Design-Based Analysis of Embedded Experiments with Applications in the Dutch Labour Force Survey

Author(s): Jan A. van den Brakel

Source: *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 171, No. 3 (2008), pp. 581-613
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/30135087
Accessed: 28-10-2018 02:27 UTC

J. R. Statist. Soc. A (2008)
171, *Part* 3, *pp.* 581–613

Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey

Jan A. van den Brakel

Statistics Netherlands, Heerlen, The Netherlands

[Received September 2006. Final revision October 2007]

Summary. Previous research has proposed a design-based analysis procedure for experiments that are embedded in complex sampling designs in which the ultimate sampling units of an on-going sample survey are randomized over different treatments according to completely randomized designs or randomized block designs. Design-based Wald and t-statistics are applied to test whether sample means that are observed under various survey implementations are significantly different. This approach is generalized to experimental designs in which clusters of sampling units are randomized over the different treatments. Furthermore, test statistics are derived to test differences between ratios of two sample estimates that are observed under alternative survey implementations. The methods are illustrated with a simulation study and real life applications of experiments that are embedded in the Dutch Labour Force Survey. The functionality of a software package that was developed to conduct these analyses is described.

Keywords: Measurement error models; Probability sampling; Randomized experiments; X-tool

1. Introduction

Randomized experiments that are embedded in sample surveys are frequently used to test the effects of one or more adjustments in a survey process on response rates or parameter estimates of an on-going survey. In survey methodology literature one finds many references to experimental studies on improving the quality or efficiency of survey processes, e.g. studies to compare the effect of different questionnaire designs, modes of data collection or approach strategies on the main outcomes of a sample survey, with the purpose of reducing response bias or improving response rates.

At national statistical offices such experiments are particularly useful to quantify discontinuities in the series of repeated surveys due to adjustments in the survey process. The Dutch Labour Force Survey (LFS), for example, is continuous and makes up a series that describes the development of indicators about the situation in the labour market. Comparability over time is a key aspect of the relevance of these figures. Modifications in the survey process should not result in unexplained differences in the series of the employed and unemployed labour force. This paper deals with a series of experiments that are embedded in the LFS which are aimed at quantifying the effect of alternative questionnaires, modes of data collection and approach strategies on the estimates of the employed and unemployed labour force. These applications

Address for correspondence: Jan A. van den Brakel, Department of Statistical Methods, Statistics Netherlands, PO Box 4481, 6401 CZ, Heerlen, The Netherlands.
E-mail: jbrl@cbs.nl

 0964–1998/08/171581

are used to illustrate the concepts for embedding experiments in on-going sample surveys and the need for a design-based theory for the analysis of such experiments.

The idea of embedding experiments in on-going sample surveys was probably first introduced by Mahalanobis (1946) to test interviewer variance in survey sampling. Fienberg and Tanur (1987, 1988, 1989, 1996) discussed the fundamentals, the parallels and the differences between randomized sampling and randomized experiments and detailed the strategies for design and analysis of embedded experiments. Two other key references are Fellegi (1964) and Hartley and Rao (1978).

Embedding experiments in sample surveys implies that first a sample is drawn from a finite target population by means of the probability sample of the sample survey. Next, the sample is randomly divided into $K \geqslant 2$ subsamples according to an experimental design. Each subsample is assigned to one of the K alternative survey procedures or treatments that are compared in the experiment. The objective in these applications is to estimate finite population parameters under the different survey implementations or treatments, and to test whether these parameter estimates are significantly different. van den Brakel (2001), van den Brakel and van Berkel (2002) and van den Brakel and Renssen (1998, 2005) developed a design-based procedure for the analysis of completely randomized designs (CRDs) and randomized block designs (RBDs). In this approach, a design-based estimator for the population parameter that is observed under each treatment, as well as the covariance matrix of the differences between these estimates, is derived by using the Horvitz–Thompson estimator or the generalized regression estimator. These estimators account for the probability structure that is imposed by the sampling design, the randomization mechanism of the experimental design and the weighting procedure applied in the on-going survey for the estimation of target parameters. This gives rise to a design-based Wald or t-statistic, to test hypotheses about differences between population means or totals that are measured under alternative survey implementations.

In this approach ultimate sampling units are randomized over the treatments. Owing to restrictions in fieldwork, clusters of sampling units instead of separate sampling units are randomized over the treatments in many practical applications. As a result, the level of randomization in the experimental design does not always coincide with the ultimate sampling units in the sampling design. For example, it may be necessary from a practical point of view to randomize primary sampling units (PSUs), or clusters of sampling units that belong to the same household or are assigned to the same interviewer over the different treatments in the experimental design. In this paper the design-based approach of van den Brakel and van Berkel (2002) and van den Brakel and Renssen (2005) is extended to experiments where clusters of sampling units are randomized over the different treatments. Furthermore, the methods are extended to test hypotheses about population parameters that are defined as the ratio of two population totals.

Section 2 deals with a series of experiments that are embedded in the Dutch LFS. Some aspects of designing embedded experiments are discussed in Section 3. The technical details of a design-based analysis, where clusters of sampling units of the sampling design are the experimental units in the experiment and hypotheses about ratios are tested, are detailed in Section 4. In Section 5 the properties of the variance estimator and Wald statistic proposed are further investigated with a simulation study. At Statistics Netherlands, a software package was developed to support the analysis procedures proposed. The functionality of this package is described in Section 6. In Section 7 the methodology is illustrated with a real life experiment that was embedded in the Dutch LFS to test the effect of different incentives on response rates and response bias. Some general remarks are made in Section 8.

2. Examples of experiments embedded in the Dutch Labour Force Survey

The survey design of the Dutch LFS is summarized in Section 2.1. Five illustrative experiments are described in the other sections.

2.1. Survey design

The LFS is based on a rotating panel survey. Each month a stratified two-stage cluster sample of about 7500 addresses is drawn from a register of all known addresses in the Netherlands. Strata are formed by geographical regions, municipalities are considered as PSUs, and addresses as secondary sampling units (SSUs). Addresses of people aged 65 years and over are undersampled, since the target parameters of the LFS concern people aged 15–64 years. All households, with a maximum of three, living at an address, are included in the sample.

In the first wave, data are collected by means of computer-assisted personal interviewing (CAPI) by using laptops. Interviewers collect data for the LFS in areas that are close to where they live. Demographic variables are observed for all members of the households selected. Only people who are aged 15 years and over are interviewed for the target variables. When a household member cannot be contacted, proxy interviewing is allowed with members of the same household. Households in which one or more selected people do not respond themselves or in a proxy interview are treated as non-responding households. The respondents are reinterviewed four times at quarterly intervals. In these four subsequent waves, data are collected by means of computer-assisted telephone interviewing (CATI). During these reinterviews a condensed questionnaire is applied to establish changes in the labour market position of the household members aged 15 years and over. Proxy interviewing is also allowed during these reinterviews.

The weighting procedure of the LFS is based on the generalized regression estimator of Särndal *et al.* (1992) which will be discussed in Section 4.2. The inclusion probabilities reflect the undersampling of addresses that was described above as well as the different response rates between geographical regions. The weighting scheme is based on a combination of different sociodemographic categorical variables. The integrated method for weighting people and families of Lemaître and Dufour (1987) is applied to obtain equal weights for people belonging to the same household.

The most important parameters of the LFS are total unemployment and the employed and unemployed labour force. The unemployed labour force is defined as the ratio of total unemployment and the total labour force. The employed labour force is defined as the ratio of total employed labour force and the total population aged 15–64 years.

2.2. Experiments with new questionnaire designs

The LFS questionnaire was revised in 1999. The questions were grouped in a different order, the wording of questions changed and a block of questions about receiving social benefits was deleted since this information is also available from registrations. It was anticipated that the introduction of the new questionnaire would change the measurement errors that are induced by the design of the questionnaire and have systematic effects on the outcomes of the LFS. Therefore a large-scale field experiment was conducted to quantify the effects of the new questionnaire on the main parameter estimates of the LFS. This enabled us to separate the real development of the employed and unemployed labour force from the systematic effect of the new questionnaire on these parameter estimates.

From April 1999 to September 1999 the monthly sample was divided in two subsamples in a randomized experiment. About 80% of the monthly samples were assigned to the normal

questionnaire. The data that were obtained in this subsample are used for official publication purposes of the LFS but also served as the control group of the experiment. The remaining 20% of the sampling units were assigned to the new questionnaire. It was decided to assign interviewers to one of the two questionnaires only, to avoid confusion of the questionnaires by the interviewers during the data collection. Furthermore, it was impossible to run both questionnaires on the same hand-held computer, since the new questionnaire was supported by a Windows version of Blaise whereas the normal questionnaire was supported by a non-Windows version of Blaise. It is not feasible for interviewers to visit households with two hand-held computers.

On the basis of these considerations, the experiment was designed as a two-treatment RBD. Each block consisted of two neighbouring interview areas of two interviewers. The two interviewers as well as the addresses in the monthly sample of the LFS in each block were randomly assigned to the experimental and the control group. Within each block the interviewer visits the addresses that are assigned to his or her treatment.

The purpose of this type of experiment is to estimate the employed and unemployed labour force by using the data that were obtained with the normal and the new questionnaire, and to test whether these estimates are significantly different. Such an analysis should typically account for the sample design and estimation procedure of the LFS as described in Section 2.1. In van den Brakel and van Berkel (2002) such a design-based approach was proposed for two-treatment experiments and was used to analyse this experiment. The revisions of the questionnaire resulted in significantly higher estimates for the unemployed labour force and those who were registered at the employment exchange. See van den Brakel and van Berkel (2002) for a detailed discussion.

At the request of the Ministry of Social Services and Employment, the LFS questionnaire was extended in 2005 with a module containing questions about combining paid employment and care activities for ill partners or other relatives. The extension of the LFS questionnaire with this module must not result in inexplicable discontinuities in the series about the employed and unemployed labour force. Therefore the effect of adding this module on the main parameters of the LFS was tested in a large-scale field experiment.

This experiment was conducted from January to June 2005. The households that were included in the monthly sample were assigned to interviewers. Subsequently three-quarters of the households of each interviewer were randomly assigned to the normal questionnaire and a quarter to the questionnaire with the additional module. This implies that the experiment is designed as an RBD with interviewers as the block variable. In this case, the interviewers collect data with both questionnaires simultaneously. It was expected that the interviewers would not confuse both questionnaires, since the only difference between the normal and the new questionnaire was an additional, clearly separated block of questions.

The data that were obtained in both subsamples are used to estimate the unemployed labour force and the total unemployment according to the estimation procedure that was described in Section 2.1. Subsequently the difference between the two estimates for each parameter obtained under the two questionnaires was tested by using the design-based approach of van den Brakel and van Berkel (2002) with the software package X-tool, which is described in Section 6. In this experiment a total of 5750 and 17 500 households completed a new and a normal questionnaire. With these subsample sizes, differences larger than or equal to 55 000 in the estimated total unemployment could be observed at a level of significance of 5%. The estimated total unemployment at the time that this experiment was conducted amounted to about 500 000. Under the new questionnaire the estimate for total unemployment dropped by 15 000 (p-value 0.60) and the estimated unemployed labour force by 0.24 percentage points (p-value 0.52), so the observed differences were not significantly different.

2.3. Experiment with different data collection modes

In 2001 an experiment was conducted to test the effect on the estimates of the employed and unemployed labour force if the data in the first wave were collected by means of CATI instead of CAPI. In this period there was a structural lack of capacity for the CAPI fieldwork organization, particularly in the more urban regions of the Netherlands. One solution to this capacity problem is to conduct the data collection in the first wave partially by means of CATI. Before the data collection in the first wave switched from a unimode design by means of CAPI to a mixed mode design through CAPI and CATI, it should be established that this produces no large mode effects on the estimates of the employed and unemployed labour force.

From August to December 2001 an experiment was conducted to quantify mode effects. Households in the monthly samples with a listed permanent telephone connection were randomized over the CAPI and CATI data collection mode by means of an RBD, using geographical regions as the block variable. About 90% of these households were assigned to the regular data collection mode, i.e. CAPI. The remaining 10% were assigned to the CATI mode. In this experiment 9900 households responded under the CAPI and 1100 under the CATI mode. On the basis of data that were observed in both subsamples, hypotheses were tested about mode effects in the estimates of the employed and unemployed labour force, again by using the design-based approach of van den Brakel and van Berkel (2002) with X-tool. The hypotheses that there are no mode effects was rejected, since the estimate for the unemployed labour force dropped by about 1.1 percentage points (p-value 0.017) under the CATI mode and the estimate for the employed labour force increased by 2.5 percentage points (p-value 0.008). Possible explanations are the increased fraction of proxy interviews under the CATI mode, differences in the perception of privacy of the respondent and differences in the speed of interview between these modes. To avoid discontinuities in the series of the employed and unemployed labour force, it was decided not to change the data collection mode in the first wave.

2.4. Experiment with a new advance letter

In an attempt to improve the LFS response rates, a new more informal advance letter for the LFS was developed. The effect on the response rates of this new advance letter was tested by means of an experiment from January to March 2004. The purpose of this experiment was to detect small differences in response and refusal rates. This could be achieved without making substantial additional costs, since in this application the households that were assigned to the standard as well as the experimental advance letter are both used for the official publication purposes of the LFS. During 3 months a third of the sample addresses of each interviewer were randomly assigned to the new letter and two-thirds to the usual letter. This resulted in an RBD with interviewers as the block variable with a gross subsample size of 8073 households for the new letter and 16155 households for the usual letter. With an average response rate of 55%, these subsample sizes give rise to an experiment where differences larger than or equal to 1.3% could be detected at a level of significance of 5%. Finally, the response rate that was obtained with the new letter was 1.4% smaller and the refusal rate was 1.8% higher than with the usual letter. A logistic regression analysis where interviewers (blocks) and letter (treatment) were used as the explanatory variables demonstrated that the new letter had a significant negative effect on response behaviour. So the usual advance letter was not replaced.

2.5. Experiment with incentives

Statistics Netherlands does not pay respondents for their participation in a survey. Therefore, there are generally no incentives in the approach strategies for surveys that are conducted by

Table 1. Overview treatments

Treatment			Subsample fraction (%)
Number	*Description*	*Value(€)*	
1	No incentive	0.0	48
2	Stamp booklet containing 5 stamps	1.95	24
3	2 stamp booklets containing 10 stamps in total	3.9	24
4	4 stamp booklets containing 20 stamps in total	7.8	4

Statistics Netherlands. There are, however, many references to experiments demonstrating the positive effect of incentives on response rates (Groves and Couper, 1998). To explore the possibilities for improving response rates and for reducing non-response bias in the Dutch LFS, an embedded experiment with small prepaid incentives was conducted in 2005.

In this experiment the effects of three differently valued incentives were compared with a control group, where no incentive was applied. Booklets of stamps of different values are used as a prepaid incentive (Table 1). The booklets of stamps were included with the advance letter that was sent to the sampled households before the interview for the first wave. This experiment was conducted in November and December 2005. The gross sample was randomly divided into four subsamples according to an RBD with interviewers as the block variable. The fractions that were used to split the sample into four subsamples are specified in Table 1.

The purpose of this experiment was to test hypotheses about incentives on response rates and response bias. A logistic regression analysis was applied to test hypotheses about effects on response and refusal rates. To investigate the effect on response bias, it is tested whether the parameter estimates of the unemployed labour force and total unemployment that were obtained with the subsamples assigned to the four different incentives are significantly different. This experiment is analysed in Section 7.

3. Design of embedded experiments

3.1. Embedding experiments in sample surveys

A major advantage of embedded experiments is the random selection of the sampling units from a finite target population. This makes them appropriate to test whether a modification in the survey process or a complete survey redesign yields a higher response rate or lower response bias, or whether cheaper methods do not yield a lower response or lower quality of data.

The examples in the previous section illustrate experiments in the LFS which were undertaken to ensure that deliberate modifications or redesigns of the survey process did not lead to unexplained discontinuities in the series of the employed and unemployed labour force. Running the normal and the new approach in parallel by means of an embedded experiment provides a safe survey transition process, since the new approach is conducted in a full-scale sample before its formal implementation. This reduces several continuity risks. Quantifying and explaining the effect of a redesign avoid the confounding of real developments that are described by the series with the effect of the adjustments in the underlying survey process. This reduces the negative effect of the redesign of a repeatedly conducted survey on the continuity of the series. Finally, if the new approach turns out to be a failure, this still leaves the possibility of keeping the old approach for official publication purposes without having a period for which no reliable figures

are available. For instance, the experiment with the advance letter showed that the new letter resulted in a reduced response rate whereas the opposite effect was expected. The experiment with the alternative data collection mode showed that the introduction of a mixed data collection mode would give problems with the integrity of the data that are collected in the first wave. On the basis of the results of these experiments, the intended modifications in the survey process were not introduced.

Embedding experiments in sample surveys is efficient from a financial point of view. In the applications in Section 2 there is one relatively large subsample that is assigned to the usual survey, which serves not only as the official publication purposes for the LFS but also as the control group in the experiment. In some situations even the data that are obtained in the subsamples assigned to the alternative treatments can be used for the official publication purposes of the on-going survey, e.g. the experiment with the advance letters in Section 2.4. Nevertheless it should be realized that two virtually competing objectives are combined in an embedded experiment. The purpose of the normal survey is to estimate population parameters as precisely as possible, so the subsample that is assigned to the normal survey should be maximized. The purpose of the experiment, in contrast, is to estimate contrasts between the population parameters that are observed under different survey implementations as precisely as possible. This implies that the subsample sizes should preferably be equal, since balanced designs maximize the power of the tests about treatment effects (see for example Montgomery (2001)). The fractions that are used to split the usual survey into subsamples in the different examples of Section 2 are always a trade-off between an acceptable loss of precision for the on-going survey, the sample size required to detect prespecified differences at a certain power and significance level, and the time that is available to conduct the experiment.

3.2. Design considerations

A CRD is the most straightforward approach to divide a sample randomly into K subsamples. However, the application of unrestricted randomization is generally not the most efficient design available. Fienberg and Tanur (1987, 1988, 1989) argued that the application of an RBD with sampling structures like strata, PSUs, clusters and interviewers as block variables might improve the precision of an experiment considerably. The response that is obtained from sampling units that are drawn from the same strata, cluster or PSU, or are assigned to the same interviewer are generally more homogeneous than sampling units from different strata, clusters or interviewers. Using these sampling structures as a block variable in an experiment increases the power of the experiment and also guarantees that each stratum or PSU is sufficiently represented in each subsample. This last property is particularly important if the subsamples that are assigned to the alternative treatments are small compared with the subsample that is assigned to the usual survey or control group. All the examples that were discussed in Section 2 are indeed designed as RBDs.

Interviewers require special attention in the planning and designing stage of an experiment. It should be considered carefully whether interviewers are assigned to all or only one of the treatments in the experiment. There is a trade-off between the increased power of the experiment if interviewers are used as a block variable, and the simplicity for organizing the fieldwork if interviewers are assigned to only one of the treatments. Using interviewers as block variables implies that each interviewer must conduct each treatment. This might result in practical problems with conducting the fieldwork. From a statistical point of view, it is worthwhile to make an all-out effort to use interviewers as a block variable in an RBD. A part of the variation in response rates is determined by the interviewers' personal abilities to persuade respondents to

participate in the survey. It is also known that interviewers induce additional variance, since they may affect the responses that are given by respondents in personal interviews. This implies that the power of an experiment can be improved if interviewers are used as the block variable in an RBD.

Whether it is achievable to use interviewers as blocks depends on the number and type of treatments and the field staff's experience with collecting data under different treatments simultaneously. For example, different wordings in different versions of a questionnaire might be mixed up easily by inexperienced interviewers and hamper the application of an RBD with interviewers as the block variable. In the first experiment with a new LFS questionnaire, which was discussed in Section 2.2, it was decided to assign interviewers to only one treatment for different reasons. The question blocks were organized in a different order, the routing changed, questions were dropped and the wording changed. It was anticipated that the interviewers would easily mix up the different treatments, since there was hardly any experience with data collection under different treatment settings within the same survey at the time that this experiment was conducted. Since then interviewers at Statistics Netherlands have frequently been used as the block variable in an RBD, as the experience of the field staff with embedded experiments increased, as in the second experiment with an alternative questionnaire that was discussed in Section 2.2 or the experiment with incentives in Section 2.5. It was not feasible to use interviewers as blocks in the experiment with different data collection modes since the data collection by means of CAPI and CATI is organized in different departments with their own interviewers. In the experiment with a new advance letter, which was discussed in Section 2.4, the interaction between interviewers and respondents is hardly affected by the different treatments, so the interviewers could be used as blocks without risk.

Assigning interviewers to only one treatment can be accomplished as follows in a CATI survey. Sampling units and interviewers are randomly assigned to the different treatments, independently of each other. Subsequent sampling units are assigned to interviewers within each subsample or treatment.

In a CAPI survey, where interviewers are working on the data collection in relatively small areas around their own place of residence, unrestricted randomization of sampling units and interviewers over the treatments is often not applicable. This randomization mechanism might result in an unacceptable increase in the distance to travel for interviewers, particularly if the sample sizes of the subsamples that are assigned to the alternative treatments are small. An alternative is to assign sampling units to interviewers. Subsequently the interviewers with their cluster of sampling units are randomized over the treatments of the experiment. Here, however, the interviewers are the experimental units instead of the sampling units, which decrease the effective sample size for variance estimation and power for testing hypotheses. One compromise is to use geographical regions which are linked adjacent interviewer regions as a block variable. The sampling units within each block are randomized over the treatment combinations, and each interviewer within each block is randomly assigned to one of the treatment combinations. This implies that the number of interviewers in each block must be equal to the number of treatments. Subsequently each interviewer visits the sampling units that are assigned to his or her treatment combination. This results in a relatively small increase in the distance to travel for the interviewers and no increase in variance, since the sampling units are the experimental units in this design. The first experiment that was discussed in Section 2.2 with a new questionnaire is an example of this design.

Another consideration is whether the interviewers should be informed that they are participating in an experiment or not. The advantage of keeping interviewers uninformed is that they do not adjust their behaviour because they are aware that their performance is supervised in an

experiment. It depends on the treatments whether it is possible to keep interviewers uninformed and to apply an RBD with interviewers as the block variable. In the LFS examples the interviewers were informed that they were participating in an experiment, not only because they had to collect data under different treatments, but also they had to share their practical knowledge in the design and analysis through debriefings. Finally there are ethical reasons. A risk of keeping interviewers uninformed is that their loyalty in future projects may suffer if they accidentally find out that they were involved in an experiment.

An experiment that is embedded in a two-stage sampling design can be designed as an RBD where PSUs are the block variable and SSUs the experimental units, or as an experiment where PSUs are the experimental units. Often, the variation between SSUs within PSUs is small compared with the variation between the PSUs. This implies that the power of the experiment is increased if PSUs are the block variables and SSUs the experimental units, since

(a) the variation between PSUs is eliminated from variance of the treatment effects and
(b) the effective sample size for variance estimation is increased because SSUs are the experimental units instead of PSUs.

It is often not feasible to apply different treatments within the same PSU from a practical point of view, e.g. if PSUs are households and the treatments concern different approach strategies. At the cost of reduced power, PSUs are randomized over the treatments and consequently the experimental units do not coincide with the ultimate sampling units of the sampling design. In the next section, a design-based approach to analyse this design is developed.

4. Analysis of experiments with clusters of sampling units as experimental units

The purpose of the experiments that were discussed in Sections 2.2, 2.3 and 2.5 is to estimate the target parameters of an on-going survey under the different treatments and to test whether these estimates are significantly different. This implies that, for each subsample, parameter and variance estimators are required that account for

(a) the sampling design of the on-going survey that is used to draw a probability sample from the finite target population,
(b) the experimental design that is used to divide this sample in subsamples and
(c) the estimation procedure of the on-going survey to estimate target parameters.

This gives rise to a design-based Wald statistic to test hypotheses about systematic differences between finite population parameters defined as means, totals and ratios that are observed under alternative survey implementations. In this section, such a design-based analysis procedure is derived for an RBD that is embedded in a two-stage sampling design where the PSUs are the experimental units. Subsequently it is indicated in Section 4.6 how results for other designs that were mentioned above are obtained as a special case, e.g. if clusters of sampling units that are assigned to the same interviewer are randomized over the treatments. van den Brakel and Renssen (2005) discussed in detail why a design-based linear regression analysis is less appropriate in these applications.

4.1. Hypothesis testing

Testing hypotheses about response bias in finite population parameter estimates due to different survey implementations implies the existence of measurement errors. Therefore the traditional notion that observations that are obtained from sampling units are true fixed values observed

without error (e.g. Cochran (1977)) is untenable, and a measurement error model is assumed to link systematic differences between a finite population parameter that is observed under different survey implementations or treatments. Consider a finite population that consists of M PSUs. The jth PSU consists of N_j SSUs. The population size is given by $N=\Sigma_{j=1}^{M}N_j$. Let y_{ijkl} denote the observations for the target parameter that is obtained from an SSU or sampling unit i belonging to PSU j that is assigned to interviewer l and treatment k. It is assumed that the observations for this parameter are a realization of the measurement error model

$$y_{ijkl}=u_{ij}+\beta_k+\gamma_l+\varepsilon_{ijk}. \tag{4.1}$$

Here u_{ij} is the true intrinsic value of sampling unit (i, j), β_k an additive fixed effect of treatment k, γ_l an effect of interviewer l and ε_{ijk} a measurement error of sampling unit (i, j) observed under treatment k. The treatment effects β_k can be interpreted as the bias that is induced by the kth treatment or survey implementation that is used to measure the population parameter. The model allows for interviewer effects, i.e. $\gamma_l=\psi+\xi_l$, where ψ denotes a systematic interviewer bias and ξ_l the random effect of the lth interviewer. Let E_m and cov_m denote the expectation and the covariance with respect to the measurement error model. It is assumed that $E_m(\xi_l)=0$, $\mathrm{var}_m(\xi_l)=\sigma_l^2$ and random interviewer effects between interviewers are independent. Furthermore, it is assumed that $E_m(\varepsilon_{ijk})=0$, $\mathrm{var}_m(\varepsilon_{ijk})=\sigma_{ijk}^2+\sigma_{jk}^2$, the covariance between measurement errors from the same PSU is $\mathrm{cov}_m(\varepsilon_{ijk},\varepsilon_{i'jk})=\sigma_{jk}^2$ and measurement errors between different PSUs are independent. Hence

$$E_m(y_{ijkl})=u_{ij}+\beta_k+\psi,$$

and

$$\mathrm{cov}_m(y_{ijkl},y_{i'j'k'l'})=\begin{cases}\sigma_{ijk}^2+\sigma_{jk}^2+\sigma_l^2 & i=i',j=j',l=l',\\ \sigma_{jk}^2+\sigma_l^2 & i\neq i',j=j',l=l',\\ \sigma_l^2 & i\neq i',j\neq j',l=l',\\ 0 & i\neq i',j\neq j',l\neq l'.\end{cases}$$

Note that the measurement errors of each interviewer might have a separate variance. Separate variances are also allowed for the measurement errors of the different PSUs and SSUs under the different treatments. The measurement error model allows for correlated responses between different sampling units that are assigned to the same interviewer. The measurement error model also allows for correlated responses between sampling units that belong to the same PSU. Such correlation arises for example if PSUs correspond to households and proxy interviewing is allowed by other members of the same household for selected people who cannot be contacted, which is for example the case in the LFS examples; see Section 2.1.

Let $\bar{Y}_k$ denote the population mean of a target parameter that is observed under treatment $k=1,\ldots,K$. Under a complete enumeration of the population under treatment k, the population mean is given by $\bar{Y}_k=\bar{u}+\beta_k+\bar{\gamma}+\bar{\varepsilon}\approx\bar{u}+\beta_k+\psi$, where $\bar{u}$, $\bar{\gamma}$ and $\bar{\varepsilon}$ are the population means of the intrinsic values, interviewer effects and measurement errors in the finite population. Then $\bar{\mathbf{Y}}=(\bar{Y}_1,\ldots,\bar{Y}_K)^{\mathrm{T}}$ denotes the K-dimensional vector with population means observed under the different treatments of the experiment.

A linear measurement error model, like model (4.1), is appropriate for quantitative variables. In the LFS examples, however, the target variables are binary. In these applications, the observations y_{ijkl} are indicators taking values 1 if the sampling unit (i, j) under the kth treatment reports being unemployed and 0 otherwise. The intrinsic variables u_{ij} might also be considered as binary variables taking values 1 if the sampling unit is unemployed and 0 otherwise. In this case, the

population mean $\bar{u}$ denotes the real fraction of unemployed people in the finite population. The treatment effects β_k can be interpreted as the average effect at this fraction if this finite population parameter is measured under the kth treatment. It might be more appealing to interpret the intrinsic variables u_{ij} as the probability that the response of the sampling unit equals 1. The real population parameter $\bar{u}$ still denotes the real fraction of unemployed people in the finite population and the treatment effects β_k can be interpreted as the average effect at the probability that the sampling units' response under the kth treatment equals 1. The interviewer effects can be interpreted in an analogous way. This approach appears to be rigid, since logistic models are more natural in the case of binary response variables. The linear model, however, is required to develop a design-based analysis that accounts for the generalized regression estimation that is used in the LFS to estimate figures about the labour market (see Section 2.1). Furthermore, the linear measurement error model (4.1) is very appropriate to link systematic differences between a finite population parameter that is observed under different survey implementations, i.e. the K different values for $\bar{Y}_k$, and the real population value $\bar{u}$.

The purpose of the experiment is to test the hypothesis that the population means that are observed under the different treatments are equal against the alternative that at least one pair is significantly different. Only systematic differences between the treatments, which are reflected by β_k, should lead to a rejection of the null hypothesis. Random deviations due to measurement errors and interviewer effects should not lead to significant differences in the analysis. This is accomplished by formulating hypotheses about $\bar{\mathbf{Y}}$ in expectation over the measurement error model, i.e.

$$\begin{aligned} H_0&: \mathbf{C}\, E_m(\bar{\mathbf{Y}}) = \mathbf{0}, \\ H_1&: \mathbf{C}\, E_m(\bar{\mathbf{Y}}) \neq \mathbf{0}. \end{aligned} \tag{4.2}$$

Here $\mathbf{C} = (\mathbf{j}|-\mathbf{I})$ denotes a $(K-1) \times K$ contrast matrix, where $\mathbf{j}$ denotes a $(K-1)$-vector with each element equal to 1 and $\mathbf{I}$ a $(K-1) \times (K-1)$ identity matrix. The contrasts between the population parameters in hypothesis (4.2) exactly correspond to the contrasts between the treatment effects β_k that are represented by measurement error model (4.1). Hypothesis (4.2) can be tested by estimating $\bar{\mathbf{Y}}$, where we account for the sampling design, the experimental design and the weighting procedure of the regular sample survey. If $\hat{\bar{\mathbf{Y}}}$ denotes such a design-unbiased estimator and $\mathbf{V}(\mathbf{C}\hat{\bar{\mathbf{Y}}})$ the covariance matrix of the contrasts between $\hat{\bar{\mathbf{Y}}}$, then hypothesis (4.2) can be tested with the Wald statistic $W = \hat{\bar{\mathbf{Y}}}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\{\mathbf{V}(\mathbf{C}\hat{\bar{\mathbf{Y}}})\}^{-1}\mathbf{C}\hat{\bar{\mathbf{Y}}}$. Parameter and variance estimators for this Wald test are worked out in the next sections.

4.2. *Parameter estimation*

In this section a design-based estimator for the population parameters that are observed under the K different treatments of the experiment is developed. Therefore, the inclusion probabilities for the sampling units must be derived first. These are the probabilities that a sampling unit is selected in the initial probability sample that is drawn from the finite population, and subsequently is assigned to one of the K subsamples according to the experimental design. Then the Horvitz–Thompson estimator can be applied to obtain design-unbiased estimates for the population parameter that is observed under the various treatments; see for example Särndal *et al.* (1992), section 2.8, for an introduction and properties of this estimator. Finally the generalized regression estimator is developed to obtain approximately design-unbiased estimates. This estimator is widely applied in survey sampling to improve the accuracy of the Horvitz–Thompson estimator by using *a priori* knowledge about the finite population. See Särndal *et al.* (1992), chapter 6, for an introduction to the generalized regression estimator. For

notational convenience the subscript l will be omitted in y_{ijkl} if possible, since there is no need to sum explicitly over the interviewer subscript in the estimators that are developed in this section.

To test hypothesis (4.2), a two-stage sample s, which is drawn from the finite target population, is available. Let π_j^{I} denote the first-order inclusion probability of the jth PSU in the first stage of the sampling design and $\pi_{i|j}^{\mathrm{II}}$ the first-order inclusion probability of the ith SSU in the second stage given the realization of the first-stage sample. In a CRD, the sample of PSUs is randomized over the K treatments. Let m_k denote the number of PSUs that are assigned to subsample s_k. Then $m_+ = \Sigma_{k=1}^{K} m_k$ denotes the total number of PSUs in s. The conditional probability that PSU j is assigned to treatment k, given the realization of the first stage, equals m_k/m_+. In the case of an RBD, the PSUs are deterministically divided into B blocks s_b, $b=1,\ldots,B$. The PSUs within each block are randomized over the K treatments. Interviewers or strata of the first-stage design are potential block variables in this situation. Let m_{bk} denote the number of PSUs that are assigned to treatment k in block b. Then $m_{b+} = \Sigma_{k=1}^{K} m_{bk}$ denotes the number of PSUs in block b, $m_{+k} = \Sigma_{b=1}^{B} m_{bk}$ the number of PSUs in subsample s_k and $m_{++} = \Sigma_{b=1}^{B}\Sigma_{k=1}^{K} m_{bk}$ the total number of PSUs in s. The conditional probability that PSU j is assigned to treatment k, given the realization of the first stage and that PSU $j \in s_b$, equals m_{bk}/m_{b+}. Each subsample s_k can be considered as a two-phase sample, where the first phase corresponds to the sampling design that is used to draw sample s and the second phase corresponds to the experimental design that is used to divide s into K subsamples s_k. Consequently it follows that the first-order inclusion probability of the jth PSU in the first stage of s_k equals $\pi_j^{*\mathrm{I}} = (m_k/m_+)\pi_j^{\mathrm{I}}$ in a CRD or $\pi_j^{*\mathrm{I}} = (m_{bk}/m_{b+})\pi_j^{\mathrm{I}}$ in an RBD. The first-order inclusion probability of the ith SSU in subsample s_k is given by $\pi_j^{*\mathrm{I}}\pi_{i|j}^{\mathrm{II}}$.

A design-unbiased estimator for $\bar{Y}_k$ that is based on the observations obtained in s_k is given by the π-estimator or Horvitz–Thompson estimator, which was developed by Narain (1951), and Horvitz and Thompson (1952) for unequal probability sampling from finite populations without replacement:

$$\hat{\bar{Y}}_k = \frac{1}{N}\sum_{j\in s_k}\sum_{i=1}^{n_j}\frac{y_{ijk}}{\pi_j^{*\mathrm{I}}\pi_{i|j}^{\mathrm{II}}} \equiv \frac{1}{N}\sum_{j\in s_k}\frac{\hat{y}_{jk}}{\pi_j^{*\mathrm{I}}}. \tag{4.3}$$

Here n_j denotes the number of SSUs that are drawn from PSU j in the second stage and $\hat{y}_{jk}$ the Horvitz–Thompson estimator for the population total of the jth PSU assigned to the kth treatment.

In many sample surveys, including the Dutch LFS, the generalized regression estimator is used to calibrate the sample weights to a set of auxiliary variables for which the population totals are known. The analysis procedure for embedded experiments should therefore be based on the generalized regression estimator. This has the additional advantage that it makes the analysis more accurate since the generalized regression estimator generally reduces the design variance of the Horvitz–Thompson estimator and corrects, at least partially, for selective non-response. Let $\mathbf{x}_{ij} = (x_{ij1},\ldots,x_{ijH})^{\mathrm{T}}$ denote a vector containing H auxiliary variables x_{ijh} of sampling unit (i,j). It is assumed that these auxiliary variables are intrinsic variables that are observed without measurement errors and are not affected by the treatments. According to the model-assisted approach of Särndal *et al.* (1992), the intrinsic values u_{ij} in the measurement error model for each unit in the population are assumed to be an independent realization of the linear regression model

$$u_{ij} = \mathbf{b}^{\mathrm{T}}\mathbf{x}_{ij} + e_{ij}, \tag{4.4}$$

where $\mathbf{b}$ denotes an H-vector with regression coefficients and e_{ij} the residuals of the regression model. Let ω_{ij}^2 denote the variance of e_{ij}. It is assumed that all ω_{ij}^2 are known up to a common scale factor, i.e. $\omega_{ij}^2 = \nu_{ij}\omega^2$, with ν_{ij} known. The generalized regression estimator for $\bar{Y}_k$ based on the observations in s_k is given by (Särndal *et al.*, 1992)

$$\hat{\bar{Y}}_{k;\text{greg}} = \hat{\bar{Y}}_k + \hat{\mathbf{b}}_k^{\mathrm{T}}(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}_k). \tag{4.5}$$

Here $\bar{\mathbf{X}}$ denotes an H-dimensional vector, containing the known population means of the auxiliary variables, and $\hat{\bar{\mathbf{X}}}_k$ the Horvitz–Thompson estimator for $\bar{\mathbf{X}}$, which is defined analogously to equation (4.3) where y_{ijk} is replaced by $\mathbf{x}_{ij}$. Finally, $\hat{\mathbf{b}}_k$ is a Horvitz–Thompson-type estimator for the regression coefficients in the regression function of y_{ijk} on $\mathbf{x}_{ij}$ based on the observations in s_k. For an expression for $\hat{\mathbf{b}}_k$ see Särndal *et al.* (1992), section 6.4, formula (6.4.13), or van den Brakel (2001), section 7.3, formula (7.21), for an expression in this specific context. Since equation (4.5) is non-linear, a linearized approximation, which is obtained by means of a Taylor series expansion about their real values in the finite population that is truncated at the first-order term, can be shown to be design unbiased. Therefore the generalized regression estimator $\hat{\bar{\mathbf{Y}}}_{\text{greg}} = (\hat{\bar{Y}}_{1;\text{greg}}, \ldots, \hat{\bar{Y}}_{K;\text{greg}})^{\mathrm{T}}$ is an approximately design-unbiased estimator for $\bar{\mathbf{Y}}$ and $E_m(\bar{\mathbf{Y}})$.

4.3. Variance estimation

The next step is the derivation of a design-based estimator for the covariance matrix of the contrasts between $\hat{\bar{\mathbf{Y}}}_{\text{greg}}$. The subsample estimates are correlated since the subsamples are drawn without replacement from a finite population. A design-based estimator for this covariance matrix requires that for each sampling unit an observation under each of the K treatments is obtained. These paired observations are, however, not available since the sampling units are assigned to one of the K treatments only. This problem can also be stated in more technical terms by noting that an estimator for the design covariances requires joint inclusion probabilities for the sampling units (i, j) and (i', j') that are assigned to treatments k and k'. The joint inclusion probability that a sampling unit is assigned to two different treatments, i.e. $i = i'$, $j = j'$ and $k \neq k'$, equals 0. This hampers a direct estimation of the design covariance matrix of $\hat{\bar{\mathbf{Y}}}_{\text{greg}}$.

To test hypothesis (4.2) it is, however, sufficient to have a design-based estimator for the covariance matrix of the $K-1$ contrasts between $\hat{\bar{\mathbf{Y}}}_{\text{greg}}$. Under measurement error model (4.1) and a weighting model for the generalized regression estimator that at least uses the size of the finite population as auxiliary information, it follows that an approximately design-unbiased estimator for the covariance matrix of the contrasts of $\hat{\bar{\mathbf{Y}}}_{\text{greg}}$ is given by $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$ where $\hat{\mathbf{D}}$ is a diagonal matrix. In the case of an RBD, the diagonal elements of $\hat{\mathbf{D}}$ are given by

$$\hat{d}_k = \frac{1}{N^2}\sum_{b=1}^{B}\frac{1}{m_{bk}}\frac{1}{m_{bk}-1}\sum_{j\in s_{bk}}\left(\frac{m_{b+}\hat{e}_{jk}}{\pi_j^{\mathrm{I}}} - \frac{1}{m_{bk}}\sum_{j'\in s_{bk}}\frac{m_{b+}\hat{e}_{j'k}}{\pi_{j'}^{\mathrm{I}}}\right)^2, \tag{4.6}$$

with

$$\hat{e}_{jk} = \sum_{i=1}^{n_j}\frac{y_{ijk} - \hat{\mathbf{b}}_k^{\mathrm{T}}\mathbf{x}_{ij}}{\pi_{i|j}^{\mathrm{II}}}.$$

An outline of the proof of this result is given in Appendix A. An expression for the diagonal elements $\hat{d}_k$ under a CRD follows as a special case from equation (4.6) by taking $B = 1$, $m_{bk} = m_k$ and $m_{b+} = m_+$.

If equation (4.6) is compared with formula (4.5.3) of Särndal *et al.* (1992), then it can be recognized that $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$ has the structure as if the K subsamples were drawn independently from each other, where the PSUs are selected with replacement, using unequal probabilities π_j^{I}/m_+ in the case of a CRD and $\pi_j^{\mathrm{I}}/m_{b+}$ in the case of an RBD. In survey sampling this variance estimator is used to approximate the variance under complex multistage sampling designs (Särndal *et al.* (1992), section 4.6). For the embedded RBDs and CRDs, this estimator is design unbiased for the variance of the contrasts between two subsample estimates. No joint inclusion probabilities and no covariances are required in this variance estimator. This is the result of the superimposition of the experimental design on the sampling design in combination with the fact that we focus on the variances about the contrasts between subsample estimates, a weighting scheme is used that meets the condition that there is a constant vector $\mathbf{a}$ of order H, such that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij} = 1$ for all units (i, j) in the population and the assumption that measurement errors between PSUs are independent.

Owing to the superimposition of the experimental design on the sampling design, the randomization mechanism of the experimental design dominates the variance structure of the $K-1$ contrasts between $\hat{\hat{\mathbf{Y}}}_{\mathrm{greg}}$. The randomization mechanism of an RBD can be considered as the selection of K subsamples by means of stratified simple random sampling without replacement where the PSUs are the sampling units and the blocks of the experimental design are the strata. In a similar way, a CRD can be considered as selecting K subsamples from the initial sample by means of simple random sampling. In the variance of the contrasts under (stratified) simple random sampling, the finite population correction of the subsample means cancels out against the covariance between these subsample means.

The covariance matrix of the contrasts between the K generalized regression estimators is a function of differences between the residuals of the generalized regression estimator, i.e. $y_{ijkl} - \mathbf{b}_k^{\mathrm{T}}\mathbf{x}_{ij} - (y_{ijk'l} - \mathbf{b}_{k'}^{\mathrm{T}}\mathbf{x}_{ij})$. If the weighting scheme of the generalized regression estimator meets the condition that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij} = 1$ for all units (i, j) in the population, then the treatment effect β_k cancels out in these residuals. The stated condition that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij} = 1$ implies that in the weighting scheme of the generalized regression estimator at least the population size N is used as *a priori* knowledge. As a result the residuals $y_{ijkl} - \mathbf{b}_k^{\mathrm{T}}\mathbf{x}_{ij}$ are composed of the residual of the linear regression model of the intrinsic value e_{ij} from expression (4.4), a term concerning the bias that is induced by the interviewer effect (this is the residual of the regression function of the interviewer bias ψ on $\mathbf{x}_{ij}$) and the measurement error ε_{ijk} from expression (4.1). In the differences between the residuals $y_{ijkl} - \mathbf{b}_k^{\mathrm{T}}\mathbf{x}_{ij}$, the residuals of the linear regression model of the intrinsic values e_{ij} and the residuals concerning the interviewer bias cancel out since they are equal under the different treatments. As a result only the measurement errors ε_{ijk} remain in the expression for the covariance matrix of the contrasts between the generalized regression estimators; see formula (A.4) in Appendix A. It is assumed that the measurement errors between the PSUs are independent. The result of these factors is that $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$ has a structure as if the K subsamples were drawn independently from each other; see also van den Brakel and Renssen (2005) for a more technical and detailed discussion.

The stated condition that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij} = 1$ for all (i, j) in the population is not very restrictive, since it implies that at least the size of the finite population is used as auxiliary information in the generalized regression estimator. This holds for weighting models that contain an intercept or at least one categorical variable that partitions the population in subpopulations and uses these subpopulation sizes as *a priori* knowledge. This condition, however, does not hold for the ratio estimator, since the ratio model contains only a single real-valued auxiliary variable; see Särndal *et al.* (1992), section 7.3. Under this weighting model the variance estimator proposed is design unbiased under the null hypothesis of no treatment effects but not under the alternative.

The minimum use of auxiliary information in the generalized regression estimator is a weighting scheme where $(x_{ij}) = 1$ and $\omega_{ij}^2 = \omega^2$. This weighting scheme is known as the common mean model and uses only the size of the finite population N as *a priori* knowledge (Särndal *et al.* (1992), section 7.4). Under this model the generalized regression estimator simplifies to

$$\hat{\tilde{Y}}_{k;\text{greg}} = \left(\sum_{j \in s_k} \sum_{i=1}^{n_j} \frac{1}{\pi_j^{*\text{I}} \pi_{i|j}^{\text{II}}} \right)^{-1} \sum_{j \in s_k} \sum_{i=1}^{n_j} \frac{y_{ijk}}{\pi_j^{*\text{I}} \pi_{i|j}^{\text{II}}} = \frac{1}{\hat{N}} \sum_{j \in s_k} \sum_{i=1}^{n_j} \frac{y_{ijk}}{\pi_j^{*\text{I}} \pi_{i|j}^{\text{II}}} \equiv \tilde{Y}_k, \tag{4.7}$$

which can be recognized as the ratio estimator for a population mean, which was originally proposed by Hájek (1971). It also follows that $\hat{\mathbf{b}}_k = \tilde{Y}_k$. An approximately design-unbiased estimator for the covariance matrix of the contrasts between the subsample estimates is given by equation (4.6), where $\hat{e}_{jk} = \hat{y}_{jk} - \tilde{Y}_k \hat{N}_j$, with $\hat{N}_j = \sum_{i=1}^{n_j} 1/\pi_{i|j}^{\text{II}}$. For situations where the sum over the design weights equals the population total, i.e. $\hat{N} = N$, it follows that equation (4.7) simplifies to the basic Horvitz–Thompson estimator that was defined in expression (4.3). Estimator (4.7) generally differs from and usually performs better than the Horvitz–Thompson estimator (4.3), since estimator (4.7) avoids the extreme estimates that are sometimes obtained with the Horvitz–Thompson estimator; Särndal *et al.* (1992), section 7.4. Unlike the Horvitz–Thompson estimator (4.3), the Hájek estimator (4.7) meets the stated condition that $\mathbf{a}^{\text{T}} \mathbf{x}_{ij} = 1$ for all units in the population which is used to derive variance estimator (4.6). Variance expressions for the Horvitz–Thompson estimator are more complicated if $\hat{N} \neq N$; see van den Brakel (2001), chapter 3.

Particularly if the number of experimental units within each block is small, the variance estimation procedure might be improved by pooling the variance estimators for the separate subsamples,

$$\hat{d}_{k,p} = \frac{1}{N^2} \sum_{b=1}^{B} \frac{1}{m_{bk}} \frac{1}{m_{b+} - K} \sum_{k'=1}^{K} \sum_{j \in s_{bk'}} \left(\frac{m_{b+} \hat{e}_{jk'}}{\pi_j^{\text{I}}} - \frac{1}{m_{bk'}} \sum_{j' \in s_{bk'}} \frac{m_{b+} \hat{e}_{j'k'}}{\pi_{j'}^{\text{I}}} \right)^2. \tag{4.8}$$

With this pooled variance estimator it is assumed that the measurement errors of the PSUs and SSUs under the different treatments have equal variances, i.e. $\sigma_{jk}^2 = \sigma_{j'k'}^2 = \sigma_{\text{I}}^2$ and $\sigma_{ijk}^2 = \sigma_{i'j'k'}^2 = \sigma_{\text{II}}^2$.

4.4. Wald test

To test hypothesis (4.2), the subsample estimates and the covariance matrix of the contrasts between the subsample estimates give rise to the following design-based Wald statistic:

$$W = \hat{\tilde{\mathbf{Y}}}_{\text{greg}}^{\text{T}} \mathbf{C}^{\text{T}} (\mathbf{C} \hat{\mathbf{D}} \mathbf{C}^{\text{T}})^{-1} \mathbf{C} \hat{\tilde{\mathbf{Y}}}_{\text{greg}}.$$

Owing to the diagonal structure of $\hat{\mathbf{D}}$ this Wald statistic can be simplified to (van den Brakel and Renssen, 2005)

$$W = \sum_{k=1}^{K} \frac{\hat{\tilde{Y}}_{k;\text{greg}}^2}{\hat{d}_k} - \left(\sum_{k=1}^{K} \frac{1}{\hat{d}_k} \right)^{-1} \left(\sum_{k=1}^{K} \frac{\hat{\tilde{Y}}_{k;\text{greg}}}{\hat{d}_k} \right)^2. \tag{4.9}$$

To calculate p-values or critical regions for W, it is usually conjectured for generally complex sampling schemes that, under the null hypothesis, W is asymptotically χ^2 distributed with $K - 1$ degrees of freedom. See van den Brakel and Renssen (2005) for a more detailed discussion about the limit distribution of statistic (4.9). The simulation results that are discussed in Section 5 also confirm this conjecture.

4.5. Analysis of ratios

In the LFS examples the main target parameters are defined as the ratio of two population totals. The unemployed labour force, for example, is defined as the total unemployment divided by the total labour force. Therefore the design-based analysis procedure that was developed in the preceding sections for population means and totals is now extended to ratios.

Let $R_k = \bar{Y}_k / \bar{Z}_k$ denote the ratio of two population means that are observed under treatment $k = 1, \ldots, K$. Then $\mathbf{R} = (R_1, \ldots, R_K)^{\mathrm{T}}$ denotes the K-dimensional vector with ratios observed under the different treatments of the experiment. The hypothesis of no treatment effects for ratios can be tested with the Wald statistic $W = \hat{\mathbf{R}}^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \{\mathbf{V}(\mathbf{C}\hat{\mathbf{R}})\}^{-1} \mathbf{C}\hat{\mathbf{R}}$, where $\hat{\mathbf{R}}$ denotes a design-based estimator for $\mathbf{R}$. Analogously to expression (4.1) hypotheses are formulated about the ratios where the numerator and the denominator both denote the population total in expectation over the measurement error model.

Let y_{ijkl} denote the observations for the parameter in the numerator and z_{ijkl} the observations for the parameter in the denominator for sampling units (i, j) assigned to the kth treatment and the lth interviewer. It is assumed that the observations z_{ijkl} are a realization of the same type of measurement error model as defined for y_{ijkl} in expression (4.1). The generalized regression estimator for $\bar{Z}_k$ that is based on the observations that are obtained in subsample s_k is defined in a similar way as $\hat{\bar{Y}}_{k;\mathrm{greg}}$ in expression (4.5). The generalized regression estimator for R_k is given by

$$\hat{R}_{k;\mathrm{greg}} = \hat{\bar{Y}}_{k;\mathrm{greg}} / \hat{\bar{Z}}_{k;\mathrm{greg}}. \tag{4.10}$$

Finally $\hat{\mathbf{R}}_{\mathrm{greg}} = (\hat{R}_{1;\mathrm{greg}}, \ldots, \hat{R}_{K;\mathrm{greg}})^{\mathrm{T}}$ denotes the generalized regression estimator for $\mathbf{R}$. In Appendix A it is derived that, under the null hypothesis and the condition that the generalized regression estimator at least uses the size of the finite population as auxiliary information, an approximately unbiased estimator for the covariance matrix of the contrasts of $\hat{\mathbf{R}}_{\mathrm{greg}}$ is given by $\mathbf{C}\hat{\mathbf{D}}^{(R)}\mathbf{C}^{\mathrm{T}}$ where $\hat{\mathbf{D}}^{(R)}$ is a diagonal matrix. For an RBD the diagonal elements of $\hat{\mathbf{D}}^{(R)}$ are defined as

$$\hat{d}_k^{(R)} = \frac{1}{N^2 \hat{\bar{Z}}_{k;\mathrm{greg}}^2} \sum_{b=1}^{B} \frac{1}{m_{bk}} \frac{1}{m_{bk} - 1} \sum_{j \in s_{bk}} \left(\frac{m_{b+} \hat{e}_{jk}}{\pi_j^{\mathrm{I}}} - \frac{1}{m_{bk}} \sum_{j' \in s_{bk}} \frac{m_{b+} \hat{e}_{j'k}}{\pi_{j'}^{\mathrm{I}}} \right)^2, \tag{4.11}$$

and

$$\hat{e}_{jk} = \sum_{i=1}^{n_j} \frac{y_{ijk} - \hat{\mathbf{b}}_k^{\mathrm{T}} \mathbf{x}_{ij} - \hat{R}_{k;\mathrm{greg}} (z_{ijk} - \hat{\mathbf{f}}_k^{\mathrm{T}} \mathbf{x}_{ij})}{\pi_{i|j}^{\mathrm{II}}}. \tag{4.12}$$

Here $\hat{\mathbf{f}}_k$ denotes the H-dimensional vector with the Horvitz–Thompson-type estimator for the regression coefficients of the regression function of z_{ijk} on $\mathbf{x}_{ij}$, which is defined in a similar way to $\hat{\mathbf{b}}_k$ in Section 4.2. An expression for $\hat{d}_k^{(R)}$ under a CRD follows as a special case from equations (4.11) and (4.12) with $B = 1$, $m_{bk} = m_k$ and $m_{b+} = m_+$. In Section 4.4 it was emphasized that the variance estimation procedure under the generalized regression estimator with a ratio model is only unbiased under the null hypothesis. Analogously to this property, the estimator for the covariance matrix of the contrasts between ratios of two population totals is unbiased under the null hypothesis, but not under the alternative hypothesis. The property that the covariance matrix of the contrasts between the ratios is a function of measurement errors only holds under the null hypothesis. In Section 4.3 it is explained that this is one of the factors that is used to derive the proposed variance estimators. In Section 5 a simulation is described that is aimed at

investigating the performance of estimator (4.11) as an estimator for the real covariance matrix of the contrasts between ratios. These simulations do not indicate that estimator (4.11) is biased under the alternative hypothesis.

Expressions for the Hájek estimator are obtained in a straightforward manner, i.e. $\tilde{R}_k = \tilde{Y}_k / \tilde{Z}_k$, where $\tilde{Y}_k$ is defined in equation (4.7) and $\tilde{Z}_k$ is defined in a similar way. An approximation for the covariance matrix of the contrasts between the subsample estimates is given by estimator (4.11), where $\hat{e}_{jk} = \hat{y}_{jk} - \tilde{Y}_k \hat{N}_j - \tilde{R}_k(\hat{z}_{jk} - \tilde{Z}_k \hat{N}_j)$, with $\hat{N}_j = \Sigma_{i=1}^{n_j} 1/\pi_{i|j}^{\mathrm{II}}$. The pooled variance estimator (4.8) can be used as an alternative to obtain more stable variance estimates if the numbers of sampling units within the blocks are small. The hypothesis of no treatment effects is tested with Wald statistic (4.9), where $\hat{\tilde{Y}}_{k;\mathrm{greg}}$ and $\hat{d}_k$ are replaced by $\hat{R}_{k;\mathrm{greg}}$ and $\hat{d}_k^{(R)}$.

4.6. *Special cases*

4.6.1. *Randomizing interviewers with their cluster of sampling units over the treatments*

Now consider an experiment where clusters of sampling units that are assigned to the same interviewer are randomized over the treatments. The analysis of this type of experiments can be conducted with the procedure that is proposed in this section by taking $\pi_j^{\mathrm{I}} = 1$ for all j and considering $\pi_{i|j}^{\mathrm{II}} = \pi_i$ as the first-order inclusion probabilities of the sampling design. Furthermore, m_{bk} denotes the number of interviewers in block b who are assigned to treatment k, m_{b+} the number of interviewers in block b, m_{+k} the number of interviewers who are assigned to treatment k and m_{++} the total number of interviewers in the experiment. This result is obtained by conceptually dividing the target population in M subpopulations, with M the number of interviewers who are available for the data collection. Each subpopulation consists of the sampling units that are interviewed by the same interviewer if they are included in the sample. These M subpopulations are included in the first stage of the sample and randomized over the treatments.

4.6.2. *Randomizing the ultimate sampling units over the treatments*

Expressions for the parameter and variance estimates for experiments where the sampling units are randomized over the treatments are obtained by taking $\pi_j^{\mathrm{I}} = 1$ for all j and considering $\pi_{i|j}^{\mathrm{II}} = \pi_i$ as the first-order inclusion probabilities of the sampling design. This result can be derived analogously to the outline of the proof that is given in Appendix A but requires a measurement error model where the measurement errors between the ultimate sampling units are independent, i.e. $\sigma_{jk}^2 = 0$ for all j and k.

4.6.3. *Two-treatment experiments*

A special case of the experiments that are discussed in this section are the two-treatment experiments. These experiments are analysed with a design-based version of the t-test; see van den Brakel and van Berkel (2002). The parameter and variance estimates obtained in this section can be inserted into this design-based t-test, for the analysis of experiments where clusters of ultimate sampling units are randomized over the treatments and to test hypotheses about ratios.

4.6.4. *Hypotheses about population totals*

Wald and t-statistics to test hypotheses about population totals follow in a straightforward manner from the results that were obtained for population means by multiplying the parameter and variance estimates by N and N^2 respectively. The test statistics for population means and totals are equivalent since they are invariant under scale transformations with a constant like the population size.

5. Simulation study

A simulation study is conducted to evaluate the performance of the variance estimator for the contrasts and the Wald statistic that was derived in Section 4. Since the variance estimator for ratios is derived under the null hypothesis, it is particularly interesting to study the behaviour of this variance estimator and the Wald statistic for this type of parameters under alternative hypotheses.

Two artificial populations of different sizes are generated. Both populations contain five strata. PSUs and SSUs are generated by drawing strictly positive values for the intrinsic values $u_{ij}^{(z)}$ and $u_{ij}^{(y)}$ for two parameters Z and Y respectively. The sizes of the PSUs are unequal between and within the strata. The intrinsic values for parameter Z are obtained as follows. First a positive value for each PSU in the population is drawn from a uniform distribution. Subsequently a positive value for each SSU that is drawn from a uniform distribution is added to the value that is obtained for the PSU in the preceding step. This is the intrinsic value $u_{ij}^{(z)}$ for parameter Z of SSU (i, j). Subsequently, a random value from the uniform distribution with interval [0.15–0.80] is drawn for each SSU in the population. The intrinsic values $u_{ij}^{(y)}$ are obtained by multiplying this fraction by the intrinsic values $u_{ij}^{(z)}$. Within each stratum different lower and upper boundaries and interval widths are used for these uniform distributions. As a result both populations can be divided into five relatively homogeneous subpopulations or strata. The intervals of the uniform distributions that are used to generate the values for the SSUs are significantly smaller than the intervals of the uniform distributions that are used to generate the values for the PSUs. As a result, two populations are obtained where the intrinsic values for the SSUs within each PSU are clustered. The randomly generated fractions, which are used to derive the intrinsic values $u_{ij}^{(y)}$, are inversely proportional to the size of the intrinsic value $u_{ij}^{(z)}$. The structure of both populations is given in Tables 2 and 3. Here S denotes the standard deviation between the SSUs of a stratum or the entire population, S_{BP} the standard deviation between the means of the PSUs, and $\bar{S}_{\mathrm{WP}}$ the mean of the standard deviations within PSUs.

A measurement error model without interviewer effects is assumed, i.e.

$$\begin{aligned} z_{ijk} &= u_{ij}^{(z)} + \beta_k^{(z)} + \varepsilon_{ijk}^{(z)}, \\ y_{ijk} &= u_{ij}^{(y)} + \beta_k^{(y)} + \varepsilon_{ijk}^{(y)}. \end{aligned} \tag{5.1}$$

In the simulation study two parameters are used. The first parameter is the population mean of Z (e.g. the population mean of the monthly income). The second parameter is defined as

Table 2. Summary statistics for population 1

Stratum	*Number of PSUs*	*Number of SSUs*	*Intrinsic value parameter Z*				*Intrinsic value parameter Y*			
			Mean	S	S_{BP}	$\bar{S}_{WP}$	*Mean*	S	S_{BP}	$\bar{S}_{WP}$
1	85	3625	44093	13546	13769	1116	9463	2858	2556	1343
2	155	8150	24858	7441	7466	644	6785	1943	1791	751
3	300	15900	8879	2705	2720	202	3362	1002	968	269
4	700	44000	4397	1163	1172	70	2308	560	548	133
5	1100	69000	2223	649	650	25	1484	343	335	67
Total	2340	140675	6046	8725	9897	142	2467	1900	2074	204

Table 3. Summary statistics for population 2

Stratum	*Number of PSUs*	*Number of SSUs*	*Intrinsic value parameter Z*				*Intrinsic value parameter Y*			
			Mean	*S*	S_{BP}	$\bar{S}_{WP}$	*Mean*	*S*	S_{BP}	$\bar{S}_{WP}$
1	255	10875	44151	12890	13008	1136	9421	2737	2424	1296
2	465	24450	24866	7509	7535	645	6789	1962	1816	743
3	900	47700	9323	2877	2890	201	3514	1037	999	281
4	1400	88000	4403	1122	1129	70	2313	542	529	133
5	2200	138000	2233	652	654	25	1491	341	333	67
Total	5220	309025	7211	9941	11139	177	2736	2140	2308	242

the ratio of the total of Y and Z (e.g. the portion of the monthly income spend on primary necessities).

Samples are drawn repeatedly from both populations by means of a stratified two-stage sampling design without replacement. Unequal inclusion probabilities, which are proportional to the size of the target parameters, are applied between and within the strata. The sample sizes for the different strata are summarized in Table 4. For each resample a new measurement error for each population element is generated. Measurement errors for the parameter Z are drawn from a normal distribution with a mean equal to 0 and a standard deviation that is proportional to the size of the intrinsic values. The range of the standard deviations varied from 500 for the SSUs with the largest intrinsic values in the first stratum to 15 for the SSUs with the smallest intrinsic values in the fifth stratum. Subsequently, a random value within the interval [0–0.80] is drawn for each SSU in the population. The measurement error for the second intrinsic value is obtained by multiplying this fraction by the measurement error that is obtained for the first parameter. Observations for the target parameters are obtained by adding a measurement error and a treatment effect to the intrinsic value according to model (5.1).

Finally the samples are randomly divided into three subsamples. Within each stratum a third of the PSUs with their cluster of SSUs are randomly assigned to three different treatments. This

Table 4. Sample and subsample sizes

Stratum	*Samples from population 1*				*Samples from population 2*			
	Number of PSUs		*Number of SSUs*		*Number of PSUs*		*Number of SSUs*	
	Sample	*Subsample*	*Sample*	*Subsample*	*Sample*	*Subsample*	*Sample*	*Subsample*
1	66	22	1188	396	198	66	3564	1188
2	102	34	1530	510	306	102	4590	1530
3	186	62	2232	744	558	186	6696	2232
4	366	122	4392	1464	732	244	8784	2928
5	519	173	6228	2076	1038	346	12456	4152
Total	1239	413	15570	5190	2832	944	36090	12030

Table 5. Simulation settings

Simulation	*Parameter*	*Treatment effects*					
		$\beta_1^{(z)}$	$\beta_2^{(z)}$	$\beta_3^{(z)}$	$\beta_1^{(y)}$	$\beta_2^{(y)}$	$\beta_3^{(y)}$
1	Mean of Z	0	0	0	—	—	—
2	Mean of Z	0	100	200	—	—	—
3	Mean of Z	0	200	400	—	—	—
4	Ratio Y/Z	0	0	0	0	0	0
5	Ratio Y/Z	0	0	0	0	10	20
6	Ratio Y/Z	0	0	0	0	20	40
7	Ratio Y/Z	0	0	0	0	40	80
8	Ratio Y/Z	0	30	60	0	0	0
9	Ratio Y/Z	0	60	120	0	0	0
10	Ratio Y/Z	0	120	240	0	0	0

resulted in an RBD with strata as the block variable and PSUs are the experimental units. As a result the effective sample size is the number of PSUs that are assigned to the subsamples as summarized in Table 4.

The data that are obtained in each resample are analysed with the Hájek estimator that is defined by equation (4.7) for the mean of Z and for the ratio of Y and Z. Formula (4.7) is the ratio estimator for a population mean, which is a generalized regression estimator with a minimum use of auxiliary information, namely the size of the target population. Generalized regression estimators with more extensive weighting models will generally have smaller design variances, but they share the same statistical properties, like the approximate design unbiasedness of the point and variance estimates. The simulation results that were obtained with formula (4.7) are therefore representative for generalized regression estimators with more extensive weighting models.

For both populations 10 sets of treatment effects are applied, which are specified in Table 5. Each simulation is based on $R = 80\,000$ resamples.

Let $\hat{Q}_k^r$ denote the subsample estimate that is obtained under the kth treatment and the rth resample for the mean of Z and the ratio of Y and Z. The vector with three subsample estimates that is obtained in the rth resample is denoted by $\hat{\mathbf{Q}}^r = (\hat{Q}_1^r, \hat{Q}_2^r, \hat{Q}_3^r)^{\mathrm{T}}$. The vector with the two contrasts in the rth resample equals $\mathbf{C}\hat{\mathbf{Q}}^r$, with $\mathbf{C} = (\mathbf{j}|-\mathbf{I})$, $\mathbf{j} = (1, 1)^{\mathrm{T}}$ and $\mathbf{I}$ a 2×2 identity matrix. Moreover, $\hat{d}_k^r$ denotes the diagonal elements of the estimated covariance matrix for the rth resample, which is obtained with expression (4.6) for the mean of Z and equation (4.11) for the ratio of Y and Z. The estimated covariance matrix of the treatment effects in the rth resample equals $\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}^{\mathrm{T}}$, with $\hat{\mathbf{D}}^r = \mathrm{diag}(\hat{d}_1^r, \hat{d}_2^r, \hat{d}_3^r)$. The Wald statistic that is obtained in the rth resample is denoted as $W^r = (\mathbf{C}\hat{\mathbf{Q}}^r)^{\mathrm{T}}(\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\mathbf{Q}}^r)$. On the basis of $R = 80\,000$ resamples the population parameters can be approximated by the simulation mean

$$\bar{\mathbf{Q}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\mathbf{Q}}^r, \tag{5.2}$$

with $\bar{\mathbf{Q}} = (\bar{Q}_1, \bar{Q}_2, \bar{Q}_3)^{\mathrm{T}}$. The treatment effects in the population can be approximated as $\mathbf{C}\bar{\mathbf{Q}}$. The mean of the estimated covariance matrices is defined as

$$\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^{\mathrm{T}} = \frac{1}{R}\sum_{r=1}^{R}\mathbf{C}\hat{\mathbf{D}}^r\mathbf{C}^{\mathrm{T}}. \tag{5.3}$$

An approximation of the real covariance matrix of the treatment effects is obtained with

$$\mathbf{CVC}^{\mathrm{T}} = \frac{1}{R-1}\sum_{r=1}^{R}\mathbf{C}(\hat{\mathbf{Q}}^r - \bar{\mathbf{Q}})(\hat{\mathbf{Q}}^r - \bar{\mathbf{Q}})^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}. \tag{5.4}$$

If the variance estimator $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$, which was derived in Section 4, is approximately design unbiased, then $\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$ must tend to the real covariance matrix $\mathbf{CVC}^{\mathrm{T}}$, for $R \to \infty$. If $\mathbf{C}\hat{\mathbf{Q}} \to N(\mathbf{CQ}, \mathbf{CVC}^{\mathrm{T}})$, then it follows that $W \to \chi^2_{[K-1],[\delta]}$, where $\chi^2_{[K-1],[\delta]}$ denotes a χ^2-distributed random variable with $K-1$ degrees of freedom and non-centrality parameter

$$\delta = \tfrac{1}{2}(\mathbf{CQ})^{\mathrm{T}}(\mathbf{CVC}^{\mathrm{T}})^{-1}(\mathbf{CQ});$$

Searle (1971), chapter 2, theorem 2. The non-centrality parameter is calculated by inserting equation (5.4) in the expression for δ. The power of the Wald test for a set of treatment effects can be calculated by

$$P(W) = P(\chi^2_{[K-1][\delta]} > \chi^2_{[1-\alpha][K-1][0]}), \tag{5.5}$$

where $\chi^2_{[1-\alpha][K-1][0]}$ denotes the $(1-\alpha)$th percentile point of the central χ^2-distribution with $K-1$ degrees of freedom. The performance of the Wald statistic is evaluated by comparing $P(W)$ with the simulated power, which is defined as the fraction of significant runs that are observed in the R resamples, i.e.

$$P^{\mathrm{sim}}(W) = \frac{1}{R}\sum_{r=1}^{R} I(W^r > \chi^2_{[1-\alpha][K-1][0]}), \tag{5.6}$$

where $I(B)$ denotes the indicator variable which is equal to 1 if B is true and 0 otherwise. The mean and the variance of the resample Wald statistics are defined as

$$E(W) = \frac{1}{R}\sum_{r=1}^{R} W^r, \tag{5.7}$$

$$S(W) = \frac{1}{R-1}\sum_{r=1}^{R}(W^r - \bar{W})^2. \tag{5.8}$$

The expected value and the variance of the χ^2-distribution are equal to

$$E(\chi^2_{[K-1][\delta]}) = K - 1 + 2\delta, \tag{5.9}$$

and

$$\mathrm{var}(\chi^2_{[K-1][\delta]}) = 2(K-1) + 8\delta \tag{5.10}$$

respectively (Searle (1971), section 2.4.h). If the resample distribution of W tends to a $\chi^2_{[K-1],[\delta]}$-distribution, then the mean (5.7) must tend to expression (5.9) and the variance (5.8) must tend to expression (5.10).

The simulation results are summarized in Tables 6–8. Results for the point estimates are not presented. We note, however, that the simulation means of the estimated population means (5.2) correspond exactly to the real population values. The simulation means of the estimated ratios are slightly smaller than their real population values, which might be expected since the ratio of two sample estimates is a biased estimate for the ratio of the two population parameters.

Table 6 summarizes the simulation results for the variance estimation procedure. The real covariance matrices $\mathbf{CVC}^{\mathrm{T}}$ that are obtained with equation (5.4) are compared with the mean of the estimated covariance matrices $\mathbf{C}\bar{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$ that are obtained with equation (5.3). The simulation

Table 6. Simulation results for the variance estimation procedure†

Simulation	*Population*	*Real covariance matrix* $\mathbf{CVC}^T$			*Mean estimated covariance matrix* $\mathbf{C\bar{D}C}^T$		
		(1,1)	*(2,2)*	*(1,2)*	*(1,1)*	*(2,2)*	*(1,2)*
1	1	25468	25402	12655	25482	25486	12740
2	1	25376	25356	12615	25490	25489	12741
3	1	25531	25383	12716	25473	25493	12735
4	1	0.25119	0.25174	0.12619	0.25026	0.25032	0.12511
5	1	0.25209	0.25532	0.12578	0.25191	0.25378	0.12514
6	1	0.25256	0.25791	0.12509	0.25374	0.25720	0.12516
7	1	0.25793	0.26480	0.12629	0.25697	0.26434	0.12510
8	1	0.24828	0.24496	0.12635	0.24696	0.24398	0.12508
9	1	0.24461	0.23953	0.12615	0.24383	0.23803	0.12507
10	1	0.23862	0.22764	0.12605	0.23788	0.22727	0.12510
1	2	15099	15023	7566	15023	15024	7511
2	2	14969	15028	7508	15026	15025	7513
3	2	14985	14984	7528	15024	15022	7509
4	2	0.08040	0.08084	0.04040	0.08080	0.08082	0.04040
5	2	0.08068	0.08168	0.04007	0.08131	0.08182	0.04041
6	2	0.08139	0.08263	0.04019	0.08182	0.08284	0.04041
7	2	0.08237	0.08503	0.04031	0.08282	0.08494	0.04038
8	2	0.07927	0.07864	0.03986	0.07991	0.07906	0.04038
9	2	0.07866	0.07720	0.04041	0.07908	0.07742	0.04041
10	2	0.07678	0.07434	0.04010	0.07740	0.07437	0.04038

†Simulation numbers refer to the simulation settings that are summarized in Table 5.

number refers to the 10 different simulation settings that are summarized in Table 5, which are applied to the two populations. With 80000 resamples the accuracy of the approximation of the real covariance matrix (5.4) is about 1%; Knottnerus (2002), section 10.5. Since the differences between the elements of the approximate real covariance matrix (5.4) and the mean of the estimated covariance matrix (5.3) are much smaller than 1%, there are no indications that the variance estimators for the contrasts are biased for both the means and the ratios. The differences under the alternative hypotheses are of the same size as under the null hypotheses. This implies that there is no indication in this simulation study that the variance estimators for the ratios under the alternative hypothesis are biased.

In Table 7 the simulated power that is obtained with equation (5.6) is compared with the real power that is obtained with equation (5.5). The simulated first and second moments (5.7) and (5.8) are compared with their real values based on equations (5.9) and (5.10) in Table 8.

The simulated power corresponds reasonably well to the real power for the simulations with the means for both populations and for the simulations with the ratios for population 2. For simulations with ratios for population 1, the simulated power is slightly larger than the real power. The simulated first and second moments (5.7) and (5.8) are also larger than their real values based on equations (5.9) and (5.10) for the simulations with ratios for population 1. This implies that the distribution of the Wald statistic has shifted to the right, which explains why the simulated power is slightly larger compared with the real power. Probably the Wald statistic for the ratios converges slower to a χ^2-distribution than the Wald statistic for the means.

6. Software

The analysis procedures proposed in Section 4 and in van den Brakel and Renssen (1998,

Table 7. Simulation results for the power of the Wald statistic†

Simulation	*Population*	*Real power* $P(W)$			*Simulated power* $P^{\text{sim}}(W)$		
		$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	1	0.05000	0.02500	0.01000	0.05300	0.02719	0.01129
2	1	0.18561	0.11860	0.06432	0.18780	0.11980	0.06649
3	1	0.60672	0.49637	0.36727	0.60771	0.49595	0.36708
4	1	0.05000	0.02500	0.01000	0.05585	0.02949	0.01239
5	1	0.08371	0.04648	0.02124	0.08966	0.05104	0.02554
6	1	0.19689	0.12713	0.06983	0.20168	0.13245	0.07471
7	1	0.62920	0.51996	0.38984	0.62695	0.51853	0.39080
8	1	0.10277	0.05922	0.02832	0.11055	0.06535	0.03263
9	1	0.28646	0.19820	0.11854	0.29296	0.20534	0.12661
10	1	0.84184	0.76554	0.65404	0.84309	0.76775	0.65718
1	2	0.05000	0.02500	0.01000	0.05114	0.02581	0.01016
2	2	0.28910	0.20038	0.12010	0.28900	0.20071	0.12114
3	2	0.84143	0.76501	0.65341	0.84048	0.76584	0.65379
4	2	0.05000	0.02500	0.01000	0.05095	0.02579	0.01069
5	2	0.12757	0.07636	0.03823	0.12776	0.07690	0.03910
6	2	0.39005	0.28721	0.18568	0.39035	0.28645	0.18596
7	2	0.93630	0.89392	0.82139	0.93598	0.89128	0.81903
8	2	0.15498	0.09596	0.05007	0.15545	0.09679	0.05089
9	2	0.50354	0.39280	0.27324	0.50415	0.39365	0.27238
10	2	0.98569	0.97212	0.94373	0.98621	0.97289	0.94465

†Simulation numbers refer to the simulation settings that are summarized in Table 5; α denotes the level of significance.

2005) and van den Brakel and van Berkel (2002) are implemented in a software package, called X-tool. This package is available as a component of the Blaise survey processing software package, which was developed by Statistics Netherlands (Statistics Netherlands, 2002).

X-tool is a software package to test hypotheses about differences between population parameters that are observed under different survey implementations in randomized experiments embedded in sample surveys. Hypotheses can be tested about means, totals and ratios for $K \geqslant 2$ treatments. It is assumed that the data are collected according to the following design. First a probability sample is drawn from a finite population, which might be generally complex. Subsequently, this sample is randomly divided into K subsamples according to an experimental design. X-tool handles experiments that are designed as CRDs and RBDs, where sampling structures like strata, clusters, PSUs or interviewers are potential block variables. It is possible to analyse experiments where the ultimate sampling units as well as clusters of sampling units (e.g. people belonging to the same cluster or assigned to the same interviewer) are randomized over the different treatments.

With each subsample, data are collected under one of the K different treatments or survey implementations. On the basis of these K subsamples, X-tool calculates design-based estimates for the population parameters that are observed under the different treatments and the covariance matrix of the $K - 1$ contrasts between these estimates. Subsample estimates for means, totals and ratios are based on the Hájek estimator or the generalized regression estimator to account for the sampling design, experimental design and the weighting procedure of the survey. The integrated method for weighting individuals and households of Lemaître and Dufour (1987) can be applied under the generalized regression estimator to obtain equal weights for

Table 8. Non-centrality parameters and first and second moments for the simulated distributions of the Wald statistics†

Simulation	*Population*	δ	$E(\chi^2_{[2,\delta]})$	$var(\chi^2_{[2,\delta]})$	$E(W)$	$S(W)$
1	1	0.0000	2.0000	4.0000	2.0286	4.1806
2	1	0.7888	3.5775	10.3102	3.5963	10.6617
3	1	3.1517	8.3034	29.2138	8.3603	30.5230
4	1	0.0000	2.0000	4.0000	2.0582	4.3959
5	1	0.2143	2.4286	5.7145	2.4865	6.2540
6	1	0.8488	3.6976	10.7905	3.7664	11.7325
7	1	3.3083	8.6166	30.4665	8.7289	33.3335
8	1	0.3282	2.6564	6.6258	2.7293	7.2417
9	1	1.3171	4.6343	14.5372	4.7524	15.9965
10	1	5.3448	12.6896	46.7586	12.8852	51.0562
1	2	0.0000	2.0000	4.0000	2.0113	4.0286
2	2	1.3309	4.6617	14.6468	4.6675	14.8167
3	2	5.3391	12.6781	46.7126	12.6877	47.2377
4	2	0.0000	2.0000	4.0000	2.0088	4.0808
5	2	0.4709	2.9419	7.7674	2.9470	7.9446
6	2	1.8623	5.7247	18.8987	5.7374	19.3803
7	2	7.2443	16.4885	61.9542	16.4965	63.6890
8	2	0.6233	3.2466	8.9862	3.2603	9.1695
9	2	2.4990	6.9981	23.9922	7.0065	24.3629
10	2	10.0588	22.1175	84.4702	22.2285	87.1798

†Simulation numbers refer to the simulation settings that are summarized in Table 5.

individuals belonging to the same household. Also a bounding algorithm that is based on Huang and Fuller (1978) can be applied to avoid negative correction weights.

For two-treatment experiments, the design-based t-statistic that was proposed by van den Brakel and van Berkel (2002) was implemented. For experiments with more than two treatments, the design-based Wald statistic from Section 4 was applied.

The interface of X-tool is based on tab sheets where the user provides the required information to load the data, and to specify the sample design, experimental design and the required estimation options. The analysis results are specified on the final tab sheet. The user must specify the design weights, i.e. the inverse of the inclusion probabilities, of the initial sample. On the basis of the experimental design, the design weights for the different subsamples are derived automatically. In the case of the generalized regression estimator, the user must specify a weighting model, which is applied to each subsample. X-tool checks whether in each subsample a prespecified minimum number of observations is available in each cell that is formed by this weighting model. If not, a reduced weighting model is proposed, which can be adjusted by the user. X-tool also checks whether a prespecified minimum block size is available. For the variance of the generalized regression estimator, the user can choose to use variance estimators (4.6) and (4.11) or to multiply the residuals in these estimators with the correction weights, as suggested by Särndal *et al.* (1992), section 6.6. The pooled variance estimator (4.8) can be selected, e.g. in the case of insufficient sample size within each block.

7. Application to the Dutch Labour Force Survey

In Section 2.5 an experiment was introduced with small prepayment incentives that was embedded in the LFS to improve response rates and to reduce non-response bias. In this section the effects of the incentive on response behaviour and response bias are analysed by using logistic

Table 9. Response account incentive experiment in the Dutch LFS†

Category	*Results for the following incentive treatments:*							
	€0.0		*€1.95*		*€3.9*		*€7.8*	
	Number	*%*	*Number*	*%*	*Number*	*%*	*Number*	*%*
Approached addresses	6200		3150		3150		500	
Frame errors	281		146		136		16	
Visited addresses	5919		3004		3014		484	
Visited households	5994	100	3060	100	3107	100	508	100
Non-contact	313	5.2	162	5.3	138	4.4	31	6.3
Complete response	3821	63.7	2124	69.4	2236	72.0	356	72.4
Partial response	56	0.9	14	0.5	17	0.5	4	0.8
Refusal	1372	23.0	543	17.7	493	15.9	62	12.6
Rest	432	7.2	217	7.1	223	7.2	39	7.9

†Frame errors contain unable to locate address, in construction, no housing unit or vacant housing unit; the rest category contains language problems, break-off or no opportunity.

regression models and the design-based procedures that were developed in the preceding sections.

Table 9 contains an overview of the response account of the four subsamples in the experiment.

It follows from the results in Table 9 that the prepayment incentives result in an obvious increase in the response rate and a decrease in the refusal rate. Response and refusal rates of the households visited are modelled in two separate logistic regression models. This analysis serves two purposes. Firstly, hypotheses about the effect of the incentives on response behaviour may be tested. Secondly, additional information may be obtained on whether the incentive increases the response across the entire target population or whether the additional response comes from specific groups. An incentive that affects the response behaviour of specific subpopulations differently might result in a less representative sample. The generalized regression estimator, however, compensates for this overrepresentation and underrepresentation as long as these subpopulations are included in the weighting model. Second- and higher order interactions between the incentive and sociodemographic categorical variables in the logistic regression models indicate that the variation in response between different subpopulations increases. These interactions also show from which subpopulations the additional response originates.

In the logistic regression model for response rates, the dependent binary variable indicates whether a household completely responded against the remaining four response categories (non-contact, partial response, refusal and the rest categories). In the logistic regression model for refusal rates, the dependent binary variable indicates whether a household refused to participate with the survey against the remaining four response categories (non-contact, partial response, complete response and the rest categories). The response behaviour in both models is assumed to depend on

(a) a general mean,
(b) the treatment (*inc*), which is a quantitative explanatory variable containing the value of the incentive,

(c) a block variable (*block*) with 13 categories (interviewers are the block variable, but adjacent interviewer regions are collapsed in 13 blocks),
(d) auxiliary variables
  (i) region with 13 categories for the 12 provinces and one category for the four major cities Amsterdam, The Hague, Rotterdam and Utrecht,
  (ii) level of urbanization (*urb*) with five categories,
  (iii) age with five categories specifying the age class of the head of the household,
  (iv) household size (*hsize*), with five categories, specifying the number of household members (1–4, and 5 or larger),
  (v) ethnicity (*ethn*) with three categories, specifying whether the head of the household has a native, western or non-western background,
  (vi) income as a quantitative explanatory variable containing standardized household income and
  (vii) marital status (*marst*) of the head of the household with four levels (married, unmarried, divorced and widowed).

Since men are regarded as the head of the household and there is no additional information which household member is contacted in the case of a non-response, it is not very useful to include gender as an auxiliary variable in the models for response behaviour of the households. Quadratic terms in the quantitative covariates and all second- and third-order interactions between the variables are initially considered for backward model selection. The final selected models had the terms that are given in the first column of Table 10 of estimation results for response rates and Table 11 for refusal rates. For brevity, the regression coefficients with their standard error and test statistics for separate categories of a categorical variable are only expressed if they interact with the treatment variable.

Table 10. Logistic regression analysis for the response rate of the households visited

Parameter	*Coefficient*	*Standard error*	*Wald statistic*	*Degrees of freedom*	*p-value*
Mean	0.941	0.272	12.013	1	0.001
Block			52.714	12	0.000
Incentive linear	0.179	0.029	39.028	1	0.000
Incentive quadratic	−0.015	0.005	10.723	1	0.001
Urbanization level			20.366	4	0.000
Ethnicity			8.302	2	0.016
Western	−0.150	0.146	1.052	1	0.305
Non-western	−0.480	0.172	7.706	1	0.006
Region			46.64	12	0.000
Household size			45.180	4	0.000
Income	0.036	0.059	0.364	1	0.546
Age			16.320	4	0.003
Marital status			8.465	3	0.037
Incentive linear × ethnicity			11.365	2	0.003
Incentive linear × western	0.065	0.087	0.561	1	0.454
Incentive linear × non-western	−0.277	0.087	10.154	1	0.001
Incentive quadratic × ethnicity			7.015	2	0.030
Incentive quadratic × western	−0.016	0.014	1.314	1	0.252
Incentive quadratic × non-western	0.031	0.014	5.020	1	0.025
Ethnicity × household size			23.805	8	0.002
Income × age			14.383	4	0.006

Table 11. Logistic regression analysis for refusal rate of the households visited

Parameter	*Coefficient*	*Standard error*	*Wald statistic*	*Degrees of freedom*	*p-value*
Mean	−1.479	0.219	45.734	1	0.000
Block			28.110	12	0.005
Incentive linear	−0.164	0.030	30.001	1	0.000
Incentive quadratic	0.011	0.005	4.315	1	0.038
Urbanization level			8.798	4	0.066
Ethnicity			14.738	2	0.001
Income	−0.147	0.029	25.291	1	0.000
Age			19.570	4	0.001
Age×urbanization level			30.329	16	0.016
Income×ethnicity			7.345	2	0.025

The incentive has a significant linear and quadratic effect on response and refusal rates. The quadratic relationship between the value of the incentive and the logit of the response behaviour implies that an optimal value for the incentive can be derived. In Fig. 1 the effect of the incentive on the odds ratio is graphically visualized for response and refusal rates, i.e. $p/(1-p) = \exp(\hat{b}x + \hat{c}x^2) \equiv f(x)$, with p the probability of response or refusal, x the value of the incentive and $\hat{b}$ and $\hat{c}$ the estimated regression coefficients for the linear and quadratic effect of the incentive. Under the assumed quadratic relationship, response rates are maximized with an incentive of a value of about €6 and refusal rates are minimized with an incentive of about €7.5, since the extremes of $f(x)$ are obtained by $x = -\hat{b}/2\hat{c}$. This interpretation must be made with due caution. Note that, in the experiment, incentives are applied with values of €0, €2, €4 and €8 only (Table 1) and that the response rate for the subsample that is assigned to €4 and €8 are both equal to 72% (Table 9). As a result, the optimal value of €6 is induced by the quadratic relationship assumed. There is, however, no empirical evidence that the response at €6 is higher than at €4 and €8. Square-root and logarithmic transformations were considered as an alternative but did not result in better model fits. If the introduction of an incentive is considered, €4 would therefore be the most likely value. These results are in line with the prevailing opinion in the literature

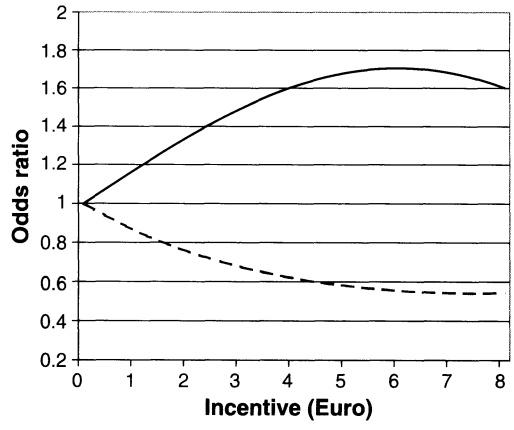

Fig. 1. Response curves for the incentive value on the odds ratio for the response rate (———) and refusal rate (— —)

Table 12. Unemployed labour force†

Treatment		*Estimate*	*Contrast*		
			Treatment	*Estimate*	*Standard error*
1	(€0.0)	6.245			
2	(€1.95)	5.972	1–2	0.273	0.55
3	(€3.9)	5.722	1–3	0.523	0.56
4	(€7.9)	6.797	1–4	−0.552	1.16

†Wald statistic, 1.357; p-value, 0.716.

that small prepaid incentives are very effective in improving response rates in household surveys (see for example Groves and Couper (1998) or Singer (2002)), and the diminishing (non-linear) effects of increasing amounts of incentives (Curtin *et al.*, 2007; Dillman, 1978, 2000).

It follows from the logistic regression analysis that ethnicity is the only auxiliary variable that interacts with incentives in the model for response rates. The response rate of the non-western population is not increased by incentives, which implies that the underrepresentation of this group increases with the introduction of an incentive.

The logistic regression analysis shows that response rates are significantly increased for almost all sociodemographic subpopulations with the exception of the non-western. The question arises whether the increased response results in a decrease in the non-response bias in the estimates of the unemployed labour force and the total unemployment. The analysis procedure that was proposed in Section 4 is applied to test the effect of the various incentives on the main parameter estimates of the LFS, i.e. the unemployed labour force and total unemployment.

The generalized regression estimator is applied to obtain estimates for both parameters under the four different treatments in the first wave, using the integrated method for weighting people and families of Lemaître and Dufour (1987). The inclusion probabilities reflect the sampling design of the LFS and the experimental design that is used to divide the sample into four subsamples. The following weighting scheme, which contains the most important auxiliary information of the regular weighting scheme of the LFS, was applied: age + region + marital status + gender + ethnicity, where the five variables are categorical. The analysis results that were obtained with X-tool are given in Tables 12 and 13. There is no indication that the increased response rates, which were obtained with the incentive treatments, result in different parameter estimates

Table 13. Total unemployment†

Treatment		*Estimate*	*Contrast*		
			Treatment	*Estimate*	*Standard error*
1	(€0.0)	459			
2	(€1.95)	444	1–2	15	41
3	(€3.9)	430	1–3	29	42
4	(€7.9)	516	1–4	−57	90

†Wald statistic, 1.089; p-value, 0.780; estimates and standard errors times 1000.

for the unemployed labour force and total unemployment. With the given sample size, there are no indications that the increased response rates affect the non-response bias in these LFS parameters.

8. Discussion

In this paper a design-based analysis procedure is presented for experiments that are embedded in complex sample surveys that accounts for the sampling design, the experimental design and the weighting procedure of the survey. This approach is particularly appropriate to analyse experiments that aim to quantify the effects of alternative survey implementations on the parameter estimates of on-going sample surveys. Experimental designs are considered, where clusters of sampling units are used as the experimental units in CRDs or RBDs to test hypotheses about parameters that are defined as means, totals and ratios. Randomizing clusters of sampling units instead of ultimate sampling units implies a decrease in the effective sample size of the experiment, resulting in larger variances for the estimated treatment effects and less power for hypothesis testing. Such designs are mainly considered to deal with limitations that are encountered in the fieldwork with data collection, e.g. to assure that all members of the same household or all sampling units that are assigned to the same interviewer are assigned to the same treatment.

The variance estimation procedure is approximately design unbiased with the exception of contrasts between ratios under the alternative hypothesis. A simulation study with a complex sampling design, i.e. stratified two-stage sampling with unequal selection probabilities and large sampling fractions, is conducted to study the analysis procedure proposed. There is no indication that the variance estimation procedures result in biased estimates, even under the alternative hypotheses for ratios. Simulated powers approximate the real powers reasonably well, although the sample distribution of the Wald statistic for ratios appears to converge more slowly to the limit distribution compared with the sample distribution of the Wald statistic for the means.

An important advantage of the variance estimation procedure proposed is that no joint inclusion probabilities and no design covariances between the subsample estimates are required. As a result a design-based analysis procedure for experiments that are embedded in complex sampling designs is obtained with the appealing relatively simple structure as if sampling units are drawn with unequal selection probabilities with replacement.

The experiment with incentives in the Dutch LFS, which is used as a numerical example, illustrates how the design-based approach can be completed with a more direct modelling approach like logistic regression analysis. The design-based analysis directly tests differences between target parameters that are observed under the different survey approaches by using the estimation procedure as applied in the on-going survey. If, however, the null hypothesis of no treatment effects is rejected, it might be difficult to decide which treatment is better. In this application, the logistic regression analysis helps to understand whether observed differences are induced by an increased response across the entire population or by specific groups and to draw better-founded conclusions. The incentives appear to increase the response across the entire population, with the exception of the non-western population. Since ethnicity is included in the weighting scheme of the LFS, the generalized regression estimator will compensate increased underrepresentation of this population. An incentive of €4 appears to be the most effective value to increase response rates and to decrease refusal rates. The increased response will result in a small reduction in the design variance but there are no indications that it results in a reduction in the response bias.

Acknowledgements

The author thanks the Associate Editor, the referees and Rita Gircour for their constructive comments on former drafts of this paper. The software package X-tool was implemented by Joeri Roels. The views that are expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

Appendix A: Variance of contrasts

A.1. Population means

An expression for the covariance matrix of the $K-1$ contrasts between $\hat{\bar{\mathbf{Y}}}_{\text{greg}}$ is derived in this appendix. Therefore measurement error model (4.1) is expressed in matrix notation first, i.e.

$$\mathbf{y}_{ijl} = \mathbf{j}u_{ij} + \boldsymbol{\beta} + \mathbf{j}\gamma_l + \boldsymbol{\varepsilon}_{ij} \tag{A.1}$$

with $\mathbf{y}_{ijl} = (y_{ij1}, \ldots, y_{ijKl})^{\mathrm{T}}$ the vector with the K potential responses for each of the K treatments of sampling unit (i, j), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ a K-vector with treatment effects, $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijK})^{\mathrm{T}}$ a K-vector with measurement errors and $\mathbf{j}$ a K-vector with each element 1. It is assumed that

$$E_m(\boldsymbol{\varepsilon}_{ij}) = \mathbf{0},$$

$$\operatorname{cov}_m(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{i'j'}) = \begin{cases} \boldsymbol{\Sigma}_{ij} + \boldsymbol{\Sigma}_j & i = i', j = j', \\ \boldsymbol{\Sigma}_j & i \neq i', j = j', \\ \mathbf{O} & i \neq i', j \neq j', \end{cases}$$

$$\operatorname{cov}_m(\xi_l, \xi_{l'}) = \begin{cases} \sigma_l^2 & l = l', \\ 0 & l \neq l', \end{cases}$$

where $\boldsymbol{\Sigma}_{ij}$ denotes a $K \times K$ covariance matrix of the measurement error for sampling unit (i, j) and $\boldsymbol{\Sigma}_j$ denotes a $K \times K$ matrix with the covariance between the measurement errors of the sampling units from the jth PSU, $\mathbf{0}$ a K-vector with each element 0 and $\mathbf{O}$ a $K \times K$ matrix with each element 0. The generalized regression estimator $\hat{\bar{Y}}_{k;\text{greg}}$ is approximated with a first-order Taylor linearization about the true population values $(E_m\bar{Y}_k, \mathbf{b}_k, \bar{\mathbf{X}})$, i.e.

$$\hat{\bar{Y}}_{k;\text{greg}} \approx \hat{\bar{Y}}_k + \mathbf{b}_k^{\mathrm{T}}(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}) = \hat{\bar{E}}_k + \mathbf{b}_k^{\mathrm{T}}\bar{\mathbf{X}},$$

with

$$\hat{\bar{E}}_k = \frac{1}{N}\sum_{j \in s_k}\sum_{i=1}^{n_j} \frac{y_{ijk} - \mathbf{b}_k^{\mathrm{T}}\mathbf{x}_{ij}}{\pi_j^{*1}\pi_{i|j}^{\mathrm{II}}} \equiv \frac{1}{N}\sum_{j \in s_k} \frac{\hat{y}_{jk} - \mathbf{b}_k^{\mathrm{T}}\hat{\mathbf{x}}_j}{\pi_j^{*1}}$$

and

$$\mathbf{b}_k = \left(\sum_{j=1}^{M}\sum_{i=1}^{N_j} \frac{\mathbf{x}_{ij}\mathbf{x}_{ij}^{\mathrm{T}}}{\omega_{ij}}\right) \sum_{j=1}^{M}\sum_{i=1}^{N_j} \frac{\mathbf{x}_{ij}E_m(y_{ijk})}{\omega_{ij}}, \tag{A.2}$$

the regression coefficients of the regression function of $E_m(y_{ijk}) = u_{ij} + \beta_k + \psi$ on $\mathbf{x}_{ij}$ in the finite population. $\mathbf{V}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\text{greg}})$ can be approximated by $\mathbf{V}(\mathbf{C}\hat{\bar{\mathbf{E}}})$, with $\hat{\bar{\mathbf{E}}} = (\hat{\bar{E}}_1, \ldots, \hat{\bar{E}}_K)^{\mathrm{T}}$. Let E_s and E_e denote the expectation with respect to the sampling design and the experimental design and let cov_s and cov_e denote the covariance with respect to the sampling design and the experimental design. Now $\mathbf{V}(\mathbf{C}\hat{\bar{\mathbf{E}}})$ can be decomposed as

$$\mathbf{V}(\mathbf{C}\hat{\bar{\mathbf{E}}}) = \operatorname{cov}_m\{E_sE_e(\mathbf{C}\hat{\bar{\mathbf{E}}}|m, s)\} + E_m \operatorname{cov}_s\{E_e\mathbf{C}\hat{\bar{\mathbf{E}}}|m, s)\} + E_mE_s \operatorname{cov}_e(\mathbf{C}\hat{\bar{\mathbf{E}}}|m, s). \tag{A.3}$$

Under the condition that there is a constant vector $\mathbf{a}$ of order H, such that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij} = 1$ for all units in the population, it follows from equation (A.2) that $\mathbf{b}_k = \tilde{\mathbf{b}} + \mathbf{a}\beta_k$ with $\tilde{\mathbf{b}}$ the regression coefficients of the regression of the intrinsic values biased with the average interviewer effect, i.e. $u_{ij} + \psi$, on $\mathbf{x}_{ij}$. Let $\mathbf{B}$ denote an $H \times K$ matrix where the columns are the vectors $\mathbf{b}_k$ that are defined in equation (A.2). It follows that the contrasts of the vector with the residuals of the generalized regression estimator equal

$$\mathbf{C}(\mathbf{y}_{ij} - \mathbf{B}^{\mathrm{T}}\mathbf{x}_{ij}) = \mathbf{C}(\mathbf{j}u_{ij} + \boldsymbol{\beta} + \mathbf{j}\gamma_l + \boldsymbol{\varepsilon}_{ij} - \mathbf{j}\tilde{\mathbf{b}}^{\mathrm{T}}\mathbf{x}_{ij} - \boldsymbol{\beta}) = \mathbf{C}\boldsymbol{\varepsilon}_{ij}. \tag{A.4}$$

According to result (A.4) the contrast of the vector with the residuals of the generalized regression estimator equals the contrast of the measurement errors which are, according to the measurement error model, independent for sampling units from different PSUs. Taking expectations and covariances of the three components on the right-hand side in equation (A.3) and taking advantage of result (A.4) give

$$\mathrm{cov}_m\{E_s E_e(\mathbf{C}\hat{\tilde{\mathbf{E}}}|m,s)\} = \frac{1}{N^2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\mathbf{C}\boldsymbol{\Sigma}_{ij}\mathbf{C}^{\mathrm{T}} + \sum_{j=1}^{M}\left(\frac{N_j}{N}\right)^2\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}},$$

$$E_m\,\mathrm{cov}_s\{E_e(\mathbf{C}\hat{\tilde{\mathbf{E}}}|m,s)\} = \frac{1}{N^2}\sum_{j=1}^{M}\sum_{j=1}^{N_j}\left(\frac{1}{\pi_j^{\mathrm{I}}\pi_{i|j}^{\mathrm{II}}}-1\right)\mathbf{C}\boldsymbol{\Sigma}_{ij}\mathbf{C}^{\mathrm{T}} + \frac{1}{N^2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\sum_{i'=1}^{N_j}\frac{\pi_{ii'|j}^{\mathrm{II}}-\pi_j^{\mathrm{I}}\pi_{i|j}^{\mathrm{II}}\pi_{i'|j}^{\mathrm{II}}}{\pi_j^{\mathrm{I}}\pi_{i|j}^{\mathrm{II}}\pi_{i'|j}^{\mathrm{II}}}\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}},$$

$$E_m E_s\,\mathrm{cov}_e(\mathbf{C}\hat{\tilde{\mathbf{E}}}|m,s) = E_m E_s\mathbf{C}\mathbf{D}\mathbf{C}^{\mathrm{T}} - \frac{1}{N^2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\frac{\mathbf{C}\boldsymbol{\Sigma}_{ij}\mathbf{C}^{\mathrm{T}}}{\pi_j^{\mathrm{I}}\pi_{i|j}^{\mathrm{II}}} - \frac{1}{N^2}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\sum_{i'=1}^{N_j}\frac{\pi_{ii'|j}^{\mathrm{II}}}{\pi_j^{\mathrm{I}}\pi_{i|j}^{\mathrm{II}}\pi_{i'|j}^{\mathrm{II}}}\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}},$$

where $\pi_{ii'|j}^{\mathrm{II}}$ denotes the joint inclusion probability that SSU i and i' are both drawn from the jth PSU in the second stage of the sampling design, and $\mathbf{D}$ is a $K \times K$ diagonal matrix. In the case of an RBD, the diagonal elements of $\mathbf{D}$ are given by

$$d_k = \frac{1}{N^2}\sum_{b=1}^{B}\frac{1}{m_{bk}}\frac{1}{m_{b+}-1}\sum_{j\in s_b}\left(\frac{m_{b+}e_{jk}}{\pi_j^{\mathrm{I}}} - \frac{1}{m_{b+}}\sum_{j'\in s_b}\frac{m_{b+}e_{j'k}}{\pi_{j'}^{\mathrm{I}}}\right)^2,$$

with

$$e_{jk} = \sum_{i=1}^{n_j}\frac{y_{ijk}-\mathbf{b}_k^{\mathrm{T}}\mathbf{x}_{ij}}{\pi_{i|j}^{\mathrm{II}}}.$$

Collecting results gives $\mathbf{V}(\mathbf{C}\hat{\tilde{\mathbf{E}}}) = E_m E_s\mathbf{C}\mathbf{D}\mathbf{C}^{\mathrm{T}}$. Conditionally on the realization of the sample and the measurement errors an approximately design-unbiased estimator for $\mathbf{D}$ can be derived directly. Therefore, $\mathbf{V}(\mathbf{C}\bar{\tilde{\mathbf{Y}}}_{\mathrm{greg}})$ can be conveniently stated implicitly as the expectation over the measurement error model and the sampling design. In the case of an RBD, the allocation of the PSUs to subsample s_k can be considered as stratified simple random sampling without replacement from s, where the block variables are the strata. Consequently a design-unbiased estimator for $\mathbf{V}(\mathbf{C}\bar{\tilde{\mathbf{Y}}}_{\mathrm{greg}})$ is given by $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^{\mathrm{T}}$ where $\hat{\mathbf{D}}$ is a diagonal matrix with elements that are defined by expression (4.6). Results for a CRD follow as a special case by taking $B=1$, $m_{bk}=m_k$ and $m_{b+}=m_+$.

The proof of this result is in essence analogous to the proof of result (28) in van den Brakel and Renssen (2005) for the analysis of means where the ultimate sampling units of the sampling design are the experimental units that are randomized over the treatments. Now the derivations are applied on the level of the PSUs (which are the experimental units) with $\hat{y}_{jk}$ defined in equation (4.3). The properties of the randomization vectors that were used in the appendix of van den Brakel and Renssen (2005) are defined at the level of the PSUs, i.e. n_{jk} and n_{j+} in equations (42)–(46) in van den Brakel and Renssen (2005) are replaced by m_{bk} and m_{b+} as defined in Section 4.2.

This variance structure holds for the Hájek estimator that is defined in expression (4.10). This estimator only uses the population total as auxiliary information and thus satisfies the condition that there is a constant H-vector such that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij}=1$ for all units in the population.

A.2. Ratios

To obtain an approximation of the covariance matrix of the $K-1$ contrasts between $\hat{\mathbf{R}}_{\mathrm{greg}}$, the elements $\hat{R}_{k;\mathrm{greg}}$, which are defined in expression (4.10), are linearized about the true population value R_k by a first-order Taylor approximation, i.e.

$$\hat{R}_{k;\mathrm{greg}} \approx R_k + \frac{1}{\bar{Z}_k}(\hat{\tilde{Y}}_{k;\mathrm{greg}} - R_k\hat{\tilde{Z}}_{k;\mathrm{greg}}). \tag{A.5}$$

Subsequently $\hat{\tilde{Y}}_{k;\mathrm{greg}}$ and $\hat{\tilde{Z}}_{k;\mathrm{greg}}$ in approximation (A.5) are linearized with a first-order Taylor approximation about $(E_m\bar{Y}_k, \mathbf{b}_k, \bar{\mathbf{X}})$ and $(E_m\bar{Z}_k, \mathbf{f}_k, \bar{\mathbf{X}})$ respectively. Here $\mathbf{b}_k$ is defined in equation (A.2) and $\mathbf{f}_k$ denotes

the regression coefficients of z_{ijk} on $\mathbf{x}_{ij}$ in the finite population, which is defined by equation (A.2) where y_{ijk} is replaced by z_{ijk}. It follows that

$$\hat{R}_{k;\text{greg}} \approx R_k + \frac{1}{\bar{Z}_k}[\hat{\bar{Y}}_k + \mathbf{b}_k^{\mathrm{T}}(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}_k) - R_k\{\hat{\bar{Z}}_k + \mathbf{f}_k^{\mathrm{T}}(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}_k)\}] \equiv R_k + \hat{\bar{E}}_k + \frac{1}{\bar{Z}_k}(\mathbf{b}_k^{\mathrm{T}}\bar{\mathbf{X}} - R_k\mathbf{f}_k^{\mathrm{T}}\bar{\mathbf{X}}),$$

where

$$\hat{\bar{E}}_k = \frac{1}{N\bar{Z}_k}\sum_{j\in s_k}\frac{\hat{y}_{jk} - \mathbf{b}_k^{\mathrm{T}}\hat{\mathbf{x}}_j - R_k(\hat{z}_{jk} - \mathbf{f}_k^{\mathrm{T}}\hat{\mathbf{x}}_j)}{\pi_j^{*\mathrm{I}}}.$$

Express the measurement error model for the observations of the target parameter of the numerator and the denominator in matrix notation. An expression for the numerator is given by equation (A.1). For the denominator an equivalent model is assumed. Let $\mathbf{Z} = \text{diag}(\bar{Z}_1, \ldots, \bar{Z}_K)$ and $\mathbf{R} = \text{diag}(R_1, \ldots, R_K)$. Under the null hypothesis and the condition that there is a constant H-vector such that $\mathbf{a}^{\mathrm{T}}\mathbf{x}_{ij} = 1$ for all units in the population it follows that the contrasts of the vector with the residuals of the generalized regression estimator equal

$$\mathbf{C}\mathbf{Z}^{-1}\{(\mathbf{y}_{ij} - \mathbf{B}^{\mathrm{T}}\mathbf{x}_{ij}) - \mathbf{R}(\mathbf{z}_{ij} - \mathbf{F}^{\mathrm{T}}\mathbf{x}_{ij})\} = \mathbf{C}\mathbf{Z}^{-1}(\boldsymbol{\varepsilon}_{ij}^{(y)} - \mathbf{R}\boldsymbol{\varepsilon}_{ij}^{(z)}). \tag{A.6}$$

Here $\mathbf{F}$ is an $H \times K$ matrix with the columns containing the vectors $\mathbf{f}_k$, and $\boldsymbol{\varepsilon}_{ij}^{(y)}$ and $\boldsymbol{\varepsilon}_{ij}^{(z)}$ are vectors with measurement errors for the target parameter of the numerator and the denominator respectively. Equivalent to the derivation for the covariance matrix of the sample means, it follows that $\mathbf{V}(\mathbf{C}\hat{\mathbf{R}}_{\text{greg}}) = E_m E_s \mathbf{C}\mathbf{D}^{(R)}\mathbf{C}^{\mathrm{T}}$. In the case of an RBD the diagonal elements of $\mathbf{D}^{(R)}$ are given by

$$d_k^{(R)} = \frac{1}{N^2\bar{Z}_{k;\text{greg}}^2}\sum_{b=1}^{B}\frac{1}{m_{bk}}\frac{1}{m_{b+}-1}\sum_{j\in s_b}\left(\frac{m_{b+}e_{jk}}{\pi_j^{\mathrm{I}}} - \frac{1}{m_{b+}}\sum_{j'\in s_b}\frac{m_{b+}e_{j'k}}{\pi_{j'}^{\mathrm{I}}}\right)^2, \tag{A.7}$$

with

$$e_{jk} = \sum_{i=1}^{n_j}\frac{y_{ijk} - \mathbf{b}_k^{\mathrm{T}}\mathbf{x}_{ij} - R_k(z_{ijk} - \mathbf{f}_k^{\mathrm{T}}\mathbf{x}_{ij})}{\pi_{i|j}^{\mathrm{II}}}.$$

For ratios this result only holds under the null hypothesis since the equality in expression (A.6) requires that the diagonal elements of $\mathbf{Z}$ as well as $\mathbf{R}$ are equal. Since the allocation of the PSUs to subsample s_k can be considered as stratified simple random sampling without replacement from s, where the block variables are the strata, it follows that expression (4.11) is a design-unbiased estimator for equation (A.7). Results for a CRD are obtained as a special case by taking $B = 1$, $m_{bk} = m_k$ and $m_{b+} = m_+$.

References

van den Brakel, J. A. (2001) Design and analysis of experiments embedded in complex sample surveys. *PhD Thesis*. Erasmus University, Rotterdam.

van den Brakel, J. A. and van Berkel, C. A. M. (2002) A design-based analysis procedure for two-treatment experiments embedded in sample surveys. *J. Off. Statist.*, **18**, 217–231.

van den Brakel, J. A. and Renssen, R. H. (1998) Design and analysis of experiments embedded in sample surveys. *J. Off. Statist.*, **14**, 277–295.

van den Brakel, J. A. and Renssen, R. H. (2005) Analysis of experiments embedded in complex sampling designs. *Surv. Methodol.*, **31**, 23–40.

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

Curtin, R., Singer, E. and Presser, S. (2007) Incentives in random digit dial telephone surveys. *J. Off. Statist.*, **23**, 91–105.

Dillman, D. A. (1978) *Mail and Telephone Surveys*. New York: Wiley.

Dillman, D. A. (2000) *Mail and Internet Surveys*. New York: Wiley.

Fellegi, I. P. (1964) Response variance and its estimation. *J. Am. Statist. Ass.*, **59**, 1016–1041.

Fienberg, S. E. and Tanur, J. M. (1987) Experimental and sampling structures: parallels diverging and meeting. *Int. Statist. Rev.*, **55**, 75–96.

Fienberg, S. E. and Tanur, J. M. (1988) From the inside out and the outside in: combining experimental and sampling structures. *Can. J. Statist.*, **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. *Int. Statist. Rev.*, **64**, 237–253.
Groves, R. M. and Couper, M. P. (1998) *Nonresponse in Household Interview Surveys*. New York: Wiley.
Hájek, J. (1971) Comment on "An essay on the logical foundations of survey sampling" by D. Basu. In *Foundations of Statistical Inference* (eds V. P. Godambe and D. A. Sprott). Toronto: Holt, Rinehart, and Winston.
Hartley, H. O. and Rao, J. N. K. (1978) Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement* (ed. N. K. Namboodiri), pp. 35–43. New York: Academic Press.
Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Ass.*, **47**, 663–685.
Huang, E. T. and Fuller, W. A. (1978) Nonnegative regression estimation for survey data. *Proc. Socl Statist. Sect. Am. Statist. Ass.*, 300–305.
Knottnerus, P. (2002) *Sample Survey Theory; Some Pythagorean Perspectives*. New York: Springer.
Lemaître, G. and Dufour, J. (1987) An integrated method for weighting persons and families. *Surv. Methodol.*, **13**, 199–207.
Mahalanobis, P. C. (1946) Recent experiments in statistical sampling in the Indian Statistical Institute. *J. R. Statist. Soc.*, **109**, 326–370.
Montgomery, D. C. (2001) *Design and Analysis of Experiments*, 5th edn. New York: Wiley.
Narain, R. (1951) On sampling without replacement with varying probabilities. *J. Ind. Soc. Agric. Statist.*, **3**, 169–174.
Särndal, C.-E., Swensson, B. and Wretman, J. H. (1992) *Model Assisted Survey Sampling*. New York: Springer.
Searle, S. R. (1971) *Linear Models*. New York: Wiley.
Singer, E. (2002) The use of incentives to reduce nonresponse in household surveys. In *Survey Nonresponse* (eds R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 163–177. New York: Wiley.
Statistics Netherlands (2002) *Blaise Developer's Guide*. Heerlen: Statistics Netherlands. (Available from `www.Blaise.com`.)