Journal of Official Statistics, Vol. 14, No. 3, 1998, pp. 277–295

Design and Analysis of Experiments Embedded in Sample Surveys

Jan van den Brakel and Robbert H. Renssen[1]

This article focuses on the design and analysis of large-scale field experiments embedded in ongoing surveys. Historically, experimental designs and sampling theory have formed two separated domains in applied statistics. However, in the design and analysis of embedded experiments, statistical methods from experimental designs and sampling theory can be combined in order to improve the quality and efficiency of these experiments. We will discuss how parallels between randomized experiments and random survey samples can support the design of field experiments embedded in ongoing surveys. This is illustrated with a series of embedded experiments conducted at Statistics Netherlands. This article focuses on a design-based approach for the analysis of such experiments. Generally, the objective of embedded field experiments is to draw inferences on finite population parameters. To this end the analysis has to include the probability structure imposed by the applied sampling design of the survey as well as the randomization of the applied experimental design used to assign the experimental units to the different treatments. We worked out such a design-based method for the analysis of the two-sample problem.

Key words: Experimental design; sampling design; Horvitz-Thompson estimator; generalized regression estimator; two-sample problem.

1. Introduction

Traditionally, the design and analysis of experiments and sampling theory form two separate domains of applied statistics. These two fields, however, come together in situations where experiments are performed in order to investigate possible improvements of a sample survey process. Statistical methods used for the design and analysis of experiments are highly appropriate for obtaining quantitative information on the effect of alternative survey methodologies on response behavior and/or estimates of population parameters of a survey. For example, at Statistics Netherlands, the effects of alternative questionnaire designs, different approach strategies, or advance letters have been tested by means of large-scale field experiments embedded in ongoing surveys. Methods from experimental designs can also support quality control of the survey process. For instance, the bias and various variance components in total measurement error models (Forsman 1989

[1] Statistics Netherlands, Department of Statistical Methods, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

Acknowledgments: This article is based on a paper presented at the International Conference on Survey Measurement and Process Quality held in Bristol, United Kingdom, on April 1–4, 1995 (Van den Brakel and Renssen 1996a). The authors wish to thank Prof. Stephen E. Fienberg (Department of Statistics, Carnegie Mellon University, Pittsburgh, U.S.A.) for his encouraging discussions and reading of drafts, and the referees for commenting on an earlier version.

and Biemer et al. 1991) can be estimated by using principles from the field of experimental designs.

Fienberg and Tanur (1987, 1988, 1989, 1996) reviewed the development of statistical methods in random sampling and randomized experimentation by Fisher and Neyman in the 1920s and 1930s. They discussed the parallels between the concepts in randomized experiments and random sample surveys and emphasized that the statistical methodology used in both fields is essentially the same. Based on these parallels, they emphasized how random samples can be embedded in randomized experiments or vice versa in order to increase the internal as well as the external validity of the conclusions drawn from experiments.

In this article we describe how statistical methods from the theory of experimental designs can support research aimed at improving the survey process. In Section 2, we briefly review the principles of experimental designs and sampling theory. We distinguish between small-scale experiments conducted in laboratory circumstances and large-scale field experiments embedded in ongoing surveys. With respect to the large-scale field experiments we are mainly interested in effects on finite population parameters. We discuss how parallels between experimental designs and sampling theory can be used in the design of efficient embedded experiments (Section 3) and illustrate this with a series of practical examples of embedded experiments conducted at Statistics Netherlands (Section 4). For the analysis of large-scale field experiments embedded in ongoing surveys, Fienberg and Tanur (1987, 1989) emphasize model-based estimation. As an alternative, we put more emphasis on design-based methods in this article. In embedded experiments, if the focus of interest is the effects of treatments on estimates of finite population parameters, then we argue that the complexity of the sampling design should be taken into account by relying on sampling theory (Section 5). Such a design-based method is proposed and worked out for the analysis of the two-sample problem embedded in complex sampling design (Section 6).

2. Concepts of Experimental Designs and Sampling Theory

2.1. Experimental designs

The objective of experimental designs is to obtain relevant quantitative information about the effects of different treatments and their mutual interactions. The principles of experimental designs, such as replication, randomization, local control for sources of variation by skillful grouping of the experimental units (e.g., randomized block designs and row and column designs), the use of factorial designs and covariance analysis, were developed mainly on the basis of Fisher's work (1935).

A minimum number of replications of the treatments is required to discover statistically significant treatment effects if they exist. Randomization is used to prevent experimental errors systematically distorting the measurement of the treatment effects, by ensuring that each treatment has an equal probability of being favored or handicapped by an extraneous source of variation. Local control by means of randomized block designs, row and column designs, or covariance analysis is applied to reduce sources of extraneous variation in order to obtain more precise estimates of the treatment effects. Simultaneous testing of different treatments by means of factorial designs is applied to increase the efficiency of experimentation and to test possible interactions between the different treatments.

Statistical methods from the theory of design and analysis of experiments are basically model dependent. Often, the observations in an experiment, taken at different levels of different factors, are assumed to be the outcome of a stochastic variable, which is modeled according to a linear regression model with the different levels (and possibly also covariables) as explanatory variables. The disturbance terms are assumed to be identically and independently distributed (IID) with expectation zero. The factors may concern treatment variables as well as local control variables for sources of extraneous variation (e.g., block variables, row and column variables or covariables). The disturbance terms concern the extraneous variation insofar as it is not explained by local control and/or by the covariables. An important question in the model assumption concerns whether the levels at which a factor is measured are fixed or whether they have to be considered as a random sample from a large population. In the former case the effects are called fixed, whereas in the latter the effects are random. A set of effects is often considered fixed when the statistical inference only concerns the levels included in the experiment, and as random when the inference extends to a population from which the levels were supposedly drawn.

The main purpose of experimental designs is to make statistical inferences about the treatment effects. Statistical models play a central role in testing hypotheses about the significance of model parameters which are assumed to reflect the treatment effects and their interactions, and in exploring relationships between variables. Based on the regression model and its assumptions, efficient test statistics are derived in order to test hypotheses concerning the corresponding regression parameters. When the regression model is correctly specified, the treatment effects are indeed reflected in the corresponding regression parameters and the stated hypotheses concerning the regression parameters also concern the treatment effects. Note that the principles of experimentation (such as replication, randomization, local control and factorial experimentation) are applied to ensure that observed treatment effects actually can be attributed to the parameters of the statistical model. If the model is misspecified, the validity of the conclusions with respect to the treatment effects depends on the robustness of the test statistics against the type of misspecification involved. In general, conclusions based on efficient test statistics derived under a specific model cannot be generalized to situations outside that model. Kempthorne (1952) and Hinkelmann and Kempthorne (1994) suggested a design-based approach for the analysis of experiments by elaborating on randomization theory in a way similar to the approach in survey analysis.

To summarize, statistical methods of experimental designs are mainly intended to guarantee a sufficient internal validity of the experiment, and to estimate differences in treatment effects as precisely as possible. The internal validity of an experiment is defined as the extent to which the observed effects in an experiment can be attributed to the differences in the treatments. It thus relates to the cause-effect relationship between the treatments and the observed effects within the experiment itself.

2.2. *Sampling theory*

The purpose of sample surveys is to gather information about a certain finite population by estimating finite population parameters such as means, totals, and fractions. The concept of random sampling has been developed, mainly on the basis of the work of Neyman (1934), as a method to obtain valid estimators for finite population parameters based on representative samples rather than on complete censuses. Neyman (1934) introduced random sampling with unequal selection probabilities by treating optimal allocation in stratified sampling. The concept of random sampling with unequal selection probabilities has been generalized by Hansen and Hurwitz (1943) for random sampling with replacement and by Horvitz and Thompson (1952) for random sampling without replacement as a method to improve the precision of population parameter estimates.

In sampling theory, observations obtained from the sampling units are regarded as fixed. The randomness is introduced because a probability sample is observed instead of the entire target population. In random sampling, the concept of random selection is applied in order to draw statistical inferences about finite population parameters and generalize results from the observed sample to the finite target population from which the sample is drawn.

Statistical methods from sampling theory can be considered as distribution free, that is, no assumptions are made regarding the frequency distribution of the finite population. A proper combination of sampling design and estimation procedure (i.e., the sampling strategy) should give unbiased or nearly unbiased estimates with a minimum variance for the finite population parameters under consideration. An important tool to achieve this is the use of auxiliary information. Such information can be utilized in the sampling design and/or the estimation procedure. In the design stage, techniques like stratification and lattice sampling are applied to increase the precision of the estimators by excluding the variation between homogeneous groups in the finite population from the sampling error. These techniques are similar to randomized block designs and row and column designs from experimental designs. In the estimation procedure, auxiliary information is utilized by means of the regression estimator (with poststratification as a special case) to obtain more precise estimators. This is equivalent to the technique of covariance analysis from experimental designs.

Although auxiliary information was originally used in the design and estimation procedure to improve the efficiency of sampling, nowadays it is an important tool to decrease the bias due to selective nonresponse. Estimators using auxiliary information are generally more robust against selective nonresponse than estimators that do not use auxiliary information (see, e.g., Särndal and Swenson 1987 and Bethlehem 1988).

Statistical models have traditionally played a minor role in the analysis of sample surveys. In the model-assisted approach (see Särndal et al. 1992) it is assumed that the value of each element of the finite population with respect to a certain target variable is a realization of a stochastic variable. This stochastic variable is modeled, e.g., according to a linear regression model, with the values of the auxiliary variables as covariates. Based on the assumed relationship between the target variable on the one hand and the auxiliary variable on the other, a general regression estimator can be derived of which most well-known estimators are special cases. After this estimator is derived, it is judged by its design-based properties, such as design expectation and design variance. The derived formulas hold irrespective of the validity of the model. If the regression model used to derive the estimator does not hold for the finite population, this will result only in higher design variances; not in invalid estimators.

In conclusion, statistical methods from sampling theory are mainly intended to guarantee a sufficient external validity of a survey, i.e., the extent to which the results of the sample can be generalized to the target population.

2.3. Comparison between design and analysis of experiments and sampling theory

Fienberg and Tanur (1987, 1988, 1989) and Van den Brakel and Renssen (1995) discussed the parallels between the statistical methods from experimental designs and sampling theory. Since we further elaborate on these parallels in this article, they are summarized in Table 1.

Besides these parallels, there are also some typical differences between experimental designs and sampling theory which are discussed in the previous sections and are summarized in Table 2. See also Fienberg and Tanur (1987, 1988, 1989) and Van den Brakel and Renssen (1995).

3. Design of Embedded Field Experiments

The conducting of small-scale experiments in laboratory settings is an appropriate and regularly used tool to develop questionnaire designs and interview procedures, or to investigate nonsampling errors in survey processes more systematically (Fienberg and Tanur 1989). The advantage of laboratory experiments is the relative ease with which the effects of a large number of factors can be tested with a high degree of internal validity. The external validity of the results of such experiments, however, is generally not assured. To test the generalization of significant results obtained in such experiments to finite

Table 1. Parallels between design and analysis of experiments and sampling theory

Experimental Designs	Sampling Theory
randomization of experimental units to treatments	random sampling of sampling units from a finite population
replication of the treatments	sample size
randomized block designs	stratified sampling designs
row and column designs	lattice sampling or deep stratification
split-plot designs	two-stage sampling designs
covariance analysis	the general regression estimator

Table 2. Differences between design and analysis of experiments and sampling theory

Experimental Designs	Sampling Theory
stochasticity introduced because observations are assumed to be the outcome of a random variable	stochasticity introduced because a random sample is drawn from a finite population
traditionally model dependent	traditionally design-based
statistical methods mainly intended to guarantee a sufficient internal validity	statistical methods mainly intended to guarantee a sufficient external validity
inference about a hypothetical infinite superpopulation	inference about a finite population

survey populations, large-scale field experiments embedded in sample surveys are very appropriate. Such experiments embedded in ongoing surveys are particularly appropriate if interest is focused on the quantification of the effect of alternative survey methodologies on estimates of finite population parameters. The statistical methods from the theory of experimental designs and sampling theory can be combined in a useful and natural way in the design and analysis of the embedded experiments (see Fienberg and Tanur 1987, 1988). To ensure internal validity, parallels between the concepts of randomized experiments and random sampling should be exploited in the design of embedded experiments to improve the accuracy of estimated treatment effects and to draw correct conclusions about the observed effects. Thus, the sampling design of the survey forms a prior framework for the design of an embedded field experiment. To ensure external validity, statistical methods from sampling theory should support the analysis of these experiments. Because experimental units are selected by means of a probability sample from a finite population, it becomes possible to draw conclusions concerning population parameters. In Sections 5 and 6 we will address the technical aspects of such embedded experiments from the design-based perspective.

3.1. Design of embedded experiments by simple random sampling designs

Suppose that field experiments embedded in ongoing surveys are designed as split sample experiments in order to test the effects of k treatments. In a split sample experiment the sample is randomly divided into k similarly designed interpenetrating subsamples, not necessarily of equal size. Each subsample can be considered as a probability sample from the population and is assigned to one of the k treatments. Now, if the original sample is simple random, then the experiment is in fact a completely randomized design (CRD) (Cochran and Cox 1957, Chapter 4). Generally, this is not the most efficient design available, because no advantage is taken of the possibilities of application of local control for sources of extraneous variation.

An important source of extraneous variation is for example the interviewer effect. In conducting an experiment, it should be avoided that treatments are systematically favored or handicapped because only experienced or inexperienced interviewers are assigned to one particular treatment. In a CRD, this is achieved by assigning the interviewers randomly over the different treatments. It is likely that respondents interviewed by the same interviewer produce more homogeneous answers than respondents interviewed by different interviewers and therefore it is efficient to apply local control for interviewers by means of randomized block designs (RBD) (Cochran and Cox 1957, Chapter 4) with interviewers as block variables (Fienberg and Tanur 1988). When the statistical inference concerning the interviewer effects has to be extended to a larger population from which the interviewers are supposedly drawn, interviewers can be modeled as random components. This leads to an RBD with blocks as random effects, which is equivalent to a split-plot design.

3.2. Design of embedded experiments within more complex sampling designs

Sampling designs are usually more complex than simple random sampling. For instance stratified sampling and two-stage or cluster sampling are frequently applied. In a stratified sampling design there are two ways to divide the sample into k subsamples. Firstly, the whole sample is randomly divided into k subsamples, irrespective of the applied stratification. This may cause some differences in the frequency distribution of the experimental units over the strata among the k subsamples. Secondly, the sample is randomly divided into k subsamples in each stratum. Here the frequency distribution of the experimental units over the strata can be held equal for the k subsamples insofar as this is not disturbed by nonresponse. Disregarding the stratification in the applied sampling design, the first option leads to a CRD. Sampling units from the same stratum are generally more homogeneous than sampling units from different strata. Consequently, the second options leads naturally to an RBD with strata as block variables (Fienberg and Tanur 1988). Also crossings between two or more control variables can be used as block variables (e.g., interviewers and strata). In the first option, local control for the stratification can be applied by means of covariance analysis.

In two-stage sampling designs, three different ways to divide the sample into k subsamples can be distinguished. Firstly, ignoring the primary sampling units, the secondary sampling units of the sample are divided into k subsamples. Secondly, the secondary sampling units within each primary sampling unit are divided into k subsamples. Thirdly, the primary sampling units are divided into k subsamples, and thus all secondary sampling units within a primary sampling unit are assigned to the treatment concerned. Disregarding the structure of the sampling design, the first option leads to a CRD where the secondary sampling units are the experimental units. In the second approach the k treatments are randomly assigned to the secondary sampling units within each primary sampling unit or cluster. Consequently, this type of randomization naturally leads to a split-plot design. Primary sampling units in the sampling design correspond to the whole plots of the split-plot design and the secondary sampling units correspond to the split plots. Fienberg and Tanur (1988) argue that this parallel can be used to design experiments embedded in two-stage samples as split-plot designs in order to eliminate the variance between the whole plots (i.e., the primary sampling units) from the analysis of the treatment effects. This approach is appropriate when sampling units from the same primary sampling units are more homogeneous than sampling units from different primary sampling units. If in the case of the first type of randomization sufficient secondary sampling units are assigned to each of the k treatments within each primary sampling unit (large primary sampling units), then it is still possible to apply local control for the primary sampling units by means of covariance analysis. The third approach leads to a CRD with the primary sampling units as the experimental units. This approach is appropriate if the variation within the primary sampling units is large and the variation between the primary sampling units is small and/or if the primary sampling units are small compared to the number of treatments.

3.3. Design of (double) blind experiments

Field experiments have the advantage that they are conducted in the natural setting of the respondents who do not necessarily know that they are participating in an experiment. If the experimental units or those who conduct the experiment (for instance the interviewers) know that they are participating in an experiment, then the behavior of the experimental units may be altered, perhaps even unconsciously. This type of biased results can be avoided by designing blind or double blind experiments. For example, in experiments where the effects of different questions or different sequences of questions in a questionnaire are compared, skilful use can be made of the possibilities of computer assisted interviewing. The alternative questions, or different orders of questions, can be implemented in the supporting software package. If the routing of the questionnaire already depends on the response of the respondent, interviewer and respondent do not have to know that they are participating in an experiment. In many situations, however, it will be difficult or even impossible to design (double) blind experiments due to the nature of the treatments.

The disadvantages of blocking of interviewers in such situations is that interviewers are aware that they are participating in an experiment. Consequently there is a danger that the interviewers will introduce a bias due to selective behavior. In such situations, the experimenter is faced with the choice between a double blind experiment and an RBD with interviewers as block variables but consequently not double blind. This choice partially depends on the number and the type of treatments and the experience of the interviewers. If the introduction of a substantial bias due to a systematic interviewer effect can be expected because interviewers know that they are participating in an experiment, then a double blind experiment where interviewers are randomly allotted over the treatments is preferable. A less precise comparison is less harmful than a systematic influence on one of the treatments.

Conducting an embedded experiment not double blind involves the danger that the regular survey which besides publication purposes also serves as the control group, will acquire priority above the experimental group and consequently distort the conclusions of the experiment. This is illustrated by an experiment conducted by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census by the redesign of the U.S. Current Population Survey (CPS) (O'Muircheartaigh 1997). In this experiment the new CPS and the old CPS were run in parallel for some period. First, the new CPS was run in parallel with the old CPS as the regular survey, indicating that the new CPS would lead to an increase of the estimated level of unemployment. After the changeover the old CPS was continued in parallel with the new CPS as the regular survey, as a further check on the observed effect. Unfortunately, in this case the old CPS showed a higher estimate of unemployment. O'Muircheartaigh attributes this to the extra effort going into the regular survey on both occasions and to the fact that the interviewers were not blinded. Nevertheless it is a good methodological practice to run the two versions in parallel both before and after the changeover.

4. Embedded Experiments Conducted at Statistics Netherlands

To illustrate the problems encountered by the design of embedded field experiments, a series of embedded field experiments conducted at Statistics Netherlands are described and discussed in the following sections. In these experiments, the sample of an ongoing survey has been randomly divided into one relatively large subsample and one or more smaller subsamples. The large subsample was in fact the regular survey and, besides publication purposes, served as the standard methodology (treatment). The other subsamples were assigned to the alternative treatments so the experiment was conducted in parallel with the regular survey.

4.1. Labour Force Survey

Several field experiments have been conducted to improve the data quality of the Dutch Labour Force Survey (LFS). The LFS is based on a stratified two-stage sampling design with households as the ultimate sampling units. The sample is self-weighted. At the first stage a stratified sample of municipalities is drawn, where strata are formed by geographic areas. At the second stage a sample of addresses is drawn, from each selected municipality. All households at a selected address (with a maximum of three) are included in the sample. The data are collected from personal interviews with hand-held computers (CAPI).

Most of the information of the LFS is gathered by means of retrospective questions, the answers to which are often biased by memory effects. An embedded experiment has been conducted to investigate whether or not the quality of these retrospective data could be improved in terms of consistency and completeness. In the experimental group, a personal calendar indicating important events like festivals, holidays, birthdays, etc. was used during the completion of the retrospective questionnaires. This is intended to give the respondent a guideline for answering the questions, which should minimize memory effects.

Each interviewer worked in a particular interview district. In the experiment, interviewers were randomly allocated to a control group and an experimental group. The interviewers in the experimental group received special instruction in the use of the calendar to assist the completion of the questionnaires. Respondents in a particular interview district were assigned to the group of the corresponding interviewer. So they were assigned to the experimental group or control group, depending on the interviewer district they live in.

A significant decrease of memory errors was observed in the experimental group. However, due to the experimental design, it is unclear if this decrease was caused by the experimental treatment or by an interviewer effect. The extra training of the interviewers in the experimental group could have favored this group systematically, due to extra motivation or attentiveness, and could consequently have distorted the analysis of this experiment. These problems could have been avoided by randomizing households within interview districts over the two different treatments, leading to an RBD with interviewers as block variables or a split-plot design where interviewers correspond to the main plots and households to the split plots. Moreover, if respondents from the same household were more homogeneous with respect to their target variables, then the precision of such an experiment could have been improved by randomizing respondents within each household over the two treatments. This would have led to a split-plot design with interviewers as (fixed) block variables and whereby households correspond to the main plots and respondents to the split plots or a split-plot design with three randomization levels.

4.2. National Travel Survey

The sample design of the Dutch National Travel Survey (NTS) is similar to the sampling design of the LFS (a stratified two-stage sample design with households as the ultimate sampling units). Only the geographical zoning of the applied stratification between the two surveys is different.

The data are collected in a telephone interview as well as a journey diary sent by mail. A few days after the sending of an advance letter, one of the members of the household is contacted by telephone and asked to provide some information about the household situation. Next, diaries are mailed to all household members. Each individual is asked to keep a record of all his/her journeys for one day.

The journey diary has been adjusted several times, for example in order to gather more detailed information concerning travel behavior or to measure carpool behavior. Before new questions are added or other changes in the journey diary are implemented as a standard, they are tested by means of embedded experiments if such changes result in significant differences in the response rates and the estimates of the population parameters of the NTS. This enables us to explain and quantify trend changes in the time series of the population parameter estimates. In another embedded experiment possible effects of the implementation of an informed consent paragraph in the advance letter of the NTS on response rates and estimates of population parameters have been tested (Van den Brakel, Luppes, and Moritz 1995).

In each of these experiments households were randomly divided into a (large) control group and a (small) experimental group. These experiments were carried out as double blind experiments. This was easily achieved, because the experimental factors concern adjustments in the journey diary (sent by mail) or in the advance letter. In the analysis of the response rates, strata were incorporated as block variables and were found highly significant (Van den Brakel, Luppes, and Moritz 1995). In the carpool experiment, a few respondents with extremely long distance travels were assigned to the experimental group. Because the experimental group was relatively small, these outliers had quite a large effect on the estimates of the population parameters and consequently resulted in a highly significant treatment effects, if the analysis is conducted with the *t*-test. In randomized experiments, the concept of randomization ensures that each treatment has an equal probability of being favored or handicapped by an extraneous source of variation, and consequently the observed effects can be assigned to the experimental treatments; however, there is always a small possibility that the results of an experiment are distorted due to an unfavorable outcome of the randomization of the experimental units over the treatments with respect to a covariate. If an experimenter believes that he/she is in this situation, the best option might be to conduct another experiment, but usually this is not feasible due to time and money constraints. As an alternative the two-sample test of Wilcoxon and the two-sample Kolmogorov-Smirnov test were applied to analyze these experiments because these tests are robust against outliers and violations of the normality assumption. These tests could not find a significant difference between the estimates of the population parameters of the experimental group and the control group.

A third experiment was used to test whether two alternative calling schedules could improve the response rates and affect estimates of population parameters of the NTS. In the alternative calling schedules, the same number of calling attempts were more equally spread over the weekdays and the times of day. Households were randomly allotted to one control group and two experimental groups within interviewers. Consequently an RBD with three treatments and with interviewers as block variables was obtained. Interviewer effects turned out to be highly significant with respect to the response rates. The implementation of the alternative calling schedules in the supporting computer system was quite complicated. To preclude distortion of the experiment due to initial problems with the computer system and the behavior of the interviewers in the new situation of the alternative treatments, one week of pretesting preceded the actual experiment. Many unforseeable practical problems which arose during this pretesting could be solved and consequently saved the experiment from systematic distortion of the experimental groups.

4.3. Justice and Security Survey

From 1980 until 1992, the Victim Survey (VS) was conducted in order to measure the frequency of occurrence of particular types of crime. The survey was kept unchanged as long as possible in order to construct crime trends. In 1993 the VS was transformed into the Justice and Security Survey (JSS). In this new survey, several necessary and unavoidable changes were simultaneously implemented (Huys and Rooduijn 1994).

The VS was originally based on a stratified three-stage sampling design with persons as the ultimate sampling units. In the first two stages, households were drawn in a manner similar to the sampling design of the LFS. In the third stage one person was randomly selected from the household. In the VS people were interviewed about events in the preceding calendar year. Interviewing was carried out in January and February by means of CAPI. In the JSS this survey approach was fundamentally changed. In the third stage of the sample design of the JSS two persons (if possible) are now randomly selected. The JSS is a continuing survey conducted every month. The figures to be published refer to the twelve months preceding the interview month. The JSS covers more items than the VS, and new topics. Finally, the wording of the questions has been modified.

To maintain the possibility of constructing crime trends, the effects of the differences between the JSS and the VS on parameter estimates were quantified by means of an experiment. During one year (1992), both surveys were conducted concurrently and treated as regular surveys. In this experiment two separate samples were drawn for the JSS and the VS, both with a sample size equal to the size of the regular VS sample in the past. Interviewers were randomly allotted over the two treatments in order to conduct the experiment double blind. Many estimates of the crime figures based on the JSS turned out to be significantly higher than in the VS. In this experiment, only the total effect of all the changes introduced simultaneously could be quantified. Simple indexations were derived in order to keep the figures based on the VS and the JSS comparable. When the effects of the separate changes and their possible interactions have to be quantified, then a factorial design should be applied.

5. Analysis of Embedded Field Experiments

In Section 4 we pointed out how the principles of experimental design can be applied in order to design efficient embedded field experiments. Here we will discuss how statistical methods from sampling theory can be used to support the analysis of embedded experiments if interest is focused on hypothesis testing concerning population parameters.

In embedded experiments, experimental units are selected by some complex sampling design from a finite population. Statistical methods traditionally used in the analysis of experimental designs are model-dependent and typically require IID observations. The stochastic assumptions underlying these techniques, however, do not reflect the complexity which is usually exhibited by the applied sampling design and the finite survey population from which the experimental units have been drawn (Skinner et al. 1989, Chapter 1). In these cases the assumption that the observations are IID is usually not tenable. Skinner et al. (1989, Chapter 3) have shown that the application of multivariate procedures based on the assumption of IID observations in the analysis of data obtained from complex sampling designs can lead to misleading results.

Fienberg and Tanur (1987, 1988, 1989) advocated a model-based approach for the analysis of embedded field experiments. The internal validity is ensured by the application of such fundamentals as randomization and local control on sampling structures such as strata, clusters or interviewers in the design as well as in the analysis of the embedded field experiment. The external validity is achieved by incorporating certain local control variables, such as interviewers or clusters, as random components in a mixed model analysis. Fienberg and Tanur (1988) showed, using statistical likelihood theory, that the weights of the applied sampling design can be ignored in the analysis when the selection of the sampling units depends only on prior variables conditioned on in the statistical model and are independent of the target variables. If the experiment is analyzed under the assumption of IID observations, then the analysis is performed conditional on the drawn sample and inferences are made about the parameters of some hypothetical superpopulation model and not about the finite population from which the sample is drawn. By this approach, the observations are assumed to be identically and independently distributed realizations of this superpopulation model. Under the postulated model it does not matter which respondents provide the information to draw statistical inferences about the parameters of this model. Skinner et al. (1989) propose various multivariate procedures for the analysis of data obtained from complex surveys. These methods are based on model-dependent procedures which require IID observations. The asymptotic distribution of the test statistics used to test hypotheses concerning model parameters is adjusted for the design effect of the sampling design in order to incorporate the complexity of the sampling design in the analysis.

As the examples mentioned in Section 4 suggest, the purpose of the embedded field experiments can be viewed as testing or quantifying the effect of alternative treatments on estimates of finite population parameters. The disadvantage of using the model-based approach for such problems is that the inference concerns model parameters from some superpopulation and not the estimates of the parameters of the finite survey population, even if the external validity is guaranteed by the use of random or mixed models. Furthermore, the validity of the inference depends on model assumptions. It is unclear how robust such an analysis is against, e.g., bias due to selective nonresponse. To cope with these disadvantages we explore a design-based approach. In embedded field experiments a large number of experimental units are selected from a finite population by means of a random sampling design. As a result, it becomes possible to draw inferences on finite population parameters that do depend on a probability structure imposed by the design of the survey and not on model parameters from a superpopulation that depends on an assumed probability distribution. To this end, the analysis can be based on the estimates of finite population parameters. From the objective of the experiment, it is possible to formulate sensible hypotheses about these finite population parameters and to construct efficient test statistics. Statistical methods from sampling theory can be used by constructing such test statistics, which take into account that experimental units are selected from a finite population by some complex sampling design with possibly unequal inclusion probabilities and/or clustering. Furthermore, such a design-based approach makes it possible to use auxiliary information by means of the generalized regression estimator. This not only increases the precision, but it also makes the analysis of embedded experiments more robust against the negative effects of selective nonresponse. In the next section we propose a method for the analysis of the two-sample problem to illustrate the possibility of developing design-based methods for embedded experiments.

6. A Design-based Approach for the Analysis of the Two-sample Problem

Consider an embedded field experiment designed to compare the effect of an alternative survey methodology with respect to a standard survey methodology on a target parameter of a survey. The population mean of a target parameter measured by using the standard treatment is denoted by $\bar{X}$ and the population mean of the same target parameter measured by using the alternative treatment is denoted by $\bar{Y}$. The objective of the experiment is to investigate whether there is a significant difference between the parameters $\bar{X}$ and $\bar{Y}$. From this objective the following hypothesis can be derived:

$$H_0: \quad \bar{X} = \bar{Y}$$

$$H_{1a}: \quad \bar{X} \neq \bar{Y} \quad \text{or} \quad H_{1b}: \bar{X} > \bar{Y} \quad \text{or} \quad H_{1c}: \quad \bar{X} < \bar{Y} \tag{1}$$

To test this hypothesis, a sample s of size n is drawn from the target population U of size N by some complex sampling design with first order inclusion expectations π_i for sampling unit i and second order inclusion expectations π_{ij} for sampling units i and j. Regardless of the structure of the sampling design, sample s is randomly divided into two subsamples s_1 and s_2 of sizes n_1 and n_2, respectively. The subsamples are not necessarily of equal size. The experimental units assigned to subsample s_1 receive the standard treatment and the experimental units assigned to subsample s_2 the alternative treatment. The observations of the first subsample are denoted by $x_i \, (i = 1, 2, \ldots, n_1)$ and the observations of the second subsample by $y_i \; (i = 1, 2, \ldots, n_2)$. In many practical situations, s_1 is relatively large compared with s_2 because s_1 serves as the regular survey for publication purposes and the control group in the experiment.

To draw inferences about finite population parameters, the analysis should explicitly take into account the probability structure of the applied complex sampling design used to draw sample s (established by the first and second order inclusion expectations π_i and π_{ij}) as well as the randomization mechanism applied to divide sample s into two subsamples. To this end it is proposed to replace the parameters of the t-statistic by their corresponding Horvitz-Thompson estimator. The following test statistic for testing hypothesis (1) is proposed:

$$\tilde{t} = \frac{\hat{\bar{X}}_{\pi_1} - \hat{\bar{Y}}_{\pi_2}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\hat{\bar{X}}_{\pi_1} - \hat{\bar{Y}}_{\pi_2})}} \tag{2}$$

with $\hat{\bar{X}}_{\pi_1}$ the Horvitz-Thompson estimator for $\bar{X}$ based on the sampling units of subsample s_1, $\hat{\bar{Y}}_{\pi_2}$ the Horvitz-Thompaon estimator for $\bar{Y}$ based on the sampling units of s_2, and $\hat{\mathrm{V}}\mathrm{ar}(\hat{\bar{X}}_{\pi_1} - \hat{\bar{Y}}_{\pi_2})$ an estimator for the variance of $(\hat{\bar{X}}_{\pi_1} - \hat{\bar{Y}}_{\pi_2})$.

The first order inclusion expectations for the sampling units in subsamples s_1 and s_2 are

$(n_1/n)\pi_i$ and $(n_2/n)\pi_i$, respectively (Van den Brakel and Renssen 1996b). Note that π_i is the first order inclusion expectation of sample s. Because sample s is randomly divided into two subsamples the n_1/n and n_2/n are introduced in the first order inclusion expectations for the sampling units of s_1 and s_2, respectively. It follows that the Horvitz-Thompson estimator for $\bar{X}$ based on subsample s_1 is given by

$$\hat{\hat{X}}_{\pi_1} = \frac{n}{Nn_1} \sum_{i \in s_1} \frac{x_i}{\pi_i} \tag{3}$$

In an equivalent way, the Horvitz-Thompson estimator for $\bar{Y}$ based on subsample s_2 is given by

$$\hat{\hat{Y}}_{\pi_2} = \frac{n}{Nn_2} \sum_{i \in s_2} \frac{y_i}{\pi_i} \tag{4}$$

Because $\hat{\hat{X}}_{\pi_1}$ and $\hat{\hat{Y}}_{\pi_2}$ are based on two interpenetrating subsamples drawn from a finite population, they are not independent. Van den Brakel and Renssen (1996b) derived an expression for $\mathrm{Var}(\hat{\hat{X}}_{\pi_1} - \hat{\hat{Y}}_{\pi_2})$, taking into account the dependency between $\hat{\hat{X}}_{\pi_1}$ and $\hat{\hat{Y}}_{\pi_2}$. Deriving a design-unbiased estimator for $\mathrm{Var}(\hat{\hat{X}}_{\pi_1} - \hat{\hat{Y}}_{\pi_2})$ requires paired observations of x_i and y_i obtained from each sampling unit. These paired observations are not available because the sampling units are assigned to either s_1 or s_2. However, it can be derived that an approximately unbiased estimator for $\mathrm{Var}(\hat{\hat{X}}_{\pi_1} - \hat{\hat{Y}}_{\pi_2})$ is given by (see Van den Brakel and Renssen (1996b) for a derivation):

$$\begin{aligned}\hat{\mathrm{V}}\mathrm{ar}(\hat{\hat{X}}_{\pi_1} - \hat{\hat{Y}}_{\pi_2}) = {} & \frac{1}{n_1} \frac{1}{(n_1 - 1)} \sum_{i \in s_1} \left(\frac{nx_i}{N\pi_i} - \frac{1}{n_1} \sum_{i \in s_1} \frac{nx_i}{N\pi_i} \right)^2 \\ & + \frac{1}{n_2} \frac{1}{(n_2 - 1)} \sum_{i \in s_2} \left(\frac{ny_i}{N\pi_i} - \frac{1}{n_2} \sum_{i \in s_2} \frac{ny_i}{N\pi_i} \right)^2 \equiv \frac{1}{n_1} \hat{S}_X^2 + \frac{1}{n_2} \hat{S}_Y^2\end{aligned} \tag{5}$$

Note that $\hat{S}_X^2/n_1$ and $\hat{S}_Y^2/n_2$ are ordinary variance estimators for the sample means as if the sample elements are selected with unequal probabilities (π_i/n) with replacement (Cochran 1977, Equation (9A.16)). These variance estimators only depend on the first order inclusion expectations. No second order inclusion expectations are required. Consequently, test statistic (2) is relatively simple to evaluate. The efficiency of the sampling design tends to vanish by the comparison of the subsample means $\hat{\hat{X}}_{\pi_1}$ and $\hat{\hat{Y}}_{\pi_2}$. This result seems to be in conformity with the results of Kish and Frankel (1974). They found that the design effect for differences between subclass means tends towards one from below for proportionate stratified sampling. Also for cluster samples they empirically found that the design effect of a positive intraclass correlation for differences between subclass means is less than for separate means.

If all first order inclusion expectations are equal (self-weighted sampling designs), the test statistic (2) reduces to Welch's t' test statistic (Miller 1986), regardless of the second order inclusion expectations of the sampling design used to draw s. Note that $\hat{S}_X^2$ and $\hat{S}_Y^2$ are estimates for the population variances for the x and y variables weighted with a factor $n/(N\pi_i)$. If it is reasonable to assume that these weighted population variances are equal, then an efficient estimate is obtained by using the pooled variance estimator:

$$\hat{S}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i \in s_1}\left(\frac{nx_i}{N\pi_i} - \frac{1}{n_1}\sum_{i \in s_1}\frac{nx_i}{N\pi_i}\right)^2 + \sum_{i \in s_2}\left(\frac{ny_i}{N\pi_i} - \frac{1}{n_2}\sum_{i \in s_2}\frac{ny_i}{N\pi_i}\right)^2\right) \tag{6}$$

In the case of a self-weighted sampling design, the test statistic (2) reduces to the *t*-test statistic.

In the analysis of the experiment, we can take advantage of auxiliary information by applying the generalized regression estimator instead of the Horvitz-Thompson estimator for the estimation of the parameters in the *t*-statistic. This increases the precision of the analysis and corrects, at least partially, for the bias due to selective nonresponse. Note that this approach very much resembles the application of covariance analysis from the theory of experimental designs. By analogy with (2), the following test statistic is obtained:

$$\hat{t} = \frac{\hat{\bar{X}}_{R_1} - \hat{\bar{Y}}_{R_2}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\hat{\bar{X}}_{R_1} - \hat{\bar{Y}}_{R_2})}} \tag{7}$$

with $\hat{\bar{X}}_{R_1}$ the generalized regression estimator for $\bar{X}$ based on s_1 and $\hat{\bar{Y}}_{R_2}$ the generalized regression estimator for $\bar{Y}$ based on s_2. Following the model assisted approach of Särndal et al. (1992) the target variables for each element in the population are to a certain extent assumed to be an independent realization of a linear regression model. In order to describe the target variables measured by means of the standard treatment as well as the experimental treatment, two different regression models are defined:

$$\begin{aligned} x_i &= \mathbf{z}_i^t \mathbf{b}_x + e_{x_i} \quad V(x_i) = \sigma_{x_i}^2 \\ y_i &= \mathbf{z}_i^t \mathbf{b}_y + e_{y_i} \quad V(y_i) = \sigma_{y_i}^2, \quad i = 1, 2, \ldots, N \end{aligned} \tag{8}$$

with $\mathbf{z}_i$ a vector with q auxiliary variables of element i, $\mathbf{b}_x$ and $\mathbf{b}_y$ vectors containing q regression coefficients, e_{x_i} and e_{y_i} the residuals, and $\sigma_{x_i}^2$ and $\sigma_{y_i}^2$ the variances of the regression models of the target variables of x_i and y_i respectively.

The generalized regression estimator for $\bar{X}$ based on subsample s_1 is given by:

$$\hat{\bar{X}}_{R_1} = \hat{\bar{X}}_{\pi_1} + \hat{\mathbf{b}}_x^t(\bar{\mathbf{Z}} - \hat{\bar{\mathbf{Z}}}_{\pi_1}) \tag{9}$$

with $\hat{\mathbf{b}}_x$ the generalized regression estimator of the regression coefficient $\mathbf{b}_x$ based on the sampling units in subsample s_1 (see Särndal et al. (1992), Equation 6.4.13), $\bar{\mathbf{Z}}$ a vector with the q population means of the auxiliary variables, and $\hat{\bar{\mathbf{Z}}}_{\pi_1}$ a vector with Horvitz-Thompson estimators of the population means of the q auxiliary variables based on subsample s_1 (with first order inclusion expectation $\frac{n_1}{n}\pi_i$). In a similar way, the generalized regression estimator for $\bar{Y}$ based on the sampling units in subsample s_2 is given by:

$$\hat{\bar{Y}}_{R_2} = \hat{\bar{Y}}_{\pi_2} + \hat{\mathbf{b}}_y^t(\bar{\mathbf{Z}} - \hat{\bar{\mathbf{Z}}}_{\pi_2})$$

with $\hat{\mathbf{b}}_y$ the generalized regression estimator of the regression coefficient $\mathbf{b}_y$ based on s_2 and $\hat{\bar{\mathbf{Z}}}_{\pi_2}$ a vector with Horvitz-Thompson estimators of the population means of the q auxiliary variables based on subsample s_2 (with first order expectation $\frac{n_2}{n}\pi_i$).

An approximately unbiased estimator for $\mathrm{Var}(\hat{\hat{X}}_{R_1} - \hat{\hat{Y}}_{R_2})$ is given by:

$$\hat{\mathrm{V}}\mathrm{ar}(\hat{\hat{X}}_{R_1} - \hat{\hat{Y}}_{R_2}) = \frac{1}{n_1}\frac{1}{(n_1-1)}\sum_{i\in s_1}\left(\frac{n\hat{e}_{x_i}}{N\pi_i} - \frac{1}{n_1}\sum_{i\in s_1}\frac{n\hat{e}_{x_i}}{N\pi_i}\right)^2 + \frac{1}{n_2}\frac{1}{(n_2-1)}\sum_{i\in s_2}\left(\frac{n\hat{e}_{y_i}}{N\pi_i} - \frac{1}{n_2}\sum_{i\in s_2}\frac{n\hat{e}_{y_i}}{N\pi_i}\right)^2 \equiv \frac{1}{n_1}\hat{S}^2_{E_x} + \frac{1}{n_2}\hat{S}^2_{E_Y} \qquad (11)$$

with $\hat{e}_{x_i} = x_i - \mathbf{z}_i^t\hat{\mathbf{b}}_x$ and $\hat{e}_{y_i} = y_i - \mathbf{z}_i^t\hat{\mathbf{b}}_y$. The derivation of this variance estimator resembles the derivation of the variance estimator in the case of the Horvitz-Thompson estimator (5). Following Särndal et al. (1992, Result 6.6.1), the g weights (Särndal et al. (1992), Equation (6.5.10)) can be attached to the residuals in the variance estimators, as an alternative. If it is reasonable to assume that the weighted population variances for both treatments are equal (under the null hypothesis), then it is more efficient to use the pooled variance estimator. This estimator has the same form as (6) with x_i and y_i replaced by $\hat{e}_{x_i}$ and $\hat{e}_{y_i}$ respectively. Instead of defining two separate regression models for both treatments in the experiment, it is possible to assume that the regression coefficients of the auxiliary variables in both treatments are equal ($\mathbf{b}_x = \mathbf{b}_y = \mathbf{b}$). Then the target variables in the population can be described with one linear regression model. Consequently, the estimates of the regression coefficients $\hat{\mathbf{b}}$, based on sample s (with first order inclusion expectation π_i), will be more accurate. Vector $\hat{\mathbf{b}}$ can be substituted, in the generalized regression estimators (9) and (10).

Hypothesis (1) can be tested with the test statistics (2) or (7). In order to construct critical regions, we have to know the probability distribution of the test statistics. In the case of simple random sampling without replacement, Lehmann (1975, Appendix 8), based on the work of Hájek (1960), gives a sufficient condition under which the joint distribution of the two-sample means tends to the bivariate normal distribution. Consequently, in the case of simple random sampling central limit theorems can be applied to derive that the limit distribution of the test statistics (2) and (7) tends to the standard normal distribution. In survey literature the normality assumption for estimators based on complex sampling designs is usually assumed to be valid. The assumption that the test statistics (2) and (7) are asymptomatically standard normally distributed has been confirmed by simulation studies for different sampling designs. Consequently, the standard normal distribution can be used to construct critical regions which yield very nearly $(1-\alpha)\%$ coverage, where α denotes the size of the test.

7. Discussion, Conclusions and Further Research

Field experiments embedded in ongoing surveys are particularly appropriate if interest is focussed on testing of hypotheses concerning the effect of alternative survey methodologies or treatments on estimates of finite population parameters. Statistical methods from experimental designs and sampling theory can be combined in order to develop efficient methods for design and analysis of such experiments. Principles of experimentation should be applied in the design and analysis of embedded experiments to improve the precision of the estimated treatment effects and to avoid distorting the cause-effect relationship between treatments and outcomes. Trying to implement experiments embedded in ongoing surveys, where the regular survey is also used as the control group (see the examples in Section 4), involves the danger that the regular survey will take priority over the alternative treatments by the conduction of the fieldwork. Therefore it may be efficient to conduct embedded experiments completely separate from the field work of regular surveys. Nevertheless, in large-scale field experiments it remains very difficult to standardize the application of treatments and to eliminate or exercise sufficient control over external influences. Usually many people (e.g., interviewers) are involved in conducting a field experiment, which makes it difficult to standardize protocols of experimentation and to supervise compliance. So there are many sources of extraneous variation that can mask or bias the results of the experiment and distort the cause-effect relationship between treatments and observed effects. The principles of experimental designs can be applied in designing embedded field experiments in order to minimize the negative effects of these disturbances. Parallels between structures of experimental designs and sampling theory can be exploited in a straightforward manner by designing efficient embedded experiments based on these principles. The structure of the survey design forms a framework for the design of the experiment, e.g., local control by means of randomization within strata, clusters or interviewers.

In this article we advocate a design-based analysis in order to draw inferences about finite population parameters of the survey. In embedded field experiments, a large number of experimental units are drawn by means of a random sample from a finite population and are, according to the experimental design, randomized to different treatments. Sensible hypotheses about finite population parameters can be formulated from the objective of the experiment. Efficient test statistics can be constructed which take into account the probability structure imposed by the applied sampling design as well as the randomization applied to assign experimental units to the different treatments. By using a design-unbiased estimator, like the Horvitz-Thompson estimator, or the generalized regression estimator for the parameters of the test statistic, the analysis takes into account the complexity of the applied sampling design so inferences on finite population parameters can be drawn. In doing so the external validity of the experiment is guaranteed. Besides the sampling design, the estimator for the test statistic must also take into account the randomization of the experimental design applied to assign the experimental units to the different treatments, which consequently guarantees the internal validity of the experiment.

Such statistical procedures are currently not generally available. A first step towards the development of such methods is given in this article by deriving a design-based method for the analysis of the two-treatment embedded experiment. In Section 6 we showed that combining the applied sampling design with the randomization according to the experimental design comes down to a reweighting of the observations using a factor $n/(N\pi_i)$. As a result, we were able to obtain a test statistic that is relatively simple to evaluate.

Van den Brakel and Renssen (1996c) generalized these results to the analysis of the k sample problem and obtained a design-based method for the analysis of the completely randomized design embedded in complex sampling designs. The results obtained so far must be generalized to experimental designs which exercise local control over sampling structures by e.g., randomization within strata, interviewers or clusters. This will lead naturally to statistical procedures for the design and analysis of embedded experiments, combining the internal validity guaranteed by methods from randomized experimentation with the external validity obtained from the theory of randomized sampling.

8. References

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias Through Regression Estimation. Journal of Official Statistics, 4, 251–260.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (eds.) (1991). Measurement Errors in Surveys. Wiley, New York.

Cochran, W.G. (1977). Sampling Techniques. Wiley, New York.

Cochran, W.G. and Cox, G.M. (1957). Experimental Designs. Wiley, New York.

Fienberg, S.E. and Tanur, J.M. (1987). Experimental and Sampling Structures: Parallels Diverging and Meeting. International Statistical Review, 55, 75–96.

Fienberg, S.E. and Tanur, J.M. (1988). From the Inside Out and the Outside In: Combining Experimental and Sampling Structures. The Canadian Journal of Statistics, 16, 135–151.

Fienberg, S.E. and Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. Science, 243, 1017–1022.

Fienberg, S.E. and Tanur, J.M. (1996). Reconsidering the Fundamental Contributions of Fisher and Neyman on Experimentation and Sampling. International Statistical Review, 64, 237–253.

Fisher, R.A. (1935). The Design of Experiments. Oliver and Boyd, Edinburgh.

Forsman, G. (1989). Early Survey Models and Their Use in Survey Quality Work. Journal of Official Statistics, 5, 41–55.

Hájek, J. (1960). Limiting Distributions in Simple Random Sampling from a Finite Population. Mathematical Institute of the Hungarian Academy of Sciences, 5, 361–374.

Hansen, M.H. and Hurwitz, W.N. (1943). On the Theory of Sampling from Finite Populations. Annals of Mathematical Statistics, 14, 333–362.

Hinkelmann, K. and Kempthorne, O. (1994). Design and Analysis of Experiments. Wiley, New York.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Huys, H. and Roodijn, J. (1994). A New Survey on Justice and Security. Netherlands Official Statistics, 9, 47–51.

Kempthorne O. (1952). The Design and Analysis of Experiments. Wiley, New York.

Kish, L. and Frankel, M.R. (1974). Inference from Complex Samples. Journal of the Royal Statistical Society, Series B, 36, 1–37.

Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks. McGraw-Hill, New York.

Miller, R.G. (1986). Beyond ANOVA, Basics of Applied Statistics. Wiley, New York.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society, Vol. 97, 558–625.

O'Muircheartaigh, C. (1997). Casm: Successes, [Failures,] and Potential. Methodology Institute, London School of Economics and Political Science.

Särndal, C.E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Application to Two-phase Sampling and Nonresponse. International Statistical Review, 55, 279–294.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag, New York.

Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). Analysis of Complex Surveys. Wiley, New York.

Van den Brakel, J.A. and Renssen, R.H. (1995). Application of Experimental Designs at Statistical Bureaus. Research paper (BPA no.: 5443-95-RSM), Department of Statistical Methods, Statistics Netherlands.

Van den Brakel, J.A., Luppes, M., and Moritz, G. (1995). Testing Effects of Informed Consent in the National Travel Survey. Netherlands Official Statistics, 10, 37–43.

Van den Brakel, J.A. and Renssen, R.H. (1996a). Application of Experimental Designs in Survey Methodology. Proceedings of the International Conference on Survey Measurement and Process Quality, American Statistical Association, 151–156.

Van den Brakel, J.A. and Renssen, R.H. (1996b). The Analysis of the Two-Sample Problem Embedded in Complex Sampling Designs. Research paper (BPA no.: 1027-96-RSM), Department of Statistical Methods, Statistics Netherlands.

Van den Brakel, J.A. and Renssen, R.H. (1996c). The Analysis of Completely Randomized Designs Embedded in Complex Sampling Designs. Research paper (BPA no.: 8090-96-RSM), Department of Statistical Methods, Statistics Netherlands.

Received August 1995
Revised August 1997