the questionnaire closed with this tamperproof tape and seal it in this special envelope.

[SHOW THE SUBJECT THE TAPE AND THE TAPED-UP QUESTIONNAIRE]

After you have done this, we will walk to the nearest mailbox and you can mail the envelope yourself.

EXHIBIT 3.7

17

QUESTIONING PHASE

1. Can you explain in your own words HOW this procedure protects the privacy of the individual who is being interviewed?

OK, I would now like you to answer the following questions by marking the scales that are provided.

****

2. If the procedure I just described was used in an actual survey, what are the chances that someone could figure out which of the two questions you had answered?

EXHIBIT 3.8

ANSWER SHEETS

2. If the procedure I just described was used in an actual survey, what are the chances that someone could figure out which of the two questions you had answered?

0% |--|--|--|--|--|--|--|--|--|--| 100%

THEY DEFINITELY COULD NOT FIGURE OUT THE QUESTION

THEY DEFINITELY COULD FIGURE OUT THE QUESTION

****

4a. What are the chances that someone WOULD figure out which question you had answered by using a trick or some kind of deception?

The envelope procedure - under the high protection level - was explained as shown in exhibit 3.6. Under the low level protection procedure the respondent was just asked to hand the sealed envelope to the interviewer. After carefully explaining the procedure and making sure that the subject understood it, the interviewer proceeded to the questioning phase in which he tried to obtain the subject's assessments of risks -- see example in exhibit 3.7. The first question tried to ascertain how the subjects understood the protection of anonymity provided by the procedure used. The remaining questions try to assess various elements of risk. The subjects were asked to quantify their assessments on scales such as those shown in exhibit 3.8.

After going through some 5-6 questions for one procedure, the subjects were then asked to repeat the whole process for the second administration procedure. The experiment was extremely complex from the point of view of administration (eight different sequences of questionnaires had to be used), but the subjects did not have any special difficulty in answering and following the procedures.

Remembering that there were only 24 subjects and that the aim was primarily exploratory rather than analytic the conclusions are necessarily rather limited. These were basically of two types - those relating to evaluation of the measurement process and those relating to the cognitive tendencies and the results of the measurements themselves.

With respect to the understanding of the protocol, basically it was found to be understood. However, as expected, some formulations presented to subjects were not completely understood as intended; many of the questions asked set a number of implicit assumptions which are not always stated in order to keep the question of a manageable length. As to conclusions related to cognitive tendencies, the main conclusion was that subjects do consider spontaneously both the costs and benefits associated with answering, as utility theory would propose. For example, they want to know why the information is being collected (will it will have some benefit to someone?), and they want to be sure that it will be collected in such a way that it is useful. Secondly, there is a tendency for respondents to focus on elements of the administration mode other than those procedural aspects which are used to provide protection. Thus, in asking sensitive questions, the appearance and demeanor of the interviewer may be as important as the procedure used to protect the individual.

Finally, some conclusions related to the specific purpose of the experiment, i.e. to determine if perceptions of risk, that are associated with matching of the name, response, and question asked, can be measured accurately. It is clear from the previous discussion, though, that the assessments that were reported tended to include elements of the surveying situation that were outside of the range intended for purposes of measurement.

The second experiment set out primarily to evaluate losses due to to disclosure of sensitive information and the potential benefits of the data collected. Note that no attempt was made to assess possible

losses due to wrong information being disclosed (e.g., when a non-user is considered a drug user, because he answers untruthfully). These types of losses are regarded as very minor relative to those involved in disclosing true information that someone does belong to the sensitive group. In this experiment some risk questions were also asked, in order to get evaluations of risks with respect to exactly the same outcomes for which we measured costs.

The design of this experiment, SQR2, was similar to that of the first experiment, except that here the sensitivity of the question, rather than the administration procedure, was designated as the within-subject factor. This was due to the feeling that question sensitivity, rather than administration procedure, was the main factor affecting losses, whereas administration procedure affected primarily the risks. In addition, the protection level factor was dropped in this experiment, since it was not considered as likely to affect losses. Other aspects of design were similar to those of the first experiment - see exhibit 3.9. Again 24 subjects were included and attempts were made to balance on socio-demographic characteristics. After a similar opening, questions on benefits, on losses and risks were asked, as shown in the examples of exhibit 3.10.

The first question assesses the benefits perceived as being derived from the data collected. Question 3 assesses the loss assigned to embarrassment towards the interviewer should he consider the respondent to belong to the sensitive group - on the basis of his answer. Question 4 relates to the losses resulting from breach of anonymity (i.e the collecting agency figuring out the answer).. The next questions asked with respect to breach of confidentiality, i.e release of data outside the collecting agency. These questions (5, 6 and 7) were asked with respect to release of information to three different recipients - family and friends, employer and other government agency. Q.5 assesses the risk of breach of confidentiality, Q.6 probes for possible outcomes and Q.7 assesses the loss.

The results of this experiment with respect to measurement problems once more pointed to the difficulties of obtaining assessments for hypothetical situations, primarily drug use, in which most of the subjects were never involved. However, by and large the measurement process was found to be viable - subjects were able to give coherent answers and there was a wide range of response to practically all questions.

With respect to the results themselves it must be remembered that they relate to only 24 subjects. However some tendencies seem to be observable. While in general differences due to the question sensitivity were small, they were consistently large with respect to the loss assigned to the embarrassment factor and to that assigned to breach of confidentiality. In both cases the loss for the drug question was judged as higher than that for the question on psychological help. The order of administration did not seem to have any effect on any of the responses, but the administration condition itself did seem to affect the assessment of loss due to invasion of privacy, when anonymity is breached. This loss was assessed as considerably higher for the envelope procedure than for randomized

EXHIBIT 3.9

SQR2 - LOSS ASSESSMENT
Experimental design

FACTOR (levels)	Between/ Within subjects
ADMINISTRATION PROCEDURE (RR/envelope)	Between
SENSITIVITY OF QUESTION (high/low)	Within
ORDER OF ADMINISTRATION (drug/psyc. first)	Between
SOCIO-DEMOGRAPHIC (sex, age, education)	Between

EXHIBIT 3.10

1. Imagine that this procedure was used to collect information from several thousand people. How helpful do you think the overall statistical information would be in improving the nation's health health?

****

3. How much embarrassment or unease would you feel if the interviewer thought that you were a drug user, after you had answered?

****

4. Some people feel that the government has no business knowing their answers to some questions. If there was some way that the Center could figure out which question you had answered, how upset would you be about this?

5a) If the Center were able to tell from the survey that someone is a drug user, how likely do you think it is that they would release this information to the person's family and friends?

****

6a) What kinds of things do you think could happen to the person if the Center released this information to the person's family and friends?

****

7a) If the Center released the information to the person's family and friends, how upset would the person be about this?

response (differences of 40% for drug use and of 23% for psychological help). There were important differences, in general, with respect to the sensitivity level, with higher losses for the drug question than for the question on psychological help, with respect to embarrassment and to breach of confidentiality, as well as higher risks for the risks and losses due to breach of confidentiality. A summary table of the mean risks and losses is given in exhibit 3.11.

The factor with the greatest influence on perceptions relating to breach of confidentiality seemed to be the identity of the recipient, both for risks and for losses. With respect to risks the difference is primarily between government, on one hand and employer and family, on the other. The risk assigned to breach of confidentiality to government agencies was assessed as about double that for employer or family. With respect to losses, those due to breach of confidentiality to employers and government were not found to be very different from each other but both were much higher than those relating to release of information to family members (by some 10 percentage points).

The third experiment (SQR3) was designed to assess decisions on participation in the survey and on answering truthfully. It was considered that for this element it would be most difficult for subjects not belonging to sensitive groups to project behavior of subjects who do belong to these groups. Thus for this experiment (as well as for the final one - SQR4), subjects were recruited from a high-risk group, in addition to those from the general public. The high-risk subjects were 25 patients at a local drug abuse treatment clinic, who volunteered for the study. An additional 15 subjects were recruited from the general public. The design of this study differed somewhat from those of the previous experiments, in that the emphasis was on the cognitive processes involved in the decision on response. Thus the only experimental factor, in addition to the type of subject (drug-user/general) and his/her demographic characteristics, was the sensitivity of the question - although in this experiment all the questions were very sensitive. Subjects were given the usual introduction and asked to project themselves in a survey situation, with the assurance of confidentiality and of high-protection anonymity (the envelope sealed with tamper-proof tape and mailed by the respondent - randomized response was not used here). The respondent was then asked to answer questions on participation in a survey relating to health problems.

The questions asked before the respondent actually saw the sensitive questions are shown in exhibit 3.12. The respondent was then given some 10-15 minutes to read the complete list of 25 questions on health, sex practices and drug use (many of which were very sensitive). Then he was asked to refer to each of several questions (or sets of questions) separately and answer whether a respondent who actually was in the relevant sensitive group would answer the question truthfully.

The results were quite surprising - 92% of the subjects said they would participate in the survey, even **before** they saw the survey questionnaire. This rose to 100% **after** they actually saw the questionnaire. The main reasons cited for agreement to participate

EXHIBIT 3.11

RISK/LOSS (Q.5 and Q.7) BY RECIPIENT AND TOPIC

	VALUE					
	LOSS			RISK		
		TOPIC			TOPIC	
RECIPIENT	TOTAL	DRUG	PSYC	TOTAL	DRUG	PSYC
TOTAL	81.8	83.8	79.8	20.7	22.8	18.5
FAMILY	74.0	72.9	75.2	15.3	16.4	14.1
EMPLOYER	85.1	87.9	82.3	14.1	15.7	12.6
GOVERNMENT	86.1	90.4	81.8	32.4	36.0	28.7

EXHIBIT 3.12

SQR3 - QUESTION PHASE

Q1: Given what you now know, how would you feel about participating in a survey of this type?

IF SUBJECT SAYS HE/SHE WOULD PARTICIPATE ASK Q1a; IF HE/SHE WOULD NOT PARTICIPATE SKIP TO Q1b.

Q1a: Why would you choose to participate? What would your reason(s) be?

Q1b: Why would you choose NOT to participate?

READ: Now, before we go any further, I would like you to read over the entire questionnaire.

Now let's take a closer look at the questionnaire. I am going to ask you to imagine how different kinds of people might react to some of the questions that appear in this questionnaire.

Q4a. Please take another look at questions 8-10. (PAUSE) If a survey respondent actually had been told by a doctor that he/she (USE GENDER OF SUBJECT) had GENITAL HERPES, how do you think he/she would respond to question 10?

Do you think he/she would REFUSE TO ANSWER the question?

Do you think he/she would ANSWER TRUTHFULLY and mark "YES"?

Or, do you think he/she would ANSWER UNTRUTHFULLY and mark "NO"?

were the strong assurances of anonymity and confidentiality, although most subjects did refer in some way or another to the sensitive nature of the questions. On the individual topics a wide range of response was found, as can be seen from the table in exhibit 3.13.

The table reports only the percentage who said that the respondent would answer truthfully. A very small proportion said they would refuse to answer individual questions, the remainder answering that they would lie or that it would depend on the specifics of the respondent or the situation. There is obviously a high degree of differentiation between topics with less frequent honest answers to the more sensitive topics (down to 20% for anal sex by general subjects). There are often considerable differences between the two groups and for the more sensitive types of behavior the drug-users are usually more ready to answer truthfully than the general subjects. It must be borne in mind that the drug-users participating in the study were receiving treatment, implying that they had faced up to their problems, at least to some degree.

The analysis of the verbal explanations given for the respondent's assumed behavior (answering truthfully or not) revealed that the main reasons for not answering truthfully were denial (blocking from conscious awareness), embarrassment and fear of isolation or rejection. There were also explanations relating to risk factors - doubts about the anonymity or confidentiality of the survey.

Overall the experiment did show that subjects could relate to the decision-making process and allow assessment of this element of the utility theory model. The relatively high number of cases reporting that they could not decide how the respondent would answer ("it depends") - up to 43% - indicates that a much more clearly defined subject must be used for projection. This was indeed the main change made in the plans for the last experiment (SQR4) which is now being analyzed. It was decided to use vignettes to describe the person to whom we want the subject to relate (for instance: "Imagine Tom, a 37 year old stock-broker who lives in Washington. He uses cocaine about once a week and his drug use is a secret to his wife, boss and co-workers"). The questions on risk, loss and decision are then asked, in a similar way to those used in the previous experiments, with respect to this person ("How would you answer if you were Tom?"). Other aspects of this experiment incorporate the main features of the previous experiments, with administration procedure and sensitivity level as within-subject factors and the vignette and subject characteristics as between-subject factors. Only envelope procedures were used (at two levels of protection) and both drug-users and general subjects were be interviewed.

To summarize - it has been established in these experiments that the elements of the utility theory based model - risks, losses and decisions - can be viably measured by some modified versions of the methods tested. Furthermore, there are indications that components of risk and loss do seem to have some relationship to the decision that the respondent makes on whether and how to respond. We expect to find out more about the way in which assessments of risk and loss combine

EXHIBIT 3.13

SQR3 SUMMARY OF RESULTS
Pct. would answer truthfully

QUESTION	DRUG USERS	GENERAL
Herpes	48	67
HIV testing	36	47
Shooting up	76	40
Sex with drug user	60	47
Sex with bisexual	56	53
Get/give $/drugs for sex	56	53
20+ hetero-partners	64	73
20+ homo-partners	36	40
Anal sex	52	20

in order to determine this decision from the last experiment. However it must be kept in mind that these experiments are extremely limited in size and their generality is very doubtful. If the results of the present experiment are positive, it is hoped that more ambitious laboratory experiments will be undertaken, followed by large-scale field experiments, to furnish the basics of a decision model. In these methods of validation will be introduced, which do not depend on projection techniques or on subjects whose sensitive situation is known to the interviewer (e.g. find drug users who do not know that the interviewer knows that they are users). If decisions can be shown to relate in a well-defined way to risks and losses which are measurable and if indeed, as surmised, risks depend primarily on the method of administration and losses depend primarily on the sensitivity of the question, then the decision model could be very useful in improving survey design. In any case, the approach developed seems promising in providing insights into the way in which respondents decide how to answer sensitive questions. We are definitely beginning to get some perception on this, but obviously what has been done so far raises more questions than answers.

4. METHODS OF RESPONSE ERROR EVALUATION AND THEIR APPLICATION TO IMPROVEMENT OF QUESTIONNAIRE DESIGN.

As already pointed out, any improvements in questionnaire design, as well as improvements in other features of survey design, have to be based on solid quantitative data. Thus the theory of questionnaire design effects, outlined in section 2, must finally be tried out on experimental data. Similarly, the research in cognitive aspects of survey methodology, treated in section 3, must be quantified in order to be useful for the improvement of survey design. In the following, we consider additional methods used in evaluating survey results, which can be useful in improving questionnaire design and survey design, in general.

We differentiate between **macro-methods** and **micro-methods**. Macro-methods use **aggregated** data to compare between survey methods, questionnaire designs or other alternative design features. The aggregation can be at different levels, but these methods do not require information at the individual or unit level. Micro-methods, on the other hand rely on the **disaggregated** data, i.e. comparisons at the individual or unit level. It is clear that while macro-methods will, in general, be less expensive and simpler to operate, they will, on the other hand, provide less evaluation information than micro-methods and the information will be less accurate. It should also be pointed out that, in general, micro-methods will only be available to the agency collecting the data, so that external research organizations will mostly have to use macro-methods.

4.1 Macro-methods of response error evaluation.

The main macro-methods of evaluation are those traditionally used when individual data are not available - comparisons of the survey data with data from alternative sources, which approximately provide similar information. The other sources could be administrative data or data from other censuses or surveys. In general, the alternative sources will not provide exactly the same data which the survey is supposed to provide and there will be differences in definitions, in frameworks and in reference times or periods. Thus raw comparisons will usually not suffice and more sophisticated methods of comparison, which take these problems into account, have to be used. The widely used techniques of demographic adjustments and the use of demographic models are a case in point.

In some cases aggregates from different parts of the survey can be compared and provide a basis for evaluation. This can be done by design, where the survey is planned in a way which will provide internal comparisons for evaluation. A good example is the split-ballot experimental design, as discussed in detail in section 2. Other ways to check internal consistency on the basis of aggregate data are also available.

The frequent use of panel surveys, in which at least part of the sample is investigated more than just once, opens the way to a good potential source for evaluation. In general, simple comparisons of aggregates for the different panels is insufficient. However, if the aggregate information on differences in classification over time for the same units is available, this can be utilized efficiently for evaluation. Examples of this can be found in Kantorowitz and Nathan (1987) and in Nathan (1987), where the Israel Labour Force Survey (a rotating four panel design) was used for inter-panel comparisons. Here we shall consider, in some detail, a further development of this on the basis of a **misclassification model**, in an application to the problem of deciding on the mode of collection (telephone or home interview).

As we have already seen, the assessment of measurement errors is often carried out on the basis of a controlled experiment of the split-ballot type, in which subjects are randomly allocated to alternative design options. Response error effects are then essentially estimated on the basis of the comparison of results obtained from the sub-sample allocated to the different alternatives, possibly with added validation information. In many cases, random allocation of subjects to modes of collection is not possible and the allocation is based on practical considerations of field organization. To deal with this case some form of repeated observation is necessary. This is possible, for instance, in panel surveys for characteristics which do not change over time. In the following we consider the assessment of measurement errors due to the mode of data collection (home visit or telephone interview) as an example of the use of misclassification models. The model proposed for this case assumes probabilities of correct classification and of misclassification depending on the mode of collection. The estimation and comparison of these parameters are used to assess the effects of mode of collection.

4.2 Evaluation of mode of collection response effects.

The study of mode of collection effects has become one of the more important elements of survey design and should be considered analogous to questionnaire design. The rising costs of survey field-work, especially when carried out by face-to-face home interviews, have led many statistical agencies to consider alternative methods of data collection. Sample selection via RDD (Random Digit Dialling) -Waksberg (1978) - and the use of CATI (Computer Assisted Telephone Interviewing) - Nicholls and Groves (1985) - have greatly enhanced the benefits of telephone surveys.There is no doubt that sampling and interviewing by telephone is indeed fast becoming "a major development in the history of survey methods" - Groves and Kahn (1979) - albeit at present primarily in North America.

In most countries outside North America the use of telephone interviewing has been developing more slowly, although Christofferson (1984) reports increasing use of telephone interviewing in the Scandinavian countries. However, the use of the telephone for surveys, outside North America, is still at the most partial or supplementary. Its primary use in many cases is for follow-up, contacting not-at-

homes and for screening.The telephone is also widely used in surveys based on non-probability samples (e.g. quota sampling). There are several reasons for the limited use of telephone interviewing outside North America.The main reason is that telephone coverage of households is far from complete in most countries, especially with respect to the rural population. Thus any probability sample survey would of necessity be a mixed-mode survey, requiring supplementing of telephone interviewing by home visits or mail follow-up for the non-telephone households. Furthermore, in many countries, especially in Europe, the use of the telephone is often still limited to important business and social matters and its use for interviewing is perceived as an unwarranted intrusion of privacy.

Another reason is the feeling, often expressed by field staff, that response to telephone interviews is, in some way, less accurate than that obtained by face-to-face interview in the respondent's home. There seems to be little empirical evidence to support this claim and several studies have found only small differences, if at all, between responses obtained by different modes of collection - Kantorowitz and Nathan (1987), Rogers (1976), Schuman and Presser (1981). Raw comparisons between responses obtained by telephone surveys with those obtained from home visits are difficult to assess because of confounding between the effects of mode of collection (telephone interview or home visit) and the effect due to differences in characteristics between households without telephones (and those not willing to answer by telephone) and households with telephones who allow telephone interviewing.

Well designed experiments of the split ballot type, in which subjects are randomly allocated to mode of collection, partially overcome this difficulty, but are difficult to implement with effective control - Hochstim (1967), Locander, Sudman and Bradburn (1974), and Rogers (1976). Moreover, while it is relatively simple to limit the comparison to households with telephones, it is more difficult to neutralize the effect of unwillingness to be interviewed by telephone. The answers to a hypothetical question on such willingness in a home interview must be treated with extreme caution. Obviously,"pure" response error effects of mode of collection can best be assessed by independent measurements for the same subject by telephone and by home interview. This has been done by reinterviewing via the telephone subjects who have previously responded to a home interview - e.g. Rogers (1976). Simple reinterviewing will allow a comparison but will not provide an assessment of the measurement errors for which, in general, at least two measurements for each mode are required for each subject.

4.3 A misclassification model for mode of collection effects.

There is an inherent difficulty in attaining independent observations from the same subject over a short period of time, while over longer periods the characteristics are usually subject to change. In the following we consider the possibility of using a multi-round survey, with relatively long time intervals between interviews, for assessment of the effect of mode of collection on characteristics which are invariant over time, at least over the period of the survey

(such as last school attended for those not studying during the survey period). This is an extension of the general treatment of response error micro-effects from repeated surveys, given in Kantorowitz and Nathan (1987), without a mode of collection effect. A misclassification model is proposed, from which, under certain assumptions, both "true" class proportions and misclassification probabilities can be estimated, both for responses obtained by telephone interview and for responses obtained by home interview, for the same set of respondents. These misclassification probabilities can be assessed for the set of respondents who exhibit willingness to respond via telephone interviews (by actually participating in at least one telephone interview), and can be compared with those estimated for the set of respondents from whom only home interviews are obtained (due to the lack of a telephone or to their unwillingness to be interviewed over the telephone).

We consider the population as being divided by a qualitative polytomous variable into categories with unknown probability, R_k, of belonging to category k, ($k = 1,\ldots,c$ and $\Sigma_k R_k = 1$). At each round of a multi-round survey, each unit is classified (correctly or not) into one of the c categories. Let $P_{kk'}^{(j)}$ denote the conditional probability that a unit reports its category as k' by mode of collection j ($j = 1,\ldots,m$), given that its true category is k ($k,k' = 1,\ldots,c$ and $\Sigma_{k'} P_{kk'}^{(j)} = 1$).

We assume that the probabilities, $P_{kk'}^{(j)}$ of misclassification to category k' (for $k \neq k'$) and of correct classification (for $k' = k$) are constant over rounds for the same mode of collection, We also assume that classification is independent (conditionally, given the true category) over rounds.

Let x_{jk} be the number of times a given unit is classified as belonging to category k by mode of collection j, ($j=1,\ldots,m$; $k=1,\ldots,c$);

let $x = (x_{11},\ldots,x_{1c},\ldots,x_{m1},\ldots,x_{mc})$ be the unit's mc-component observation vector;

and let $r_j(x) = \Sigma_k x_{jk}$ be the number of rounds the unit is observed by mode of collection j.

Then, under the above assumptions, the probability of observing x can be expressed as:

$$\pi(x) = \Sigma_k R_k \Pi_j r_j(x)! \left\{ \prod_{k'} \frac{\left[P_{kk'}^{(j)}\right]^{x_{jk'}}}{x_{jk'}!} \right\} \qquad (4.1)$$

If we assume that a simple random sample of n independent observations on x is obtained, then the distribution of the frequencies of observing x, f(x), is multinomial with probabilities $\pi(x)$, where $\Sigma_x f(x) = n$ and $\Sigma_x \pi(x) = 1$.

Thus the kernel of the log likelihood of the observations is given by:

$$\ln L = \Sigma_x f(x) \ln p(x) \qquad (4.2)$$

The maximum likelihood estimators of the parameters R_k and $P_{kk}^{(j)}$ are then obtained by equating the appropriate partial derivatives of the Lagrange function for (4.2) to zero. Explicit solutions are, in general, not available but a simple iterative procedure is obtained by successive substitution of the values of f(x) from one set of parameter values to obtain the next set. This procedure can be shown to be equivalent to the EM algorithm and to converge to maximum likelihood estimators of the parameters - see details in Nathan and Eliav (1988).

4.4 Application of the misclassification model.

The misclassification model described above was applied to a set of data on educational characteristics derived from four rounds of a single panel of the Israel Labour Force Survey. The data set was derived for the general analysis of response error effects in Kantorowitz and Nathan (1987) and its characteristics and sources are described in detail therein. Following is a brief summary of the aspects of the data set most salient to the present analysis. As pointed out previously, the models can only be applied to characteristics which are invariant over time. Most of the variables measured in Labour Force Surveys or other current surveys either change over time (e.g. labour force characteristics), or are obtained only at the first interview (e.g. date of birth). In the Israel Labour Force Surveys the set of educational questions shown in exhibit 4.1 is asked independently at each of the rounds (up to four) in which the household participates, for each individual aged 14 and over. For those who did not attend school during the whole period of the survey (ascertainable via question 8), the variables years of study (question 9) and type of last school attended (question 10) can be regarded as invariant over the rounds in which the respondent participated.

The Israel Labour Force Survey - Central Bureau of Statistics (1987) - is a current rotating panel survey, with four panels investigated each quarter for urban localities. Each dwelling unit in the survey is investigated for two consecutive quarters and after a break of two quarters, for two additional consecutive quarters. The sample design is single-stage stratified for large localities, each locality serving as a stratum, and two-stage for smaller localities, with stratified PPS selection of localities in the first stage. Selection of dwelling units within selected localities is random systematic from lists of units, with equal final inclusion probabilities. The final sample for urban localities can be considered as approximately a simple random sample of dwelling units. Since all persons aged 14 and over in selected dwelling units are investigated,the sample of individuals is clustered. Face-to-face home interviews are carried out for most households in the first and last rounds. Telephone interviews are often used in the second and third rounds, at the discretion of the interviewer and after receiving the

EXHIBIT 4.1

Educational questions in Israel Labour Force Survey

8. Have you attended school or do you attend school now?

1 Attended only in the past
2 Presently attending (even if on vacation)
3 Never attended school ----> skip to question 11

9. For how many years have you attended school? ☐☐

10. What is the type of school last attended?

01 Primary school
02 Intermediate school
03 Vocational or agricultural secondary school
04 Secondary school
05 Yeshiva
06 Teacher training college
07 Technical post-secondary school
08 Other post-secondary school
09 Academic institution
10 Other, specify

household's consent during the first interview. A small number of responses are obtained by mail.The respondent rule allows any adult member of the household to answer for all others. Overall non-response runs to about 13%.

The data set was based on the population in urban Jewish localities (about 86 percent of the total) who participated in the panel first investigated in the last quarter of 1980. Only the 4084 persons matched by survey identity number, sex and age and who reported consistently that they attended school only in the past (question 8) were retained. The distribution of responses obtained from these persons in each round, by mode of collection, is given in the following table:

Table 1: Responses by mode of collection in each round

		Mode of Collection		
Round	Total	Home visit	Telephone	Mail
1	3105	3020	67	18
2	3216	1767	1435	14
3	3109	1554	1542	13
4	3065	2932	110	23

For the present analysis, only respondents who were investigated in at least two rounds were included, to ensure identifiability, and mail responses were excluded. In addition, 69 persons who responded only be telephone interview were excluded. The final data set included 11,901 responses for 3435 persons.

This sample was divided into two: (1) those who responded at least once by home visit and at least once by telephone interview - "telephone households"- representing the population of households with telephones who could be reached and were willing to respond both by home visit and by telephone; and (2) those only responding by home visits- "non-telephone households" - representing the population of households without telephone or unwilling to respond by telephone. The breakdown is given in table 2 (exhibit 4.2).

It should be emphasized that responses obtained over the different rounds for the same unit (whether by home visit or by telephone) were usually obtained by the same interviewer.This implies that the confounding of interviewer effect and that of mode of collection or of round is limited. This fact could, on the other hand, cause between-round response dependence. However, it should be noted that the time lags between rounds, ranging from three to fifteen months, are such that this dependence must be very small. Clustering of units (an average of 2.7 persons per household) may cause some departure from the assumption of independence between the units.

The results of applying the misclassification model to each of the sub-populations are given in exhibit 4.3 - table 3 for groups of years

EXHIBIT 4.2

Table 2: Persons responding (by home visit or telephone) in two or more rounds and their responses by mode of collection*

	Persons	Responses		
		Total	Home visit	Telephone
Total	3381	11901	8896	3005
"Telephone households" (both modes)	1840	6696	3691	3005
"Non-telephone households" (only home visits)	1541	5205	5205	--

* Responses by mail and those responding only by telephone excluded.

Source: Nathan and Eliav (1988)

EXHIBIT 4.3

Table 3. Estimates of misclassification probabilities for groups of years of study by type of household and by mode of collection (percentages)

True category k	Estimated percentage $\hat{R}_k$	Estimated conditional probability of reporting category k' – $\hat{P}(l)_{kk'}$			
		1–8	9–10	11–12	13+
Telephone households – response by telephone interview					
1–8	28.2	94.7	3.6	1.3	0.3
9–10	16.2	5.1	85.2	8.6	1.1
11–12	32.5	1.5	1.1	94.4	2.9
13+	23.1	0.2	0.4	1.8	97.6
Telephone households – response by home visit					
1–8	28.2	90.5	5.3	3.4	0.7
9–10	16.2	10.1	75.9	12.6	1.5
11–12	32.5	1.5	4.4	89.5	4.6
13+	23.1	0.0	0.9	4.1	95.0
Non-telephone households – (response by home visit)					
1–8	30.8	95.5	3.0	1.4	0.2
9–10	21.7	9.2	80.9	8.8	1.2
11–12	29.7	2.8	5.2	87.8	4.2
13+	17.9	0.9	0.8	4.8	93.5

Table 4. Estimates of misclassification probabilities for last school attended by type of household and by mode of collection (percentages)

True category k	Estimated percentage $\hat{R}_k$	Estimated conditional probability of reporting category k' – $\hat{P}(l)_{kk'}$			
		Primary	Vocational	Secondary[1]	Academic
Telephone households – response by telephone interview					
Primary	28.8	92.9	1.7	5.3	0.1
Vocational	23.7	3.0	86.7	9.9	0.4
Secondary[1]	30.7	2.7	5.5	89.4	2.4
Academic	16.8	0.2	0.4	5.0	94.4
Telephone households – response by home visit					
Primary	28.8	90.4	3.5	5.9	0.2
Vocational	23.7	5.2	80.8	13.8	0.1
Secondary[1]	30.7	3.2	7.1	87.9	1.7
Academic	16.8	0.2	0.4	6.6	92.9
Non-telephone households – (response by home visit)					
Primary	33.0	93.9	2.2	4.0	0.0
Vocational	28.2	4.7	84.3	10.6	0.4
Secondary[1]	28.8	4.3	8.9	84.8	2.1
Academic	10.0	0.0	1.0	2.7	96.3

[1] Includes post-secondary schools.

Source: Nathan and Eliav (1988)

of study and table 4 for last school attended. The results for groups of years of study show clearly that, for telephone households, telephone interviewing results in less misclassification than home visits. For all five categories the probabilities of correct classification (underlined) are higher for telephone interview than for home visit, the greatest difference being that for the group with less then 8 years of study (almost 9%). The overall probability of misclassification, for the telephone households is estimated as 14.1% for home visits as against 8.7% for telephone interviews. For non-telephone households the overall probability of misclassification (for home visit) is 12.7%, somewhat lower than that for telephone households, but still much higher than for telephone interviews (for telephone households). Probabilities of correct classification in categories of groups of years of study for non-telephone households are mostly lower than those for telephone households by telephone interview and higher than those for telephone households by home visits.

Similar but less striking results are obtained for last school attended, although for one category (post-secondary schools) the probability of correct classification by home visit (88.3%) is somewhat higher than that for telephone interviews (83.4%).However, this is a small category (6.2% of the total) and overall the advantage of telephone interviewing is still clear, with its estimated misclassification rate of 10.3%, against 12.6% for home visits. For non-telephone households the overall misclassification is again intermediate -11.7%. Correct classification probabilities range from a low of 82.3% for secondary school (lower than those for telephone households by both modes) to a high of 96.3% for academic (higher than those for telephone households).

The misclassification model provides, besides estimates of misclassification probabilities, also estimates of the true proportion of units in each category, $\hat{R}_k$. These are given (see exhibit 4.3) in the first column of table 3 for groups of years of study and of table 4 for last school attended, separately for telephone households and for non-telephone households.

As might be expected, telephone households have, overall, a somewhat higher educational level than non-telephone households. Thus telephone households have higher proportions than non-telephone households in the categories 12+ years of study and for last school academic and secondary (but not for post-secondary). The distribution estimated by taking misclassification into account can be compared with the expectation of the distribution as reported by each mode of collection and for each sub-population, and the relative biases of the estimates of proportions can be estimated. The results are shown in tables 6 and 7 (exhibit 4.4) and they show, for the sub-population of telephone households, the superiority of telephone interviewing over home visits, with respect to biases in estimating a distribution, under the assumptions of the model. Both for groups of years of study and for last school of study, absolute relative biases in the proportions which would have been obtained by telephone are lower than those which would have been obtained by home visits in all but one category. Overall there is a reduction in the mean absolute relative

EXHIBIT 4.4

Table 6. Effect of mode of collection on biases in distribution by groups of years of study

Category	Estimated percentage	Expected percentage reported by		Relative bias (percentage)	
		Telephone interview	Home visit	Telephone interview	Home visit
Telephone households					
1–8	28.2	28.1	27.6	−0.5	−2.0
9–10	16.2	15.3	15.4	−5.8	−4.8
11–12	32.5	32.0	33.0	+1.1	+1.6
13+	23.1	23.8	23.9	+2.8	+3.4
				(2.6)[1]	(2.9)[1]
Non-telephone households					
1–8	30.8	·	32.4	·	+5.2
9–10	21.7	·	20.2	·	−7.1
11–12	29.7	·	29.3	·	−1.4
13+	17.9	·	18.3	·	+2.3
					(4.0)[1]

[1] Average, absolute values.

Table 7. Effect of mode of collection on biases in distribution by last school attended

Category	Estimated percentage	Expected percentage reported by		Relative bias (percentage)	
		Telephone interview	Home visit	Telephone interview	Home visit
		Telephone households			
Primary	28.8	28.3	28.3	−1.6	−1.8
Vocational	23.7	22.8	22.4	−3.8	−5.5
Secondary	30.7	32.2	33.1	+4.8	+7.7
Academic	16.8	16.7	16.2	−0.5	−3.5
				(2.7)[1]	(4.6)[1]
		Non-telephone households			
Primary	33.0	.	33.6	.	+1.7
Vocational	28.2	.	27.2	.	−3.7
Secondary	28.8	.	29.0	.	+0.7
Academic	10.0	.	10.3	.	+3.5
					(2.4)[1]

[1] Average, absolute values.

Source: Nathan and Eliav (1988)

bias for both variables - from 3.4% to 1.6% for groups of years of study and from 4.7% to 2.6% for last school attended.

Since telephone households represent only about a half of all households, an attempt was made to assess the overall impact of using telephone interviews for telephone households. The expected proportions were estimated separately for non-telephone households and the resultant biases estimated as above. The results show that the biases for non-telephone households are neither consistently higher nor lower than those attained by home visits from telephone households. The estimated proportions and the expected proportions for the sub-populations were combined to obtain estimates for the total population, by weighting according to the sample representation (54% - telephone households). The results still indicate the superiority of telephone interviews even if only used partially. For last school attended absolute biases for telephone interviewing are consistently lower than for home visits - with a mean absolute biases of 1.7% against 2.7%. For groups of years of study absolute biases for telephone interviews are lower than those for home visits for all but one category - with a mean absolute bias of 2.3% against 2.8%.

To summarize,its limitations notwithstanding, the misclassification model indicates systematic differences between responses obtained for the same units by home visits and by telephone interviews. Although the differences are not substantial, they are clear indications that inconsistency in response - as measured by the rate of misclassification for qualitative variables - is higher for responses obtained by home visits than for those obtained by telephone interviews and that biases in estimating a distribution can be reduced by the use of telephone surveying.

4.5 Micro-methods of response error evaluation.

As pointed out already, when individual data are available, comparisons based on these data will, in general, be preferable to evaluation based on aggregate data. The main micro-method of evaluation is the well-known technique of re-interview. This has already been treated in some detail in section 1. It still remains the prime method of census evaluation, although its use in survey work is considerably limited due to the small sample sizes, usually employed. In countries with an extensive network of administrative data, which are accessible and available for evaluation, the use of administrative record checks may often be a reliable and relative inexpensive alternative. In Israel, a well-established register, which is maintained relatively well on a current basis, has often served as a source for record checks. This could be attained by utilizing the unique identity number every inhabitant receives at birth or on immigration, which serves a variety of administrative purposes as a base for record-linkage. Details on some of these uses can be found in Bachi, Baron and Nathan (1967).

A less well-known micro-method of evaluation is based on the technique known as surveys with multiplicity. The basic idea is that, for instance in a survey on demographic events, instead of each household reporting only on events which occurred in the household (or

to its members), respondents are asked to report also about events occurring to persons outside the household, to which they are linked by some well-defined counting rule. For example, each respondent may be asked to report on births to his siblings, whether they reside in the same household or not. By providing also information about the number of possible links (e.g the number of different households in which the respondent's siblings reside), unbiased estimates of the total number of events can easily be obtained - see Sirken (1970). While the original idea of surveys with multiplicity was to reduce sampling variance, which indeed it usually does, it was soon realized that this might well be at the expense of increased response bias and variance. This is due to the fact that more distant relatives may not be able to report as completely on events occurring outside their household as they could on closer events.

It turns out that surveys with multiplicity contain a built-in mechanism, which enhances the possibility of carrying out an evaluation survey in conjunction with the survey. This is done by actually following-up a sub-sample of the designated relatives of the respondents in the surveys and getting from them reports in the same way as in the original survey. A detailed model, which ensures that unbiased estimates of the components of the mean square error (response bias and variance and the sampling variance) can be obtained at relatively low costs - see Nathan (1976). The model was applied to an experimental survey carried out by the Israeli Central Bureau of Statistics with a sample of some 4,000 households. Reports were requested on births and marriages which occurred during a calendar year, either to persons in the survey household or to a designated set of relatives (daughters and sisters of women in the household for births; offspring and siblings of all household members for marriages). Full details can be found in Nathan, Schmelz and Kenvin (1977).

The main results are given in Table 1 (exhibit 4.5). They give the components of mean square error for the conventional survey (without multiplicity) and for two variations of the multiplicity survey (restricted and full counting rules). They show, both for births and for deaths, that while response variance and bias are larger for the multiplicity survey, the reduction in sampling variance offsets this and there is an overall gain in mean square error of 40-50%.

EXHIBIT 4.5

1. Estimates of Components of Error by Counting Rule

Component	Counting rule					
	Births			Marriages (persons marrying)		
	Conventional	Restricted multiplicity	Full multiplicity	Conventional	Restricted multiplicity	Full multiplicity
			Estimate			
Current demographic estimate	64,250	64,250	64,250	49,820	49,820	49,820
Revised survey estimate	59,970	60,526	58,016	45,130	42,781	42,518
Basic survey estimate	60,018	62,305	58,952	43,916	46,325	43,586
Estimate of net bias	+48	+1,780	+935	−1,214	+3,544	+1,068
−Undercoverage bias	−1,223	−2,857	−4,766	−1,214	−1,437	−3,618
−Overcoverage bias	+1,271	+4,637	+5,701	0	+4,980	+4,687
			Components of MSE ($\times 10^{-3}$)			
Total MSE	10,131	10,106	6,274	8,997	18,365	5,485
Squared bias	2	3,168	874	1,474	12,557	1,141
Total variance	10,129	6,938	5,400	7,523	5,808	4,344
Sampling variance	9,825	5,900	3,815	7,307	4,969	3,154
Response variance	303	1,038	1,584	216	839	1,190
			Relative standard errors (percentages)			
Total root MSE	5.31	5.25	4.32	6.65	10.02	5.51
Bias (absolute value)	0.08	2.94	1.61	2.69	8.28	2.51
Sampling standard error	5.23	4.01	3.37	5.99	5.21	4.18
Response standard error	0.92	1.68	2.17	1.03	2.14	2.57

Source: Nathan (1967)

REFERENCES

Andersen, R., Kasper, J., Frankel, M.R. et al. (1979). **Total Survey Error.** San Francisco: Jossey-Bass.

Bachi, R., Baron, R. and Nathan, G. (1967). Methods of record linkage and applications in Israel. **Bulletin of the International Statistical Institute, 41,** 766-786.

Belson, W. A. (1981). **The Design and Understanding of Survey Questions.** London, Gower.

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

Brewer, K.R.W. (1981). Estimating marihuana usage using randomized response - some paradoxical findings. **Australian Journal of Statistics, 23,** 139-148.

Brown, G.H. and Harding, F.D. (1973). **A Comparison of Methods of Studying Illicit Drug Usage.** Technical Report 73-9, Human Resources Research Organization, Alexandria, VA.

Central Bureau of Statistics (1987). **Labour Force Surveys 1985.** Special Series No. 801, Jerusalem.

Christofferson, M.N. (1984). The quality of data collected at telephone interviews. **Statistisk Tidskrift,** 1984:1, 27- 35.

Dalenius, T. (1977). Bibliography on non-sampling errors in surveys. **International Statistical Review, 45,** 71-89; 181-197; 303-317.

Fellegi, I.P. (1964). Response variance and its estimation. **Journal of the American Statistical Association, 59,** 1016-1041.

Fienberg, S. E. and Tanur, J. M. (1989). Combining cognitive and statistical approaches to survey design. **Science, 243,** 1017-1022.

Groves, R. M. and Kahn, R. L. (1979). **Surveys by Telephone: A National Comparison with Personal Interviews.** New York: Academic Press.

Goodstadt, M.S. and Gruson, V. (1975). The randomized response technique: a test on drug use. **Journal of the American Statistical Association, 70,** 814-818.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G. (1977). Respondent hazards in the unrelated randomized response model. **Journal of Statistical Planning and Inference, 1,** 53-60.

Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961). Measurement errors in censuses and surveys. **Bulletin of the International Statistical Institute, 38(2),** 359-374.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. **Journal of the American Statistical Association, 62,** 976-989.

Kantorowitz, M. (1969). **Evaluation of the Census Data.** 1961 Census of Population and Housing No. 40, Central Bureau of Statistics, Jerusalem.

Kantorowitz, M. and Nathan, G. (1987). The estimation of response error micro-effects from repeated surveys for invariant characteristics. **Proceedings of the Third Annual Research Conference,** Bureau of the Census, 359-390.

Krotki, K.J. and Fox, B. (1974). The randomized response technique, the interview and the self administered questionnaire: an empirical comparison of fertility reports. **American Statistical Association, Proceedings of the Social Statistics Section,** 367-371.

Lessler, J. T., and Sirken, M. G. (1985). Laboratory-based research on the cognitive aspects of survey methodology: The goals and methods of the National Center for Health Statistics study. **Milbank Memorial Fund Quarterly/Health and Society, 63,** 565-581.

Leysieffer, F.W. and Warner, S . (1976). Respondent jeopardy and optimal designs in randomized response models. **Journal of the American Statistical Association, 71,** 649-656.

Locander, W., Sudman, S. and Bradburn, N. (1976). An investigation of interview method, threat and response distortion. **Journal of the American Statistical Association, 71,** 269-275.

Lord, F. and Novick, R. N. (1968). **Statistical Theories of Mental Test Scores.** Reading, MA: Addison-Wesley.

Mahalonobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. **Journal of the Royal Statistical Society, 109,** 325-370.

Marquis, K. H., Marquis, M. S. and Polich, J. M. (1986). Response bias and reliability in sensitive topic surveys. **Journal of the American Statistical Association, 81,** 381-389.

Nathan, G. (1973). Utilization of information on sampling and non-sampling errors for survey design. **Bulletin of the International Statistical Institute, 45 (3),** 393-406.

Nathan, G. (1976). An empirical study of response and sampling errors for multiplicity estimates with different counting rules. **Journal of the American Statistical Association, 71,** 808-815.

Nathan, G. (1987). The use of linear models and misclassification models to assess response error effects of mode of collection. **Bulletin of the International Statistical Institute, 52,** 213-230.

Nathan, G. (1988). A bibliography on randomized response: 1965-1987. **Survey Methodology (Canada), 14,** 331-346.

Nathan, G. and Eliav, T. (1988). Comparison of measurement errors for telephone interviewing and home visits by misclassification models. **Journal of Official Statistics, 4,** 363-374.

Nathan, G., Schmelz, U. O. and Kenvin, J. (1977). **Multiplicity Study of Births and Marriages in Israel.** National Center for Health Statistics, Ser. 2., No. 70, Washington, D.C.

Nathan, G. and Sirken, M. G. (1986). Response error effects of survey questionnaire design. **Proceedings of the Section on Survey Research Methods, American Statistical Association,** 493- 498.

Nathan, G. and Sirken, M. G. (1987). Optimal allocation to control questionnaire design variance in sample surveys. **Proceedings of the Section on Survey Research Methods, American Statistical Association,** 252-255.

Nathan, G. and Sirken, M. G. (1988). Cognitive aspects of randomized response. **Proceedings of the Section on Survey Research Methods, American Statistical Association,** 173-178.

Nicholls, W.L. and Groves, R.M. (1985). The status of computer assisted telephone interviewing. **Bulletin of the International Statistical Institute, 51,** 18.1.

Pratt, J.W., Raiffa, H. and Schlaifer, R. (1964). The foundations of decision under uncertainty: an elementary exposition. **Journal of the American Statistical Association, 59,** 353-375.

Pritzker, L. and Hanson, R. H. (1962). Measurement Errors in the 1960 census of population. **Proceedings of the Social Statistics Section, American Statistical Association,** 80-90.

Rogers, T.F. (1976). Interviews by telephone and in person: quality of responses and field performance. **Public Opinion Quarterly, 40,** 51-65.

Rolnick, S.J., Gross, C. R., Garrard, J. and Gibson, R. W. (1989). A comparison of respones rate, data quality, and cost in the collection of data on sexual history and personal behaviors, **American Journal of Epidemiology, 129,** 1052-1061.

Schuman, H. and Presser, S. (1981). **Questions and answers in attitude surveys.** New York: Academic Press.

Sirken, M.G. (1970). Household surveys with multiplicity. **Journal of the American Statistical Association, 65**, 257-266.

Sirken, M.G., Mingay, D.J., Royston, P.N., Bercini, D.H. and Jobe, J.B. (1988). Interdisciplinary research in cognition and survey measurement. In **Practical Aspects of Memory: Current Research and Issues: vol 1. Memory in Everyday Life**, M. M. Greenberg, P. E. Morris and R. N. Sykes - eds., Chichester, England: Wiley, pp. 531- 536.

Sirken, M.G., Nathan, G. and Thornberry, O. (1989). Evaluation of questionnaire design effects in a national health survey. **Bulletin of the International Statistical Institute, 53**, 557-576.

Soeken, K.L. and Macready, G.B. (1982). Respondents' perceived protection when using randomized response. **Psychological Bulletin, 92**, 487-489.

Tourangeau, R. (1984). Cognitive science and survey methods. In: T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau (Eds.), **Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines**, Washington, DC: National Academy Press, pp. 73-100.

U. S. Bureau of Census (1960). **The Post-Enumeration Survey: 1950**, Technical Paper No. 4, Washington, DC.

von Neumann, J. and Morgenstern, O. (1944). **Theory of Games and Economic Behavior**. Princeton, NJ: Princeton University Press.

Waksberg, J. (1978). Sampling methods for random digit dialling. **Journal of the American Statistical Association, 73**, 40-46.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. **Journal of the American Statistical Association, 60**, 63-69.

Willis, G. B., Royston, P. and Bercini, D. (1989). The use of verbal report methods in the development and testing of survey questionnaires, (unpublished).