Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 217–231

A Design-based Analysis Procedure for Two-treatment Experiments Embedded in Sample Surveys. An Application in the Dutch Labor Force Survey

Jan A. van den Brakel and C.A.M. van Berkel[1]

Recently in the Dutch Labor Force Survey (LFS), the questionnaire had to be changed. Before the new questionnaire was implemented as standard, its effects on the outcomes of the LFS as compared with the effects of the regular one were investigated by means of an experiment embedded in the regular LFS. To test hypotheses about possible effects on estimates of finite population parameters estimated by the sample survey a two-treatment randomized block design was analyzed using a design-based analysis procedure. The results from this analysis are compared with the results obtained with a model-based analysis.

Key words: Completely randomized designs; complex sampling designs; embedded field experiments; questionnaire effects; randomized block designs.

1. Introduction

At official statistical bureaus sample surveys are conducted in order to publish various estimates of finite population characteristics. These sample survey processes contain many sources of non-sampling errors, which affect the accuracy of the sample survey estimates. Part of the research aimed at the improvement of the quality and efficiency of sample survey processes, is to consider and test alternative survey methodologies. Experiments embedded in ongoing sample surveys are particularly appropriate to test effects of alternative survey methodologies on estimates of finite population parameters (see Fienberg and Tanur 1987, 1988, 1989; Van den Brakel and Renssen 1998). In statistical bureaus sample surveys, like the Dutch Labor Force Survey (LFS), are generally kept unchanged as long as possible in order to construct uninterrupted time series of the estimated population parameters. It remains inevitable, however, that survey processes be adjusted from time to time. When they are, embedded experiments can be applied to detect and quantify possible trend disruptions in time series of estimated population parameters due to adjustments of a survey process before they are implemented as standard.

The typical situation, considered in this article, is an experiment designed to compare the effect of an alternative survey approach on the main estimates of the finite population

[1] Department of Statistical Methods, Statistics Netherlands, 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: The authors wish to thank the referees, in particular Dr. F. Scheuren, for commenting on an earlier version. Jan also thanks Prof. Stephen E. Fienberg (Department of Statistics, Carnegie Mellon University, Pittsburgh, U.S.A.) for his unreserved support during the PhD research that was a partial basis for this article.

© Statistics Sweden

parameters of an ongoing survey. To this end, a sample drawn from a finite population, is randomly divided into two subsamples according to some experimental design. In many practical situations there is one large subsample that is assigned to the regular survey, and serves besides the official publication purposes also as the control group in the experiment. The other subsample, which is generally smaller, is assigned to the alternative approach.

Generally, the purpose of such experiments is the estimation of finite population parameters obtained under the regular and the alternative survey implementation and to test hypotheses about the differences between these estimated population parameters. The application of model-based analysis procedures that do not allow for the sampling design and the weighting procedure of the survey might result in design-biased parameter and variance estimates. Moreover, in a model-based analysis the estimated treatment effects concern the parameters of a linear regression model, which do not necessarily coincide with the target population parameters of the sample survey. Therefore the analysis results obtained in a model-based procedure might be incommensurable with the parameter and variance estimates of the ongoing survey. Van den Brakel and Renssen (1998) derived a design-based procedure for the analysis of embedded two-treatment experiments designed as a completely randomized design, that does not have these limitations.

A design-based approach for the analysis of embedded two-treatment experiments is proposed in Section 2, extending the analysis done by Van den Brakel and Renssen (1998) of embedded two-treatment randomized block designs where sampling structures like primary sampling units, clusters, strata, and interviewers and the like are used as block variables. This method is applied to the analysis of an experiment embedded in the LFS aimed to test the effects of a new questionnaire. The survey design of the LFS, the changes in its questionnaire and the experimental design used to test the effects of the new questionnaire are described in Section 3. In Section 4 the experiment is analyzed with the design-based methods derived in Section 2 and compared with standard model-based methods.

2. Methods

2.1. *Hypothesis testing*

Consider an experiment embedded in an ongoing sample survey, designed to compare the effect of an alternative survey methodology to a standard survey methodology in respect of the estimates of the main target parameters. Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the finite population means observed by means of the alternative and the regular survey approach, respectively. With regard to the purpose of the experiment the following hypotheses are of interest

$$H_0 : \bar{Y}_1 = \bar{Y}_2$$
$$H_{1a} : \bar{Y}_1 \neq \bar{Y}_2 \text{ or } H_{1b} : \bar{Y}_1 > \bar{Y}_2 \text{ or } H_{1c} : \bar{Y}_1 < \bar{Y}_2 \quad (1)$$

To test these hypotheses, a sample s is drawn from the target population of size N by means of a generally complex sampling design. According to the experimental design, the sample s is randomly divided into two subsamples. Let s_1 denote the subsample assigned to the new survey approach and s_2 the subsample assigned to the regular survey approach. In

the case of a completely randomized design (CRD), the sampling units are randomly divided over the two treatments, regardless of the structure of the sampling design used to draw s. In the case of a randomized block design (RBD), the sampling units are deterministically divided into homogeneous blocks. The sampling units within each block are randomized over the two treatments. In the case of an RBD the variance between the blocks can be eliminated from the variance of the estimated treatment effects, which might increase the accuracy of an experiment considerably. Therefore Fienberg and Tanur (1987, 1988, 1989) and Van den Brakel and Renssen (1998) advocated using sampling structures such as strata, clusters, primary sampling units, and interviewers and the like as blocking variables in an RBD if experiments are embedded in ongoing sample surveys.

The way this works, based on the observations obtained in the experiment, two estimates are obtained for each population parameter of the sample survey. One estimate is based on the subsample where data were collected with the regular survey approach, the other on the subsample where data were collected with the new survey approach. To test Hypotheses (1) a design-based t-statistic is proposed in Section 2.3. To this end a generalized regression estimator or a Horvitz-Thompson estimator for the population parameters and for the variance of the differences between these two estimators are derived. These estimators are design-unbiased for the finite population parameters, since they are derived under the joint probability structure of the sampling design and the experimental design. As a result, an analysis procedure that draws inference on the finite population parameters specified in Hypotheses (1) is obtained.

2.2. *Measurement error models*

Design-based sampling theory is largely based on the traditional notion (e.g., Cochran 1977) that observations obtained from sampling units are true fixed values observed without error. This approach, however, is not tenable if systematic differences can arise between estimates of finite population parameters due to different survey implementations or non-sampling errors. A measurement error model for the observations obtained from the sampling units has to be introduced for such settings. This measurement error model conveniently fits in a design-based theory for the analysis of embedded experiments and, as we will see, enables us to relate systematic differences between estimates of finite population parameters to different survey implementations or treatment effects.

It is assumed that the observations obtained from the sampling units are a realization of the following measurement error model:

$$y_{jkl} = u_j + \beta_k + \gamma_l + \epsilon_{jk} \tag{2}$$

where y_{jkl} is the observation obtained from sampling unit j assigned to treatment k and interviewer l, u_j is the true, intrinsic value of sampling unit j, β_k an additive effect of treatment k, γ_l an effect of interviewer l and ϵ_{jk} an error component. This model allows for mixed interviewer effects, i.e., $\gamma_l = \psi_l + \xi_l$, where ψ_l and ξ_l denote the fixed and random effect of interviewer l, respectively. It is assumed that $\mathrm{E}(\xi_l) = 0$ and that the random interviewer effects between interviewers are independent. Furthermore, it is assumed that $\mathrm{E}(\epsilon_{jk}) = 0$ and that measurement errors between different sampling units are independent.

Hence,

$$\mathrm{E}(y_{jkl}) = u_j + \beta_k + \psi_l$$

$$\mathrm{Cov}(y_{jkl}, y_{j'k'l'}) = \begin{cases} \mathrm{Var}(\epsilon_{jk}) + \mathrm{Var}(\xi_l) & : \quad j = j',\ l = l' \\ \mathrm{Var}(\xi_l) & : \quad j \neq j',\ l = l' \\ 0 & : \quad j \neq j',\ l \neq l' \end{cases}$$

Obviously, any correlation between the responses of different individuals assigned to the same interviewer can be modeled by means of random interviewer effects. Any fixed interviewer effect influences the bias of the response values. Under Measurement error model (2), the application of an RBD with interviewers as block variables will be efficient since this eliminates the variance between interviewers from the estimated treatment effects. If, however, under Model (2) interviewers are not used as block variables, it is still possible to obtain design unbiased estimates for the treatment effects as long as interviewers are randomly assigned to the different treatments (see Van den Brakel 2001).

For reasons mentioned in the introduction, Hypotheses (1) are tested by estimating $\bar{Y}_k$, taking into account the sampling design and the weighting procedure of the sample survey as well as the experimental design. To allow for the weighting procedure of the regular sample survey, the analysis must be based on the generalized regression estimator. The use of auxiliary information by means of this estimator has the advantages that it might reduce the design variance of the parameter estimates and that it corrects, at least partially, for the design bias due to selective nonresponse (see Bethlehem 1988; Bethlehem and Keller 1987; Särndal et al. 1992). Let $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jH})^t$ denote a vector of order H with each element x_{jh} an auxiliary variable of sampling unit j. According to the model-assisted approach of Särndal et al. (1992), a linear regression model for the intrinsic values of the sampling units is defined by

$$u_j = \mathbf{b}^t \mathbf{x}_j + e_j \tag{3}$$

where $\mathbf{b}$ denotes a vector of order H with regression coefficients, and e_j the residuals of the regression model. It is assumed that $\mathrm{E}(e_j) = 0$ and that the residuals between sampling units are independent. For the remainder, $\mathrm{Var}(e_j)$ is denoted by ω_j^2.

2.3. *Parameter and variance estimation*

The analysis of an RBD is considered first. Let B denote the number of blocks in the experiment and n_{bk} the number of sampling units in block b assigned to treatment k. Let $n_{+k} = \Sigma_{b=1}^{B} n_{bk}$ denote the number of sampling units in subsample s_k. Then $n_{++} = n_{+1} + n_{+2}$ denotes the number of sampling units in the entire sample s. Further $n_{b+} = n_{b1} + n_{b2}$ is the number of sampling units in block b. Both subsamples can be considered as the realization of a two-phase sample where the first phase corresponds with the sampling design used to draw s and the second phase with the experimental design used to divide s into two subsamples s_1 and s_2. As a result the first order inclusion probabilities with respect to the subsamples are given by $\pi_j^* = (n_{bk}/n_{b+})\ \pi_j$ where π_j denotes the first order inclusion probability for the jth sampling unit, used to draw s.

The t-statistic for Hypotheses (1) based on the generalized regression estimator is given

by

$$t = \frac{\hat{\bar{Y}}_{1R} - \hat{\bar{Y}}_{2R}}{\sqrt{\widehat{\mathrm{Var}}(\hat{\bar{Y}}_{1R} - \hat{\bar{Y}}_{2R})}} \tag{4}$$

The generalized regression estimators for the population means in (4) are defined by

$$\hat{\bar{Y}}_{k_R} = \hat{\bar{Y}}_k + \hat{\mathbf{b}}_k^t(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}_k),\ k = 1, 2$$

Here

$$\hat{\bar{Y}}_k = \frac{1}{N}\sum_{b=1}^{B}\sum_{j=1}^{n_{bk}} \frac{n_{b+}}{n_{bk}}\frac{y_{jk}}{\pi_j} \equiv \frac{1}{N}\sum_{j=1}^{n_{+k}} \frac{y_{jk}}{\pi_j^*}$$

denotes the Horvitz-Thompson estimator for the population means $\bar{Y}_k$. Further, $\bar{\mathbf{X}}$ denotes the vector of order H containing the known population means of the auxiliary variables. The Horvitz-Thompson estimator for $\bar{\mathbf{X}}$, based on the n_{+k} sampling units in subsample s_k, is given by

$$\hat{\bar{\mathbf{X}}}_k = \frac{1}{N}\sum_{b=1}^{B}\sum_{j=1}^{n_{bk}} \frac{n_{b+}}{n_{bk}}\frac{\mathbf{x}_j}{\pi_j} \equiv \frac{1}{N}\sum_{j=1}^{n_{+k}} \frac{\mathbf{x}_j}{\pi_j^*}$$

An estimator for the regression coefficients $\mathbf{b}$ based on the n_{+k} sampling units in subsample s_k is given by

$$\hat{\mathbf{b}}_k = \left(\sum_{b=1}^{B}\sum_{j=1}^{n_{bk}} \frac{n_{b+}}{n_{bk}}\frac{\mathbf{x}_j\mathbf{x}_j^t}{\omega_j^2\pi_j}\right)^{-1} \sum_{b=1}^{B}\sum_{j=1}^{n_{bk}} \frac{n_{b+}}{n_{bk}}\frac{\mathbf{x}_j y_{jk}}{\omega_j^2\pi_j}$$

The denominator of (4) concerns the squared root of the variance estimator for the difference between the two generalized regression estimators. In Van den Brakel (2001) it is proved that

$$\widehat{\mathrm{Var}}(\hat{\bar{Y}}_{1R} - \hat{\bar{Y}}_{2R}) = \sum_{k=1}^{2}\sum_{b=1}^{B}\frac{1}{n_{bk}}\frac{1}{(n_{bk}-1)}\sum_{j=1}^{n_{bk}}\left(\frac{n_{b+}\hat{e}_{jk}}{N\pi_j} - \frac{1}{n_{bk}}\sum_{j=1}^{n_{bk}}\frac{n_{b+}\hat{e}_{jk}}{N\pi_j}\right)^2 \tag{5}$$

where $\hat{e}_{jk} = y_{jk} - \hat{\mathbf{b}}_k^t\mathbf{x}_j$ are the estimated residuals of the Regression model (3). This variance estimator is design unbiased under generally complex sampling designs, despite the fact that no second order inclusion probabilities are involved in (5). Under the assumption that $\mathrm{Var}(\epsilon_{j1}) = \mathrm{Var}(\epsilon_{j2})$ in Model (2), Van den Brakel (2001) proposed the following variance estimator, where the two treatment groups within each block are pooled:

$$\widehat{\mathrm{Var}}(\hat{\bar{Y}}_{1R} - \hat{\bar{Y}}_{2R}) = \sum_{b=1}^{B}\left(\frac{1}{n_{b1}} + \frac{1}{n_{b2}}\right)\frac{1}{(n_{b1}+n_{b2}-2)}\sum_{k=1}^{2}\sum_{j=1}^{n_{bk}}\left(\frac{n_{b+}\hat{e}_{jk}}{N\pi_j} - \frac{1}{n_{bk}}\sum_{j=1}^{n_{bk}}\frac{n_{b+}\hat{e}_{jk}}{N\pi_j}\right)^2 \tag{6}$$

An expression for the t-statistic analyzed with the Horvitz-Thompson estimator follows as a special case from the results obtained for the generalized regression estimator. The minimum use of auxiliary information is a weighting scheme where $\mathbf{x}_j = (1)$ for all elements in the population. Furthermore it is assumed that $\omega_j^2 = \omega^2$. This weighting scheme

corresponds with the common mean model (Särndal et al. 1992, Section 7.4). Under this model it follows that

$$\hat{\bar{Y}}_{k_R} = \left(\sum_{b=1}^{B} \sum_{j=1}^{n_{bk}} \frac{n_{b+}}{n_{bk}\pi_j} \right)^{-1} \left(\sum_{b=1}^{B} \sum_{j=1}^{n_{bk}} \frac{n_{b+}\, y_{jk}}{n_{bk}\pi_j} \right) \equiv \tilde{y}_{s_k} \tag{7}$$

which can be recognized as the ratio estimator for a population mean, originally proposed by Hájek (1971). It also follows that $\hat{\mathbf{b}}_k = (\tilde{y}_{s_k})$ and that an approximately design unbiased estimator for the variance of the treatment effects is given by (5) or (6) with $\hat{e}_{jk} = y_{jk} - \tilde{y}_{s_k}$.

An embedded two-treatment experiment designed as a CRD can be considered as an RBD with one block. Therefore a design-based t-statistic for the embedded two-treatment experiment designed as a CRD follows as a special case from the results obtained for an RBD by taking $B = 1$, $n_{b+} = n_{++}$ and $n_{bk} = n_{+k}$. Van den Brakel and Renssen (1998) show under which situations the traditional t-statistic and Welch's t-statistic follow as a special case from the design-based t-statistic for embedded two-treatment experiments designed as a CRD.

It is assumed that a finite population central limit theorem holds so that the two estimated population parameters have a bivariate normal distribution. For more details, see Van den Brakel (2001). Consequently, under the null-hypothesis of no treatment effects the t-statistic is standard-normally distributed.

3. An Experiment in the Dutch Labor Force Survey

3.1. Design of the Dutch Labor Force Survey

The objective of the Dutch Labor Force Survey (LFS) is to provide reliable information about the labor market. The LFS has been carried out as a continuing survey since 1987. Each month a sample of addresses is selected from which during the data collection households are identified that can be regarded as the ultimate sampling units. The target population of the LFS consists of the non-institutionalised population aged 15 years and over residing in the Netherlands. The sampling frame is derived from a register of all known addresses in the Netherlands. The LFS is based on a stratified two-stage cluster design of addresses. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. In the first stage a sample of municipalities is drawn with first order inclusion probabilities proportional to the number of addresses. At the second stage a sample of at least twelve addresses is drawn without replacement from each selected municipality. Principally, all households residing at an address, up to a maximum of three, are included in the sample. Until 1999, the sample size averaged about 10,000 addresses monthly.

Since the LFS has to provide accurate outcomes on unemployment, addresses that occur in the register of the Employment Exchange are oversampled. Since most target parameters of the LFS concern people aged 15 through 64 years, addresses with only persons aged 65 years and over are undersampled. Due to reduced capacity of the interview staff during the holiday season, the sample size is halved in July and August.

Data are collected by means of computer assisted personal interviewing using hand-held computers. Interviewers are working on the data collection of the LFS in areas around their

places of residence. For all members of the selected households, demographic variables are observed. For the target variables only persons aged 15 years and over are interviewed. When a household member cannot be contacted, proxy interviewing is allowed with members of the same household. Households in which one or more of the selected individuals does not respond in person or in a proxy interview are treated as nonresponding households.

The weighting procedure of the LFS is based on the generalized regression estimator. The inclusion probabilities reflect the over- and under-sampling of addresses described above as well as the different response rates between geographical regions. The weighting scheme is based on a combination of different social-demographical categorical variables. The integrated method for weighting individuals and families of Lemaître and Dufour (1987) is applied to obtain equal weights for individuals belonging to the same household. Finally a bounding algorithm is applied to avoid negative weights. A detailed description of the methodology of the LFS is given by Hilbink, Van Berkel, and Van den Brakel (2000).

3.2. Changing the LFS and its questionnaire

From 1987 through 1999, the LFS mainly produced annual figures on the labour market. During the last couple of years, the demand for information on short-term trends has strongly increased. This is demonstrated by a Eurostat regulation on labor force surveys, according to which member states have to supply quarterly data. To provide accurate quarterly figures with a cross-sectional design, the sample would have to be enlarged extensively. This implies an undesirable increase of costs. With a rotating panel design reliable quarterly figures can be obtained without running into extra costs (see Van Berkel and Van der Valk 1999). Therefore, from October 1999 the LFS changed into a rotating panel survey. Every month a sample of some eight thousand households are visited for a face to face interview. The respondents are reinterviewed four times by telephone at quarterly intervals.

The redesign of the LFS was an opportune occasion to reconstruct the questionnaire. First all questions were grouped by subject, resulting in a modular questionnaire with a different order of questions. Particularly questions about social position were put at the end of the questionnaire instead of at the beginning and questions about wanting a job, wanting more or fewer working hours, searching activities and availability to start a new job were brought together in one module. Next, questions about benefits were skipped because of available information in registrations. Furthermore, some wordings were adapted, yielding smooth interviews. Finally, the new questionnaire was developed in a Windows version of Blaise 4, while the old questionnaire was programmed in a DOS version of Blaise 2. The revision of the questionnaire should preferably not result in a systematic difference in the main LFS figures.

3.3. Experimental design

To investigate the effects of the new questionnaire on the main outcomes of the LFS, a large-scale field experiment was conducted from April through September 1999. The experiment was performed as a two-treatment embedded RBD. Ninety experimental areas

or blocks were selected. Each block consisted of the union of two neighboring interview areas of two interviewers. The two interviewers in each block were randomly assigned to the regular or the new questionnaire. The addresses in the monthly sample of the LFS in each block were randomly divided into two subsamples of equal size, one for each interviewer.

Blocks were selected as follows. From the available interviewers with at least one year of experience with the LFS, a list of couples was formed. The union of the interview areas of each couple formed a block. From these blocks a stratified sample of 90 blocks was drawn, using geographical regions as a stratification variable. Since the selected blocks were equally divided across the Netherlands by region, and since a random sample of addresses was drawn in each block for both treatments, the results achieved in the experiment can be generalized to the entire target LFS eligible population. During the experiment, approximately 15 per cent of the LFS sample was assigned to the new questionnaire each month.

Under a measurement error model with interviewer effects it is efficient to use interviewers as block variables, since this eliminates the interviewer variance from the variance of the estimated treatment effects. Using interviewers as block variables would imply that each interviewer had to conduct the LFS with both the new and the regular questionnaire. This has the important drawback that interviewers could confuse the questionnaires during the conduction of the fieldwork, which would disturb the experiment. Confusion of the different treatments by the interviewers can be avoided, however, by a more intensive interviewer training preceding the experiment. This also requires special attention to convince the field staff of the advantages of a design where each interviewer conducts both treatments. In this experiment, however, it was decided not to use interviewers as block variables since it was not possible to run both questionnaires on one hand-held computer. Consequently, interviewers would be forced to visit addresses with two different hand-held computers, which was considered an undesirable increase of the interviewer's workload.

One month of pretesting, during March, preceded the experiment to preclude distortion of the experiment due to initial problems with the new computer system that supports the Windows version of Blaise 4 on the new hand-held computers, and in the way the interviewers dealt with the new questionnaire. Many unexpected problems with the new software were solved during this month of pretesting.

4. Results

4.1. Response

In the sample of the LFS, 16,647 addresses were selected for this experiment. From this sample, 8,262 addresses were visited with the regular questionnaire and 8,385 addresses with the new questionnaire. Table 1 contains the response account of the subsamples of the two treatments in the experiment.

From Table 1 it can be seen that only about 85 per cent of the addresses in the gross sample were visited by an interviewer. This is because of field staff capacity problems and sample frame errors. During the nineties, Statistics Netherlands's field staff faced

Table 1. Response account

Category	Regular questionnaire		New questionnaire	
	Number	Per cent	Number	Per cent
Gross sample (addresses)	8,262	100	8,385	100
Not visited addresses	822	10	706	8
Uninhabited or untraceable addresses	496	6	473	6
Visited addresses	6,944	84	7,206	86
Visited households	7,132	100	7,367	100
Refusals	1,785	25	1,759	24
No contacts	918	13	965	13
Other nonresponse	583	8	524	7
Responding households	3,846	54	4,119	56
Households finally used in the analysis	3,037	43	3,159	43

increasing capacity problems. As a result, a gradually increasing part of the sample addresses were not assigned to or visited by interviewers. In the experiment this caused a loss of about ten per cent of the sample addresses. Moreover about six per cent of the gross sample addresses were uninhabited or untraceable.

According to the response definition used by the Data Collection Department, about 55 per cent of the visited households responded in the experiment. The response rate obtained in the regular annual LFS in 1999 was also about 55 per cent. This response rate is a continuation of the downward trend of the Dutch LFS response rates during the last decades reported by De Heer (1999).

A part of the completed questionnaires obtained from the responding households are not used in the analysis of the experiment for the following reasons. Firstly, in the regular LFS only completely responding households are used. About five per cent of the completed questionnaires concerned partially responding households and were therefore excluded from the experiment. Secondly, the response obtained in a block during one month was used in the analysis only if each of the two interviewers in that block obtained a response of at least three households. In several blocks one interviewer did not have any response at all during a month, because of for example illness or vacation. In such situations the response obtained in a block for the other interviewer during that month was excluded from the analysis in order to avoid any seasonal effects that might influence the analysis. As a result, finally 3,037 households interviewed with the regular questionnaire and 3,159 households interviewed with the new questionnaire were used in the analysis of the experiment. After these exclusions 86 blocks remained usable.

The low response rates might result in biased parameter estimates. However, there is no evidence of differential nonresponse bias in the parameter estimates across treatments. First, no significant differences could be found across the two subsamples with respect to background variables like age class, marital status, and sex, since the p-values of the corresponding chi-square tests varied between 0.5 and 0.3. Second, the response rates in the two subsamples are comparable (see Table 1). Consequently, possible nonresponse bias in the parameter estimates under both treatments conveniently cancels out by calculating treatment effects or contrasts between the parameter estimates. Therefore, the analysis results of the experiment will be more robust against nonresponse bias than the

separate parameter estimates. Moreover, the analysis is based on the generalized regression estimator that also corrects, at least partially, for nonresponse bias.

4.2. *Effects of the new questionnaire*

The purpose of this experiment is to quantify possible effects on the estimates of the population parameters of the LFS induced by the different questionnaires. This implies that the null hypothesis of no treatment effects is tested against the unrestricted alternative hypothesis, specified by H_{1a} in (1). The following five parameters of the LFS were analyzed: the Employed Labor Force, the Unemployed Labor Force, the Registered Unemployment, the Employed Labor Force according to the International Labour Organisation (abbreviated as ILO Employed) and the Registered at the Employment Exchange. All parameters are expressed as percentages of the population aged 15 through 64 years.

The analysis of the experiment was based on the design-based t-statistic (4), where the population parameters and the variance of the corresponding contrasts are estimated with the generalized regression estimator for an RBD. This is the most appropriate analysis since it takes both the experimental design and the design of the LFS, including the estimation and weighting procedure, into account.

The observations y_{jk} and the auxiliary variables $\mathbf{x}_{jk}$ in the formulas of Section 2 are on the level of households, since they are the experimental units in this application. Let y_{ijk} denote the observation obtained from person i in household j assigned to treatment k. Then $y_{jk} = \sum_{i=1}^{m_j} y_{ijk}$ is the observation obtained from household j assigned to treatment k, with m_j the number of individuals in household j aged 15 years and over. Let $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijH})^t$ denote a vector of order H with each element x_{ijh} an auxiliary variable of individual i in household j. Then $\mathbf{x}_j = \sum_{i=1}^{m_j} \mathbf{x}_{ij}$ is the vector of order H with the household totals of the auxiliary variables of the individuals aged 15 years and over. The weighting procedure is carried out at the level of household totals. In Linear regression model (3) it is assumed that $\mathrm{Var}(e_j) = m_j\omega^2$. Nieuwenbroek (1993) showed that in this situation the weighting procedure corresponds to the integrated method for weighting individuals and households proposed by Lemaître and Dufour (1987). The net sample sizes of both subsamples are small compared with the sample of the regular LFS. Based on these sample sizes, the following weighting scheme, which contains the most important auxiliary information from the regular weighting scheme of the LFS, was applied in the analysis of this experiment: $age \times region + sex \times region + marital\ status \times region$, where the four variables are categorical. Furthermore Variance estimator (5) was applied. Results are given in Table 2.

Tests were conducted at a significance level of 0.05. The Employed Labor Force and the ILO Employed observed with the new questionnaire are about 1.7 percentage points smaller than with the regular one. Although the tests of no treatment effects for these parameters are not rejected, the differences are substantial. Further investigation showed that these differences were caused by the Self-Employed. With the new questionnaire the differences were found to be 1.8 percentage points smaller than with the regular one. Since the corresponding p-value amounts to 0.01, this difference is significant. Changes in the routing of the new questionnaire caused many self-employed people to answer the question ''Do you currently have a paid job?'' in the negative. Consequently these

Table 2. Results with the generalized regression estimator for an RBD

Parameter	Regular	New	Diff.	Std. error	*t*-statistic	*p*-value
Employed labor force	64.710	63.046	−1.663	1.006	−1.653	0.098
Unemployed labor force	2.403	3.051	0.649	0.337	1.923	0.054
Registered unemployment	2.136	2.612	0.476	0.324	1.471	0.141
ILO employed	71.490	69.741	−1.749	0.985	−1.776	0.076
Registered at empl. exc.	7.640	9.072	1.433	0.568	2.524	0.012

self-employed were erroneously not classified in the Labor Force. This was confirmed by the debriefings of the interviewers. This resulted in the observed differences for the Employed Labor Force and the ILO Employed. The new questionnaire was adjusted in order to avoid this under-reporting before it was actually implemented as the standard questionnaire in the LFS.

The Unemployed Labor Force and the Registered Unemployment measured with the new questionnaire are about 0.5 percentage points higher than with the regular questionnaire. These differences are substantial and almost significant for the Unemployed Labor Force.

For the Registered at the Employment Exchange the observed difference of 1.4 percentage points is significant. In the regular questionnaire, the question about registration at the Employment Exchange was preceded by several questions concerning benefits. In the new questionnaire the questions about benefits were skipped. Possible interactions between the questions concerning benefits and registration at the Employment Exchange might explain the observed difference. Despite this difference, the questions about benefits are not included in the new questionnaire since this information will be available from registrations.

4.3. Comparisons with other analyses

In this section, the above analysis is compared with some other possible design- and model-based analyses. First consider a design-based analysis where the population parameters and the variance of the corresponding contrasts are estimated with the Horvitz-Thompson estimator for an RBD. In this application the common mean model, which results in the ratio estimator for a population mean (7), is specified by $\mathbf{x}_{ij} = (1)$, hence $\mathbf{x}_j = (m_j)$, and $\mathrm{Var}(e_j) = m_j\omega^2$. An approximately design-unbiased estimator for the variance of the treatment effects is given by (5) with $\hat{e}_{jk} = y_{jk} - m_j\,\tilde{y}_{s_k}$. Results are given in Table 3.

Compared with the Horvitz-Thompson estimator the generalized regression estimator, discussed in Section 4.2, has two effects on the analysis. Firstly, the size of the treatment effects is reduced. The estimates of the parameters under both treatments are more similar due to the application of the auxiliary information in the weighting scheme. Secondly, as expected, the variance of the estimated treatment effects is reduced. For the Unemployed Labor Force and the Registered Unemployment this finally resulted in higher *p*-values and for the other parameters in lower *p*-values for the treatment effects.

In order to compare the efficiency of designing this experiment as an RBD, the experiment was analyzed as if it were designed as a CRD. Results for the generalized regression estimator are given in Table 4 and for the Horvitz-Thompson estimator in Table 5.

Table 3. Results with the Horvitz-Thompson estimator for an RBD

Parameter	Regular	New	Diff.	Std. error	*t*-statistic	*p*-value
Employed labor force	65.306	62.667	−2.639	1.763	−1.497	0.134
Unemployed labor force	2.246	2.954	0.709	0.337	2.103	0.035
Registered unemployment	1.935	2.515	0.580	0.327	1.773	0.076
ILO employed	72.197	69.263	−2.934	1.877	−1.563	0.118
Registered at empl. exc.	7.265	8.542	1.277	0.543	2.351	0.019

Comparing the results of an analysis for an RBD with those for a CRD using the generalized regression estimator (Tables 2 and 4), it follows that the variances of the estimated treatment effects are approximately equal or even slightly smaller under a CRD. Comparing the results of an analysis for an RBD with those for a CRD, both analyzed with the Horvitz-Thompson estimator (Tables 3 and 5), it follows that the variance of the estimated treatment effects under a CRD is slightly larger.

The efficiency of blocking is small in this application, which can be explained as follows. Firstly, the regional variation between the parameters of the LFS appears to be small. Secondly, only if the fraction of households assigned to a treatment within each block is equal for each block (i.e., $n_{bk}/n_{b+} = n_{b'k}/n_{b'+}$), can it be proved that the variance of the estimated treatment effects for a CRD is larger than such variance for an RBD (Van den Brakel 2001, Section 6.3). For the gross sample these fractions are equal by the design of this experiment. However, due to unequal response rates of the two interviewers within several blocks, the equality of these fractions for the net sample was disturbed. This resulted in an extra variation of the sample weights $n_{b+}/(n_{bk}\pi_j)$ in the analysis of an RBD. This may be an explanation for the small variance reduction due to the application of an RBD for the Horvitz-Thompson estimator. The weighting scheme of the generalized regression estimator eliminates some regional variation from the estimated treatment effects. Therefore, the negative influence of the unequal allocation of the households over the treatments within the blocks of an RBD becomes more obvious if the analysis is conducted with the generalized regression estimator. This might explain why the variances of the treatment effects are even slightly smaller under a CRD for the generalized regression estimator. In summary, due to different response rates obtained by the interviewers, the optimality of the RBD was at least partially disturbed. This might have been avoided if interviewers had been used as block variables.

The effect on the analysis of the application of the generalized regression estimator compared with the Horvitz-Thompson estimator in the case of CRD can be seen from Tables 4 and 5. As in the case of an RBD (Tables 2 and 3), it follows that the generalized

Table 4. Results with the generalized regression estimator for a CRD

Parameter	Regular	New	Diff.	Std. error	*t*-statistic	*p*-value
Employed labor force	64.711	63.067	−1.643	0.986	−1.666	0.096
Unemployed labor force	2.396	3.025	0.629	0.338	1.860	0.063
Registered unemployment	2.152	2.511	0.359	0.313	1.147	0.251
ILO employed	71.499	69.694	−1.805	0.968	−1.865	0.062
Registered at empl. exc.	7.495	9.132	1.636	0.546	2.997	0.003

Table 5. Results with the Horvitz-Thompson estimator for a CRD

Parameter	Regular	New	Diff.	Std. error	*t*-statistic	*p*-value
Employed labor force	64.779	63.014	−1.764	1.871	−0.943	0.346
Unemployed labor force	2.265	2.975	0.710	0.343	2.069	0.039
Registered unemployment	1.913	2.471	0.558	0.343	1.627	0.104
ILO employed	71.721	69.495	−2.226	1.993	−1.117	0.264
Registered at empl. exc.	7.062	8.732	1.670	0.552	3.023	0.003

regression estimator reduces the differences between the treatments as well as the variance of the estimated treatment effects.

To compare the results obtained by a design-based analysis with the results obtained by a model-based analysis, this experiment was also analyzed with the model-based *t*-statistic and an ANOVA for an RBD. For both analyses a fixed effect model was applied. To approximate the target variables of the LFS as much as possible in these analyses, the dependent variables are the household means of the target parameters. Results are summarized in Tables 6 and 7.

The influence of a model-based versus a design-based analysis differs for the five parameters of the LFS. To see this, some of the possible comparisons between Tables 2 and 7 (RBD's) and between Tables 4 and 6 (CRD's) are highlighted. It can be seen from Table 6 that the Unemployed Labor Force, the Registered Unemployment and the Registered at Employment Exchange are overestimated while the Employed Labor Force and the ILO Employed are underestimated. This is mainly caused by the fact that the oversampling of addresses which occurs in the register of the Employment Exchange is ignored in the model-based analysis. This results in biased estimates for the population parameters. Especially the treatment effects for the Unemployed Labor Force, the Registered Unemployment and the ILO Employed are larger and have smaller *p*-values in a model-based analysis. In a design-based analysis the variance of the estimated treatment effects might be increased due to the fluctuation of the sampling weights. Nevertheless, the variance of the estimated treatment effects for the Unemployed Labor Force, the Registered Unemployment and the ILO Employed in the model-based analysis is slightly higher than in the design-based analysis. It is unclear why, contrary to the results observed for the Unemployed Labor Force and the Registered Unemployment, the estimated treatment effect of the Registered at Employment Exchange is smaller and has a larger *p*-value in a model-based approach.

The efficiency of an RBD in a model-based analysis can be seen by comparing Tables 6 and 7. Although in the ANOVA for an RBD, the *F*-statistic for the blocks is significant at a level of 0.05, the reduction in the error sum of squares is very small. Consequently, the

Table 6. Results with the model-based t-statistic (unweighted)

Parameter	Regular	New	Diff.	Std. error	*t*-statistic	*p*-value
Employed labor force	62.13	60.49	−1.646	0.991	−1.661	0.097
Unemployed labor force	3.716	4.708	0.992	0.419	2.370	0.018
Registered unemployment	3.836	4.839	1.003	0.441	2.275	0.023
ILO employed	68.68	66.44	−2.240	0.983	−2.279	0.023
Registered at empl. exc.	14.68	16.04	1.361	0.787	1.728	0.084

Table 7. Results with the ANOVA for an RBD (unweighted)

Parameter	Contrast	Std. error	*t*-statistic	*p*-value
Employed labor force	−1.372	1.00	−1.370	0.171
Unemployed labor force	1.025	0.40	2.413	0.016
Registered unemployment	1.028	0.40	2.306	0.021
ILO employed	−1.795	1.00	−1.818	0.069
Registered at empl. exc.	0.678	0.80	0.860	0.390

variance reduction of the estimated treatment effects due to the application of an RBD is negligible. For the Employed Labor Force, the ILO Employed and the Registered at the Employment Exchange, the contrasts are smaller due to the correction for the block variables, resulting in larger *p*-values for the treatment effects.

In conclusion, it follows that the results obtained with the model-based analysis with the *t*-test as well as the ANOVA F-test for an RBD are biased since the specific features of the weighting procedure of the LFS, especially the oversampling of addresses with individuals registered at an Employment Exchange, is ignored in these analyses. Because of these biases, clearly the application of a design-based approach has demonstrated its considerable merit in this setting.

5. References

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

Bethlehem, J.G. and Keller, W.G. (1987). Linear Weighting of Sample Survey Data. Journal of Official Statistics, 3, 141–153.

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

De Heer, W. (1999). International Response Trends: Results of an International Survey. Journal of Official Statistics, 15, 129–142.

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.

Hájek, J. (1971). Comment on a paper by D. Basu. In: Foundations of Statistical Inference, ed. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart, and Winston, 236.

Hilbink, K., Van Berkel, C.A.M., and Van den Brakel, J.A. (2000). Methodology of the Dutch Labour Force Survey, 1987–1999. Research paper, BPA nr.: 2297-00-RSM, Department of Statistical Methods, Statistics Netherlands, Heerlen.

Lemaître, G. and Dufour, J. (1987). An Integrated Method for Weighting Persons and Families. Survey Methodology, 13, 199–207.

Nieuwenbroek, N.J. (1993). An Integrated Method for Weighting Characteristics of Persons and Households Using the Generalized Regression Estimator. Research paper, BPA nr.: 8445-93-M1, Department of Statistical Methods, Statistics Netherlands, Heerlen.

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

Van Berkel, C.A.M. and Van der Valk, J. (1999). Restructuring the Dutch Labour Force Survey. Netherlands Official Statistics, 14, autumn 1999, 18–20.

Van den Brakel, J.A. and Renssen, R.H. (1998). Design and Analysis of Experiments Embedded in Sample Surveys. Journal of Official Statistics, 14, 277–295.

Van den Brakel, J.A. (2001). Design and Analysis of Experiments Embedded in Complex Sampling Designs. PhD Thesis, Erasmus University, Rotterdam.

Received May 2001
Revised April 2002