Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Design-based analysis of factorial designs embedded in probability samples

by Jan A. van den Brakel

January 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2013
Vol. 39, No. 2, pp. 323-349
Statistics Canada, Catalogue No. 12-001-X

Design-based analysis of factorial designs embedded in probability samples

Jan A. van den Brakel[1]

Abstract

At national statistical institutes experiments embedded in ongoing sample surveys are frequently conducted, for example to test the effect of modifications in the survey process on the main parameter estimates of the survey, to quantify the effect of alternative survey implementations on these estimates, or to obtain insight into the various sources of non-sampling errors. A design-based analysis procedure for factorial completely randomized designs and factorial randomized block designs embedded in probability samples is proposed in this paper. Design-based Wald statistics are developed to test whether estimated population parameters, like means, totals and ratios of two population totals, that are observed under the different treatment combinations of the experiment are significantly different. The methods are illustrated with a real life application of an experiment embedded in the Dutch Labor Force Survey.

Key Words: Completely randomized designs; Design-based inference; Embedded experiments; Measurement error models; Model-assisted inference; Randomized block designs.

1 Introduction

The fields of randomized experiments and probability sampling are traditionally two separated domains of applied statistics. Both, however, come together if experiments are embedded in ongoing sample surveys. Randomized experiments embedded in ongoing sample surveys are frequently conducted to compare and test the effect of alternative survey implementations on the outcomes of a sample survey. The purpose of such empirical research is to improve the quality and efficiency of the underlying survey processes or to obtain more quantitative insight into the various sources of non-sampling errors. Many experiments conducted in this context are small scaled or conducted with specific groups. The value of empirical research into survey methods is strengthened as conclusions can be generalized to populations larger than the sample that is included in the experiment. Selecting experimental units randomly from a larger target population, is an important tool to secure that results of an experiment can be generalized to populations larger than the group of people included in the experiment, as emphasized by Fienberg and Tanur (1987, 1988, 1989 and 1996). This naturally leads to randomized experiments embedded in ongoing sample surveys. In the survey literature, such experiments are also referred to as split-ballot designs or interpenetrating subsampling, and date back to Mahalanobis (1946).

At national statistical offices such experiments are particularly useful to quantify discontinuities in the series of repeated surveys due to adjustments to the survey process. Repeatedly conducted surveys make up series that describe the development of target parameters. Embedded experiments can be used to avoid one or more modifications in the survey process resulting in unexplained differences in the series of a survey.

1. Jan A. van den Brakel, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands and Department of Quantitative Economics, Maastricht University School of Business and Economics, P.O. Box 616, 6200 MD, Maastricht, The Netherlands. E-mail: jbrl@cbs.nl.

An important issue in the analysis of this kind of experiment is to find the right mode of inference. The statistical inference in survey sampling is traditionally design based or model assisted. This implies that the inference is predominantly based on the stochastic structure induced by the sampling design. A well-known design-based estimator is the Horvitz-Thompson (HT) estimator, developed by Narain (1951) and Horvitz and Thompson (1952) for unequal probability sampling from finite populations without replacement. Under the model assisted approach developed by Särndal, Swensson and Wretman (1992), the accuracy of the HT estimator is improved by taking advantage of available auxiliary information about the complete target population, resulting in the generalized regression (GREG) estimator. Many national statistical institutes rely on this design-based and model-assisted approach to compile official statistics.

The statistical inference that is traditionally employed in the theory of design and analysis of randomized experiments is predominantly model-based. The observations that are obtained in the experiment are assumed to be the realization of a linear model. To test hypotheses about treatment effects, F-tests are derived under the assumption of normally and independently distributed observations. An exception is Kempthorne (1955), where a randomization approach is proposed in a way that is similar to the design-based inference approach in sampling theory. The F-test is used as an approximation of the randomization test. The model-based inference for randomized experiments is not necessarily appropriate for the analysis of embedded experiments, particularly if a design-based or model-assisted inference is used in the ongoing survey to compile official statistics.

In an embedded experiment the probability sample of the ongoing survey is randomly divided into different subsamples according to an experimental design. Each subsample can be considered as a probability sample drawn from the finite target population and can be used to estimate parameters such as means, totals and ratios, that are observed under the different survey implementations or treatments of the experiment using the estimation procedure that is applied in the regular survey to compile official statistics. The purpose of such embedded experiments is to compare the effect of alternative survey implementations on the main parameter estimates of the ongoing survey and to test whether the observed differences between these parameter estimates are statistically significant. This is obtained with a design-based approach where point and variance estimates for the population parameters, are (approximately) design-unbiased with respect to the sample design used to draw an initial probability sample from the target population, and the experimental design used to randomize this sample over the different subsamples. This analysis must also reflect the specific details of the regular estimation approach used to compile official statistics, as far as this is possible with the available sample size under the different treatments.

Previous research has proposed such a design-based theory for the analysis of single-factor experiments that are designed as completely randomized designs (CRDs) or randomized block designs (RBDs) to test the effect of one factor on $K \geq 2$ levels (van den Brakel (2008); van den Brakel and Renssen (1998, 2005); van den Brakel and van Berkel (2002)). In their approach the GREG estimator is applied to derive design-based Wald- and t-statistics to test whether the differences between finite population parameter estimates observed under the different survey implementations are significantly different. This theory is further extended to the experiments embedded in rotating panel designs by Chipperfield and Bell (2010).

From standard experimental design theory it is well known that it is efficient to test different treatment factors simultaneously in one factorial design instead of conducting separate single-factor experiments

(Hinkelmann and Kempthorne (1994); Montgomery (2001)). It can be expected that different design parameters in a survey process interact with each other, *e.g.*, when different questionnaire designs and data collection modes are compared empirically. Factorial setups are indeed appropriate if more than one factor in the survey is adjusted and tested in an embedded experiment, since fewer experimental units are required to test the main effects of the treatment factors whereas interactions between the factors can be analyzed. Another advantage of testing different treatments simultaneously in a factorial design is that the validity of the observed results is extended, since the effects are observed over a wider range of conditions (Hinkelmann and Kempthorne (1994)). Therefore the design-based theory for the analysis of embedded experiments is extended to factorial designs in this paper.

The theory for factorial designs where the effect of two factors is tested simultaneously is developed in section 2. Subsequently the methodology is extended to higher order factorial designs in section 3. In section 4, the methodology is extended to test hypotheses about ratios of population totals and designs where clusters of sampling units are randomized over the treatment combinations. In section 5 these methods are applied to a factorial experiment with advance letters in the Dutch Labor Force Survey (LFS). The paper concludes with a discussion in section 6.

2 Analysis of embedded $K \times L$ factorial experiments

2.1 Experimental designs embedded in probability samples

In a $K \times L$ factorial design, the effects of two factors are tested simultaneously. The first factor, denoted A contains $K \geq 2$ levels. The second factor, denoted B contains $L \geq 2$ levels. The purpose of the experiment is to test the main effects of the two factors and the interactions between both factors on the main parameter estimates of the ongoing survey. To this end a probability sample s of size n is drawn from a finite target population U of size N according the sample design of the regular survey. This sample design can be generally complex, and is described by its first order inclusion probabilities π_i for unit i and second order inclusion probabilities $\pi_{ii'}$ for units i and i'.

Subsequently, this sample is randomly divided into KL subsamples according to a randomized experiment. In the case of a CRD, the sample s of size n is randomly divided into KL subsamples s_{kl}, each with a size of n_{kl} sampling units. The sampling units of each subsample are assigned to one of the KL treatment combinations. Under a CRD, $n_{++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}$ denotes the total number of sampling units in the sample s. The probability that sampling unit i is assigned to subsample s_{kl}, conditionally on the realization of s, equals n_{kl} / n_{++}. The unconditional probability that sampling unit i is selected in subsample s_{kl} equals $\pi_i^* = \pi_i (n_{kl} / n_{++})$.

The power of an experiment might be improved by using sampling structures such as strata, clusters or interviewers as block variables in an RBD since restricted randomization removes the variance between the blocks from the analysis of the experiment (Fienberg and Tanur (1987, 1988)). In the case of an RBD, the sampling units are deterministically grouped in B more or less homogeneous blocks s_b. Within each block, the sampling units are randomly assigned to one of the KL treatment combinations. Let n_{bkl} denote the number of sampling units in block b assigned to treatment combination kl, and

$n_{b++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{bkl}$ the number of sampling units in block b. The probability that sampling unit i is assigned to subsample s_{kl}, conditionally on the realization of s and $i \in s_b$, equals n_{bkl} / n_{b++}, $i \in s_b$. The unconditional probability that sampling unit i is selected in subsample s_{kl} equals $\pi_i^* = \pi_i (n_{bkl} / n_{b++})$.

In many practical applications one of the KL subsamples is assigned to the regular survey and serves, besides being used to produce estimates for the regular publication, as the control group in the experiment. In such situations, the size of this subsample will be substantially larger than the other subsamples.

There are a lot of issues in the planning and design stage of embedded experiments. The field staff, for example, requires special attention, since an embedded experiment can have a large impact on their daily routine of data collection, to which they are accustomed. See van den Brakel and Renssen (1998) and van den Brakel (2008) for more details about such design issues.

Although factorial designs are efficient from a statistical point of view, there might be strong practical arguments against a factorial set-up. The number of treatment combinations increases rapidly with the number of factors in full factorial designs, which might be difficult to implement in the data collection of a survey process. A general solution, known from standard experimental design theory, is to confound higher order interactions with blocks or to apply fractional factorial designs (Hinkelmann and Kempthorne (2005); Montgomery (2001)). These balanced designs, however, are generally hard to combine with the fieldwork restrictions encountered in the daily practice of survey sampling. In many applications the factors that changed in a survey redesign are therefore combined into one treatment. The total effect of these modifications is tested against the standard alternative in a two-treatment experiment. This implies that the effects of all factors in the experiment are confounded and cannot be separately estimated.

2.2 Testing hypotheses about finite population parameters

The purpose of embedded experiments is to test whether alternative survey implementations result in significantly different estimates for finite population parameters. Such differences are the result of non-sampling errors, like measurement errors and response bias. A measurement error model is required to link systematic differences between finite population parameters due to different survey implementations or treatments. Therefore the measurement error model for single-factor experiments proposed by van den Brakel and Renssen (2005) and van den Brakel (2008) is extended to factorial designs.

Let y_{iqkl} denote the observation obtained from the i^{th} individual observed under the kl^{th} treatment combination and the q^{th} interviewer. It is assumed that the observations are a realization of the measurement error model

$$y_{iqkl} = u_i + \beta_{kl} + \gamma_q + \varepsilon_{ikl}. \tag{2.1}$$

Here u_i is the true intrinsic value of the i^{th} individual, β_{kl} the effect of the kl^{th} treatment combination and ε_{ikl} an error component. The model also allows for interviewer effects, *i.e.*, $\gamma_q = \psi + \xi_q$, where ψ denotes a systematic interviewer bias and ξ_q the random effect of the q^{th} interviewer, respectively. Let E_m and cov_m denote the expectation and the covariance with respect to the measurement error model. It

is assumed that $\mathrm{E}_m(\varepsilon_{ikl}) = 0$, $\mathrm{var}_m(\varepsilon_{ikl}) = \sigma^2_{ikl}$, and that measurement errors between sampling units are independent. Furthermore it is assumed that $\mathrm{E}_m(\xi_q) = 0$, $\mathrm{var}_m(\xi_q) = \tau^2_q$ and that random interviewer effects between interviewers are independent. As a result the model allows for correlated response between sampling units that are interviewed by the same interviewer. The measurement error model allows for separate variances for measurement errors under different treatment combinations and separate variances for interviewers.

The treatment effects β_{kl} can be interpreted as the bias in the estimated population parameter if the true intrinsic population value of u is measured by means of the kl^{th} survey implementation. The treatment effect can be decomposed in the traditional way of an analysis of variance for a two-way layout:

$$\beta_{kl} = u + A_k + B_l + AB_{kl}, \tag{2.2}$$

with u the overall effect, A_k and B_l the main effects of treatment factors A and B and AB_{kl} the interactions between treatment factors A and B. If the treatment effects are defined as fixed deviations from the individuals' intrinsic value u_i, then the overall mean u equals zero. In that case A_k corresponds with the bias associated with the k^{th} level of factor A averaged over all levels of factor B, B_l the bias associated with the l^{th} level of factor B, averaged over all levels of factor A, and AB_{kl} the additional bias associated with the combination of the k^{th} level of factor A and the l^{th} level of factor B on top of A_k and B_l.

The following restrictions are required to identify model (2.2):

$$\sum_{k=1}^{K} A_k = 0, \ \sum_{l=1}^{L} B_l = 0, \tag{2.3}$$

and

$$\sum_{k=1}^{K} AB_{kl} = 0, l = 1, 2, \ldots, L, \ \sum_{l=1}^{L} AB_{kl} = 0, k = 1, 2, \ldots, K. \tag{2.4}$$

For each sampling unit, a potential response variable is defined under each of the KL treatment combinations. Therefore the measurement error model can be expressed in matrix notation as:

$$\mathbf{y}_{iq} = \mathbf{j}_{KL}\, u_i + \boldsymbol{\beta} + \mathbf{j}_{KL}\, \gamma_q + \boldsymbol{\varepsilon}_i, \tag{2.5}$$

where $\mathbf{y}_{iq} = (y_{iq11}, ..., y_{iqkl}, ..., y_{iqKL})^t$, $\boldsymbol{\beta} = (\beta_{11}, ..., \beta_{kl}, ..., \beta_{KL})^t$, $\mathbf{j}_{KL}$ a vector of order KL with each element equal to one and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i11}, ..., \varepsilon_{ikl}, ..., \varepsilon_{iKL})^t$. The sampling units are assigned to one of the treatment combinations only, so only one of the responses of $\mathbf{y}_{iq}$ is actually observed. The model assumptions specified above are stated as:

$$\mathrm{E}_m\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{0}, \tag{2.6}$$

$$\mathrm{cov}_m\left(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}\right) = \begin{cases} \boldsymbol{\Sigma}_i & : \ i = i' \\ \mathbf{O} & : \ i \neq i' \end{cases}, \tag{2.7}$$

$$\mathrm{E}_m\left(\xi_q\right) = 0, \tag{2.8}$$

$$\mathrm{cov}_m\left(\xi_q, \xi_{q'}\right) = \begin{cases} \tau_q^2 & : \quad q = q' \\ 0 & : \quad q \neq q' \end{cases}, \tag{2.9}$$

$$\mathrm{cov}_m\left(\varepsilon_{ikl}, \xi_q\right) = 0, \tag{2.10}$$

where $\mathbf{0}$ is a vector of order KL with each element zero, $\boldsymbol{\Sigma}_i$ a matrix of order $KL \times KL$ containing the variances of the measurement errors σ_{ikl}^2, and $\mathbf{O}$ a matrix of order $KL \times KL$ with each element zero.

Let $\bar{\mathbf{Y}} = (\bar{Y}_{11}, ..., \bar{Y}_{1L}, ..., \bar{Y}_{kl}, ..., \bar{Y}_{K1}, ..., \bar{Y}_{KL})^t$ denote the KL dimensional vector of population means of $\mathbf{y}_{iq}$ defined by (2.5). These are the values obtained under a complete enumeration of the finite population under each of the treatment combinations and are defined as:

$$\bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL}\, \psi + \mathbf{j}_{KL} \sum_{q=1}^{Q} \frac{N_q}{N} \xi_q + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varepsilon}_i\,, \tag{2.11}$$

where Q denotes the total number of interviewers available for the data collection and N_q the number of units assigned to the q^{th} interviewer in the case of a complete enumeration.

Only systematic differences between the population parameters that are reflected by the treatment effects $\boldsymbol{\beta}$ should lead to a rejection of the null hypotheses of no treatment effects. This is accomplished by formulating hypotheses about $\bar{\mathbf{Y}}$ in expectation over the measurement error model, *i.e.*,

$$\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL}\, \psi\,. \tag{2.12}$$

Consequently, hypotheses about main effects and interactions are formulated as.

$$\begin{aligned} H_0&: \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{0}, \\ H_1&: \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{2.13}$$

where $\mathbf{C}$ denotes an appropriate contrast matrix, and $\mathbf{0}$ a vector with elements equal to one and a dimension that is equal to the number of contrasts (rows) defined by $\mathbf{C}$. The contrast matrix for the hypothesis about the main effects of factor A is defined as

$$\mathbf{C}_{\mathbf{A}} = \frac{1}{L}\left(\mathbf{j}_{(K-1)} \mid -\mathbf{I}_{(K-1)}\right) \otimes \mathbf{j}_L^t \equiv \frac{1}{L} \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \mathbf{j}_L^t, \tag{2.14}$$

with $\mathbf{I}_{(K-1)}$ the identity matrix of order $K-1$. Matrix $\tilde{\mathbf{C}}_{\mathbf{A}}$ defines the $K-1$ contrasts between the K levels of factor A, averaged over the L levels of factor B. From (2.12) and due to restrictions (2.3) and (2.4) it follows that the contrasts between the population parameters exactly correspond to the contrasts between the main effects of the first factor:

$$\tilde{\mathbf{C}}_{\mathbf{A}} \mathrm{E}_m \bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{A}} \boldsymbol{\beta} = (A_1 - A_2, ..., A_1 - A_K)^t\,.$$

The contrast matrix for the hypothesis about the main effects of factor B is defined as

$$\mathbf{C}_{\mathbf{B}} = \frac{1}{K}\mathbf{j}_K^t \otimes \left(\mathbf{j}_{(L-1)} \mid -\mathbf{I}_{(L-1)}\right) \equiv \frac{1}{K}\mathbf{j}_K^t \otimes \tilde{\mathbf{C}}_{\mathbf{B}}. \tag{2.15}$$

This matrix defines the $L-1$ contrasts between the L levels of factor B, averaged over the K levels of factor A. From (2.12) and due to restrictions (2.3) and (2.4) it follows that the contrasts between the population parameters exactly correspond to the contrasts between the main effects of the second factor:

$$\tilde{\mathbf{C}}_{\mathbf{B}}\mathrm{E}_m\bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{B}}\boldsymbol{\beta} = (B_1 - B_2, ..., B_1 - B_L)^t.$$

The contrast matrices for the main effects use the first level of factors A and B as the reference category. This implies that treatment combination $A_1 \times B_1$ is considered as the control group in the experiment.

Interactions between the two treatment factors are defined as the $L-1$ contrasts of factor B between the $K-1$ contrasts of factor A or, equivalently, as the $K-1$ contrasts of factor A between the $L-1$ contrasts of factor B, Hinkelmann and Kempthorne (1994, chapter 11). Therefore the contrast matrix for the hypothesis about the interactions between factor A and B can be defined as

$$\mathbf{C}_{\mathbf{AB}} = \left(\mathbf{j}_{(K-1)}\Big| - \mathbf{I}_{(K-1)}\right) \otimes \left(\mathbf{j}_{(L-1)}\Big| - \mathbf{I}_{(L-1)}\right) = \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \tilde{\mathbf{C}}_{\mathbf{B}}. \tag{2.16}$$

This matrix contains the $(K-1)(L-1)$ contrasts that define the interactions between factor A and B. The contrasts between the population parameters exactly correspond to the interactions between the first and the second factor, since

$$\begin{aligned}\tilde{\mathbf{C}}_{\mathbf{AB}}\mathrm{E}_m\bar{\mathbf{Y}} = \tilde{\mathbf{C}}_{\mathbf{AB}}\boldsymbol{\beta} = (&AB_{11} - AB_{12} - AB_{21} + AB_{22}, ...,\\ &AB_{11} - AB_{1L} - AB_{21} + AB_{2L}, ...,\\ &AB_{11} - AB_{12} - AB_{K1} + AB_{K2}, ...,\\ &AB_{11} - AB_{1L} - AB_{K1} + AB_{KL})^t.\end{aligned}$$

Each element of this $(K-1)(L-1)$ vector defines one of the $(K-1)(L-1)$ interactions, which neatly corresponds to the contrasts between the interaction effects defined by (2.2). The first element *e.g.*, can be interpreted as the deviation of the treatment effect of the particular combination of factor A at level 2 and factor B at level 2 from the two main effects of these factors.

2.3 Wald test

The hypotheses specified in section 2.2, can be tested with a Wald test (Wald 1943), which is frequently applied in design-based testing procedures, see for example Skinner, Holt and Smith (1989) or Chambers and Skinner (2003). If $\hat{\bar{\mathbf{Y}}}$ denotes a design-unbiased estimator for $\bar{\mathbf{Y}}$, $\mathbf{C}$ the contrast matrix $\mathbf{C}_{\mathbf{A}}$, $\mathbf{C}_{\mathbf{B}}$, or $\mathbf{C}_{\mathbf{AB}}$ defined in (2.14), (2.15) and (2.16), and $\mathrm{cov}(\mathbf{C}\hat{\bar{\mathbf{Y}}})$ the covariance matrix of the contrasts between $\hat{\bar{\mathbf{Y}}}$, then hypotheses can be tested with the Wald statistic $W = \hat{\bar{\mathbf{Y}}}^t\mathbf{C}^t\{\mathrm{cov}(\mathbf{C}\hat{\bar{\mathbf{Y}}})\}^{-1}\mathbf{C}\hat{\bar{\mathbf{Y}}}$.

The GREG estimators, proposed by van den Brakel and Renssen (2005) and van den Brakel (2008) for single-factor experiments are extended to embedded factorial designs in this section. For notational convenience, the subscript q will be omitted in y_{iqkl}, since there is no need to sum explicitly over the interviewer subscript in most of the formulas developed in the rest of this paper.

To apply the model-assisted mode of inference to the analysis of embedded experiments, it is assumed for each unit in the population that the intrinsic value u_i in measurement error model (2.5) is an independent realization of the following linear regression model:

$$u_i = \beta^t \mathbf{x}_i + e_i, \tag{2.17}$$

where $\mathbf{x}_i$ H-vector with auxiliary information, β a H-vector with the regression coefficients and e_i the residuals, which are independent random variables with variance ω_i^2. It is required that all ω_i^2 are known up to a common scale factor, that is $\omega_i^2 = \omega^2 \nu_i$, with ν_i known. The GREG estimator for $\bar{Y}_{kl}$, based on the n_{kl} observations of subsample s_{kl}, is defined as (Särndal *et al.* 1992)

$$\hat{\bar{Y}}_{kl;greg} = \hat{\bar{Y}}_{kl} + \hat{\mathbf{b}}_{kl}^t(\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}), k = 1, 2, \ldots, K, \text{ and } l = 1, 2, \ldots, L, \tag{2.18}$$

where,

$$\hat{\bar{Y}}_{kl} = \frac{1}{N}\sum_{i=1}^{n_{kl}} \frac{y_{ikl}}{\pi_i^*}, \tag{2.19}$$

denotes the HT estimator for $\bar{Y}_{kl}$, $\bar{\mathbf{X}}$ the finite population means of the auxiliary variables $\mathbf{x}$, and $\hat{\bar{\mathbf{X}}}$ the HT estimator for $\bar{\mathbf{X}}$ based on the n_{kl} sample units of subsample s_{kl}. Furthermore,

$$\hat{\mathbf{b}}_{kl} = \left(\sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*}\right)^{-1} \sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}, \tag{2.20}$$

denotes the HT-type estimator for the regression coefficients in (2.17) based on the n_{kl} sampling units in subsample s_{kl}. In (2.19) and (2.20), π_i^* are the first order inclusion probabilities for the sampling units in the KL different subsamples, derived in subsection 2.1. Now $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}} = (\hat{\bar{Y}}_{11;greg}, ..., \hat{\bar{Y}}_{KL;greg})^t$ is an approximately design-unbiased estimator for $\bar{\mathbf{Y}}$ and also for $\mathrm{E}_m \bar{\mathbf{Y}}$ by definition.

Under the null hypotheses that there are no treatment effects and no interactions, it follows that $\mathbf{b}_{kl} = \mathbf{b}_{k'l'}$. In that case, it might be efficient to substitute for $\hat{\mathbf{b}}_{kl}$ in the GREG estimator (2.18) the pooled estimator

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*}\right)^{-1} \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}. \tag{2.21}$$

Since H instead of $KL \times H$ regression coefficients have to be estimated, the pooled estimates of the regression coefficients $\hat{\mathbf{b}}$ will be more precise, particularly in the case of small subsamples. Note,

however, that many commonly used weighting schemes meet the condition that a constant vector λ exists such that $\omega_i^2 = \lambda \mathbf{x}_i$ for all $i \in U$. In this situation the GREG estimator reduces to the simplified form $\hat{\bar{Y}}_{kl;greg} = \hat{\mathbf{b}}_{kl}^t \bar{\mathbf{X}}$ (Särndal *et al.* 1992, section 6.5). Under this simplified form, the treatment effects are completely included in the regression coefficients. In case of the pooled estimator (2.21), the KL GREG estimators are exactly equal by definition, since $\hat{\bar{Y}}_{kl;greg} = \hat{\mathbf{b}}^t \bar{\mathbf{X}}$ for all k and l.

An expression for the covariance matrix of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ where the covariance is taken over the sampling design, the experimental design and the measurement error model, is given by

$$\text{cov}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}) = \text{E}_m \text{E}_s \mathbf{C}\mathbf{D}\mathbf{C}^t, \tag{2.22}$$

where E_s denotes the expectation with respect to the sampling design, and $\mathbf{D}$ a $KL \times KL$ diagonal matrix with diagonal elements

$$d_{kl} = \frac{1}{n_{kl}\left(n_{++}-1\right)} \sum_{i=1}^{n_{++}} \left(\frac{n_{++}\left(y_{ikl} - \mathbf{b}_{kl}^t \mathbf{x}_i\right)}{N\pi_i} - \frac{1}{n_{++}} \sum_{i'=1}^{n_{++}} \frac{n_{++}\left(y_{i'kl} - \mathbf{b}_{kl}^t \mathbf{x}_{i'}\right)}{N\pi_{i'}} \right)^2, \tag{2.23}$$

in the case of a CRD and

$$d_{kl} = \sum_{b=1}^{B} \frac{1}{n_{bkl}\left(n_{b++}-1\right)} \sum_{i=1}^{n_{b++}} \left(\frac{n_{b++}\left(y_{ikl} - \mathbf{b}_{kl}^t \mathbf{x}_i\right)}{N\pi_i} - \frac{1}{n_{b++}} \sum_{i'=1}^{n_{b++}} \frac{n_{b++}\left(y_{i'kl} - \mathbf{b}_{kl}^t \mathbf{x}_{i'}\right)}{N\pi_{i'}} \right)^2, \tag{2.24}$$

in the case of an RBD. An estimator for $\mathbf{D}$ can be derived from the experimental design, conditionally on the measurement error model and the sampling design. Therefore the covariance matrix (2.22) is conveniently stated implicitly as the expectation over the measurement error model and the sampling design. A design-based estimator for this covariance matrix is given by

$$\hat{\text{cov}}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}) = \text{E}_m \text{E}_s \mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t, \tag{2.25}$$

with $\hat{\mathbf{D}}$ a $KL \times KL$ diagonal matrix with elements

$$\hat{d}_{kl} = \frac{1}{n_{kl}(n_{kl}-1)} \sum_{i=1}^{n_{kl}} \left(\frac{n_{++}(y_{ikl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{kl}} \sum_{i'=1}^{n_{kl}} \frac{n_{++}(y_{i'kl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{2.26}$$

in the case of a CRD and

$$\hat{d}_{kl} = \sum_{b=1}^{B} \frac{1}{n_{bkl}(n_{bkl}-1)} \sum_{i=1}^{n_{bkl}} \left(\frac{n_{b++}(y_{ikl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bkl}} \sum_{i'=1}^{n_{bkl}} \frac{n_{b++}(y_{i'kl} - \hat{\mathbf{b}}_{kl}^t \mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{2.27}$$

in the case of an RBD. Proofs for (2.22) and (2.25) are given by van den Brakel (2010) and resemble the derivation of the covariance matrix for single factor experiments, given by van den Brakel and Renssen (2005) and van den Brakel (2008).

The results for (2.22) and (2.25) are obtained under the condition that a constant H-vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$. This is a rather weak condition, since it implies that a weighting model is used that at least uses the size of the finite population as a priori information. See van den Brakel and Renssen (2005) or van den Brakel (2008) for a more detailed discussion.

Since the KL subsamples are drawn without replacement from a finite population, there is a nonzero design covariance between elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$. From that point of view, it is remarkable that (2.25) has a structure as if the subsamples are drawn independently through sampling with replacement using unequal selection probabilities. This gives rise to an attractive variance estimation procedure for embedded experiments, since no design covariances between the subsample estimates appear in (2.25) and no second order inclusion probabilities are required in the variance estimators (2.26) and (2.27). This result is obtained since the covariance matrix of the contrasts between $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ is derived instead of the covariance matrix of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ itself. A detailed interpretation of this result is given by van den Brakel and Renssen (2005) or van den Brakel (2008). See van den Brakel and Binder (2000) and Hidiroglou and Lavallée (2005) for approximations of the covariance matrix of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$.

The design-based estimators $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ and $\hat{\text{cov}}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}})$ can be used to construct a design-based Wald statistic to test the hypotheses described in section 2.2:

$$W = \hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}^t \mathbf{C}^t (\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t)^{-1} \mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}. \tag{2.28}$$

Design-based inferences are generally based on normal large-sample approximations to construct confidence intervals for point estimates or *p*-values and critical regions for test statistics. Under this approach it follows under the null hypothesis that the Wald statistic is asymptotically distributed as a central chi-squared random variable, where the number of degrees of freedom equals the number of contrasts specified in the hypothesis.

The Wald statistic for the hypotheses about the main effects and interactions are given by (2.28) using the contrast matrix $\mathbf{C}_\mathbf{A}$, $\mathbf{C}_\mathbf{B}$, or $\mathbf{C}_\mathbf{AB}$. Under the null hypothesis, it follows that $W \rightarrow \chi^2_{[K-1]}$ for the test about the main effects of factor A, $W \rightarrow \chi^2_{[L-1]}$ for the test about the main effects of factor B and $W \rightarrow \chi^2_{[(K-1)(L-1)]}$ for the test about interactions, where $\chi^2_{[p]}$ denotes a central chi-squared distributed random variable with *p* degrees of freedom.

The Wald test for the main effects can be further simplified. Expressions are developed for the Wald test for the main effects for factor A. Similar expressions can be derived for the main effects of factor B. Denote

$$\begin{aligned} \hat{\bar{\mathbf{Y}}}_{\mathbf{A;GREG}} &= (\hat{\bar{Y}}_{1.;greg}, ..., \hat{\bar{Y}}_{K.;greg})^t, \text{ with } \hat{\bar{Y}}_{k.;greg} = \frac{1}{L}\sum_{l=1}^{L} \hat{\bar{Y}}_{kl;greg}, \\ \hat{\mathbf{D}}_\mathbf{A} &= \text{Diag}(\hat{d}_{1.}, ..., \hat{d}_{K.}), \text{ with } \hat{d}_{k.} = \frac{1}{L^2}\sum_{l=1}^{L} \hat{d}_{kl}. \end{aligned} \tag{2.29}$$

It follows that $\mathbf{C}_\mathbf{A}\hat{\tilde{\mathbf{Y}}}_{\mathbf{GREG}} = \tilde{\mathbf{C}}_\mathbf{A}\hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}}$ and $\mathbf{C}_\mathbf{A}\hat{\mathbf{D}}\mathbf{C}_\mathbf{A}^t = \tilde{\mathbf{C}}_\mathbf{A}\hat{\mathbf{D}}_\mathbf{A}\tilde{\mathbf{C}}_\mathbf{A}^t$. With the matrix inversion lemma, the Wald statistic for the main effects of factor A can be simplified to:

$$\begin{aligned} W &= \hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}}^t\tilde{\mathbf{C}}_\mathbf{A}^t(\tilde{\mathbf{C}}_\mathbf{A}\hat{\mathbf{D}}_\mathbf{A}\tilde{\mathbf{C}}_\mathbf{A}^t)^{-1}\tilde{\mathbf{C}}_\mathbf{A}\hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}} \\ &= \hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}}^t\left(\hat{\mathbf{D}}_\mathbf{A}^{-1} - \frac{1}{\text{Trace}(\hat{\mathbf{D}}_\mathbf{A}^{-1})}\hat{\mathbf{D}}_\mathbf{A}^{-1}\mathbf{j}_{(K-1)}\mathbf{j}_{(K-1)}^t\hat{\mathbf{D}}_\mathbf{A}^{-1}\right)\hat{\tilde{\mathbf{Y}}}_{\mathbf{A;GREG}} \\ &= \sum_{k=1}^{K}\frac{\hat{\tilde{Y}}_{k.;greg}^2}{\hat{d}_{k.}} - \left(\sum_{k=1}^{K}\frac{1}{\hat{d}_{k.}}\right)^{-1}\left(\sum_{k=1}^{K}\frac{\hat{\tilde{Y}}_{k.;greg}^2}{\hat{d}_{k.}}\right)^2. \end{aligned} \tag{2.30}$$

Finally note that the HT estimator (2.19) does not meet the condition that a constant H-vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$. The minimum use of auxiliary information used in the GREG estimator is obtained with a weighting scheme that only uses the size of the finite population as a priori knowledge, *i.e.*, $(x_i) = 1$ and $\omega_i^2 = \omega^2$ (Särndal *et al*. 1992, section 7.4). Under this weighting scheme it follows that

$$\hat{\tilde{Y}}_{kl;greg} = \left(\sum_{i=1}^{n_{kl}}\frac{1}{\pi_i^*}\right)^{-1}\left(\sum_{i=1}^{n_{kl}}\frac{y_{ikl}}{\pi_i^*}\right) \equiv \tilde{y}_{kl}, \tag{2.31}$$

and $(\hat{\mathbf{b}}_{kl}) = \tilde{y}_{kl}$. Expression (2.31) can be recognized as Hájek's ratio estimator for a population mean (Hájek 1971). This weighting scheme satisfies the condition that a constant H-vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$. Therefore an approximately design-unbiased estimator for the covariance matrix of the contrasts between subsample estimates is given by (2.26) and (2.27) for a CRD and an RBD respectively, where $\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i = \tilde{y}_{kl}$. Estimator (2.31) is preferable above the HT estimator (2.19), since (2.31) is more stable and the covariance matrix of the contrasts between (2.31) always has the relatively simple form of (2.25).

2.4 Special cases

It will be shown for two special cases that the design-based Wald statistic is equal to the F-test of a standard analysis of variance. Therefore, an ANOVA-type pooled variance estimator for the diagonal elements of $\hat{\mathbf{D}}$ should be considered as an alternative for (2.26) or (2.27). Such a pooled variance estimator for a CRD is given by

$$\hat{d}_{kl}^p = \frac{1}{n_{kl}(n_{++} - KL)}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\sum_{i=1}^{n_{k'l'}}\left(\frac{n_{++}(y_{ik'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{k'l'}}\sum_{i'=1}^{n_{k'l'}}\frac{n_{++}(y_{i'k'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{2.32}$$

and for an RBD by

$$\hat{d}_{kl}^p = \sum_{b=1}^{B}\frac{1}{n_{bkl}(n_{b++} - KL)}\sum_{k'=1}^{K}\sum_{l'=1}^{L}\sum_{i=1}^{n_{bk'l'}}\left(\frac{n_{b++}(y_{ik'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bk'l'}}\sum_{i'=1}^{n_{bk'l'}}\frac{n_{b++}(y_{i'k'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2. \tag{2.33}$$

Now consider a CRD that is embedded in a self-weighted sample, *i.e.*, $\pi_i = n_{++} / N$, with equally sized subsamples, *i.e.*, $n_{kl} = n_{k'l'} = n_s$. The inclusion probabilities for all units in the KL subsamples are given by $\pi_i^* = n_s / N$. Let $\bar{y} = (1 / n_s)\sum_{i=1}^{n_s} y_{ikl}$. Under Hájek's ratio estimator (2.31) and the pooled variance estimator (2.32) it follows that $\hat{\bar{Y}}_{kl;greg} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, and

$$\hat{d}_{kl}^{\,p} = \frac{1}{n_s(n_{++} - KL)} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_s} \left(y_{ik'l'} - \bar{y}_{k'l'}\right)^2 \equiv \frac{\hat{S}^2_{p;CRD}}{n_s}.$$

The parameter estimates of the K levels of factor A averaged over the L levels of factor B are denoted as

$$\bar{y}_{k.} = \frac{1}{L}\sum_{l=1}^{L} \bar{y}_{kl} = \frac{1}{n_{k+}} \sum_{l=1}^{L} \sum_{i=1}^{n_s} y_{ikl}, k = 1, ..., K, \tag{2.34}$$

with $n_{k+} = \sum_{l=1}^{L} n_{kl}$. The diagonal elements of $\hat{\mathbf{D}}_{\mathbf{A}}$ are now given by

$$\hat{d}_{k.}^{\,p} = \frac{1}{L^2}\sum_{l=1}^{L} \hat{d}_{kl}^{\,p} = \frac{1}{L^2}\sum_{l=1}^{L} \frac{\hat{S}^2_{p;\,\mathrm{CRD}}}{n_s} = \frac{\hat{S}^2_{p;\,\mathrm{CRD}}}{n_{k+}}, k = 1, \ldots, K. \tag{2.35}$$

Let $\bar{y}_{..} = (1 / n_{++})\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_s} y_{ikl}$. Inserting (2.34) and (2.35) into (2.30), gives rise to the following expression for the Wald statistic of the main effects of factor A

$$W = \frac{1}{\hat{S}^2_{p;CRD}} \left(\sum_{k=1}^{K} n_{k+} \bar{y}^2_{k.} - n_{++} \bar{y}^2_{..} \right). \tag{2.36}$$

Note that $W / (K - 1)$ in (2.36) corresponds with the F-statistic for the main effects of an analysis of variance for the two-way layout with interactions (Scheffé 1959, chapter 4). Under the null hypothesis and the assumption of normally and independently distributed errors, the F-statistic in the two-way layout follows an F-distribution with $(K - 1)$ and $(n_{++} - KL)$ degrees of freedom, which is denoted as $F^{[K-1]}_{[n_{++}-KL]}$. If $n_{++} \to \infty$, then $F^{[K-1]}_{[n_{++}-KL]} \to \chi^2_{[K-1]} / (K - 1)$. Consequently the F-statistic and the Wald statistic have the same limit distribution.

Now consider an RBD that is embedded in a self-weighted sampling design with equal subsample sizes, thus $\pi_i = n_{+++} / N$ and $n_{kl} = n_{k'l'} = n_s$, with $n_{+++} = \sum_{b=1}^{B} n_{b++}$. Let $\bar{y}_{bkl} = (1 / n_{bkl})\sum_{i=1}^{n_{bkl}} y_{ikl}$. Furthermore, it is assumed that the fraction of sampling units assigned to each treatment combination within each block is equal, *i.e.*, $n_{bkl} / n_{b++} = n_s / n_{+++}$, and that the block sizes are sufficiently large to assume that $n_{b++} / (n_{b++} - KL) \approx 1$. Under Hájek's ratio estimator (2.31) and the pooled variance estimator (2.33) it follows that $\hat{\bar{Y}}_{kl;greg} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, and

$$\hat{d}_{kl}^{\,p} = \sum_{b=1}^{B} \frac{1}{n_{bkl}(n_{b++} - KL)} \left(\frac{n_{b++}}{n_{+++}} \right)^2 \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(y_{ik'l'} - \bar{y}_{bk'l'} \right)^2$$
$$\approx \frac{1}{n_s n_{+++}} \sum_{b=1}^{B} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(y_{ik'l'} - \bar{y}_{bk'l'} \right)^2 \equiv \frac{\hat{S}^2_{p;RBD}}{n_s}.$$

The parameter estimates of the K levels of factor A averaged over the L levels of factor B and the blocks are denoted as

$$\bar{y}_{.k.} = \frac{1}{L} \sum_{l=1}^{L} \bar{y}_{kl} = \frac{1}{n_{+k+}} \sum_{b=1}^{B} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}, k = 1, \ldots, K, \tag{2.37}$$

where $n_{+k+} = \sum_{b=1}^{B} \sum_{l=1}^{L} n_{bkl}$. The diagonal elements of $\hat{\mathbf{D}}_{\mathbf{A}}$ are given by

$$\hat{d}_{k.}^{\,p} = \frac{1}{L^2} \sum_{l=1}^{L} \hat{d}_{kl}^{\,p} = \frac{\hat{S}^2_{p;\,\mathrm{RBD}}}{n_{+k+}}, k = 1, \ldots, K. \tag{2.38}$$

Let $\bar{y}_{...} = (1 / n_{+++}) \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}$. If these results are inserted into (2.30), then the expression for the Wald statistic of the main effects of factor A can be simplified to

$$W = \frac{1}{\hat{S}^2_{p;RBD}} \left(\sum_{k=1}^{K} n_{+k+} \bar{y}^2_{.k.} - n_{+++} \bar{y}^2_{...} \right). \tag{2.39}$$

It can be recognized that $W / (K - 1)$ in (2.39) corresponds with the F-statistic for the main effects of an analysis of variance for the three-way layout with interactions, (Scheffé 1959, chapter 4). As in the case of a CRD, this Wald and F-statistic have the same limit distribution.

3 Factorial designs with more than two factors

The results developed for $K \times L$ factorial designs are extended to designs with more than two factors. A more general notation for the treatment factors is introduced first. Let A_g denote the g^{th} treatment factor in the experiment with levels $a_g = 1, ..., M_g$. In the general case there are g = 1,...,G factors included in the experiment. The population parameters observed under the $M_1 M_2 ... M_G$ treatment combinations are collected in the vector $\bar{\mathbf{Y}} = (\bar{Y}_{11...1}, ..., \bar{Y}_{a_1 a_2 ... a_G}, ..., \bar{Y}_{M_1 M_2 ... M_G})^t$. The index for the levels of a factor runs within each level of its preceding factor. Thus index a_g runs from $a_g = 1, ..., M_g$ within each level of $a_{(g-1)}$. Hypotheses about the main effects and interactions are, as motivated in section 2.2, formulated about $\bar{\mathbf{Y}}$ in expectation over the measurement error model.

The contrast matrices for the main effects and interactions in (2.13) are developed for the general case of a $M_1 \times M_2 \times ... \times M_G$ factorial design. Let $\mathcal{A} = \{1, ..., G\}$ denote the set of labels for the factors and $\tilde{\mathbf{C}}_{\mathbf{A_g}} = (\mathbf{j}_{(M_g - 1)} \mid -\mathbf{I}_{(M_g - 1)})$. The following three functions are defined first:

$$\mathbf{J}_{1_g} = \begin{cases} \mathbf{j}^t_{M_1} \otimes ... \otimes \mathbf{j}^t_{M(g-1)} & : \ g > 1 \\ 1 & : \ g = 1 \end{cases},$$

$$\mathbf{J}_{2_g} = \begin{cases} \mathbf{j}^t_{M(g+1)} \otimes ... \otimes \mathbf{j}^t_{M_G} & : \ g < G \\ 1 & : \ g = G \end{cases},$$

$$\mathbf{J}_{3_{g,g'}} = \begin{cases} \mathbf{j}^t_{M(g+1)} \otimes ... \otimes \mathbf{j}^t_{M(g'-1)} & : \ g' - g > 1 \\ 1 & : \ g' = g + 1 \end{cases}.$$

The main effect of factor A_g is defined as the $M_g - 1$ contrasts between the M_g levels, averaged over the levels of the other $G - 1$ factors and is given by:

$$\mathbf{C}_{\mathbf{A}_{g_1}} = \left(\prod\nolimits_{g \in \mathcal{A} \setminus \{g_1\}} M_g \right)^{-1} \mathbf{J}_{1_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \mathbf{J}_{2_{g_1}}, g_1 = 1, \ldots, G.$$

Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{2_{g_1}}$ sums over the levels of the factors $A_{(g_1+1)}...A_G$ that are nested within each level of A_{g_1}.Subsequently, $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ defines the $M_{g_1} - 1$ contrasts between the levels of A_{g_1} that are nested within each combination of the levels of $A_1...A_{(g_1-1)}$. Premultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{1_{g_1}}$ adds the contrast matrices $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ that are nested within all combinations of the levels of $A_1...A_{(g_1-1)}$.

The interaction between A_{g_1} and A_{g_2} is defined as the $M_{g_2} - 1$ contrasts of factor A_{g_2} between the $M_{g_1} - 1$ contrasts of A_{g_1} averaged over the levels of the other $G - 2$ factors and is given by:

$$\mathbf{C}_{\mathbf{A}_{g_1}\mathbf{A}_{g_2}} = \left(\prod\nolimits_{g \in \mathcal{A} \setminus \{g_1, g_2\}} M_g \right)^{-1} \mathbf{J}_{1_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_1}} \otimes \mathbf{J}_{3_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \mathbf{J}_{2_{g_2}},$$
$$g_1 = 1, \ldots, G-1, g_2 = 1, \ldots, G, g_1 < g_2.$$

Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_2}}$ by $\mathbf{J}_{2_{g_2}}$ adds the levels of the factors $A_{(g_2+1)}...A_G$ that are nested within each level of A_{g_2}. $\tilde{\mathbf{C}}_{\mathbf{A}_{g_2}}$ defines the contrasts of the main effect of factor A_{g_2} which are nested within each combination of the levels of $A_1...A_{(g_2-1)}$. Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{3_{g_1,g_2}}$ sums the contrast matrices $\tilde{\mathbf{C}}_{\mathbf{A}_{g_2}}$ over the levels of $A_{(g_1+1)}...A_{(g_2-1)}$ that are nested within each combination of the levels of $A_1...A_{g_1}$. Premultiplication of $\mathbf{J}_{3_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g_2}} \otimes \mathbf{J}_{2_{g_2}}$ with $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ defines the contrasts of the interactions between A_{g_1} and A_{g_2}, within each combination of the levels of $A_1...A_{(g_1-1)}$. Finally, Premultiplication of $\tilde{\mathbf{C}}_{\mathbf{A}_{g_1}}$ by $\mathbf{J}_{1_{g_1}}$ sums the contrasts of the interactions between A_{g_1} and A_{g_2} over the levels of $A_1...A_{(g_1-1)}$.

The interaction between A_{g_1}, A_{g_2} and A_{g_3} is defined as the $M_{g_3} - 1$ contrasts of factor A_{g_3} between the interactions of A_{g_1} and A_{g_2}, averaged over the levels of the other $G - 3$ factors. This process expands in a similar way to higher order interactions, which results in the following definitions of the higher order interactions:

$$\mathbf{C}_{\mathbf{A}_{g1}\mathbf{A}_{g2}\mathbf{A}_{g3}} = \left(\prod_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3\}} M_g\right)^{-1} \mathbf{J}_{1_{g1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g1}} \otimes \mathbf{J}_{3_{g1,g2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g2}} \otimes \mathbf{J}_{3_{g2,g3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g3}} \otimes \mathbf{J}_{2_{g3}},$$

$$g_1 = 1,\ldots,G-2, g_2 = 2,\ldots,G-1, g_3 = 3,\ldots,G, g_1 < g_2 < g_3,$$

$$\mathbf{C}_{\mathbf{A}_{g1}\mathbf{A}_{g2}\mathbf{A}_{g3}\mathbf{A}_{g4}} = \left(\prod_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3,g_4\}} M_g\right)^{-1} \mathbf{J}_{1_{g1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g1}} \otimes \mathbf{J}_{3_{g1,g2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g2}} \otimes \mathbf{J}_{3_{g2,g3}} \otimes$$

$$\tilde{\mathbf{C}}_{\mathbf{A}_{g3}} \otimes \mathbf{J}_{3_{g3,g4}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g4}} \otimes \mathbf{J}_{2_{g3}},$$

$$g_1 = 1,\ldots,G-3, g_2 = 2,\ldots,G-2, g_3 = 3,\ldots,G-1,$$

$$g_4 = 4,\ldots,G, g_1 < g_2 < g_3 < g_4,$$

$$\vdots$$

$$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\ldots\mathbf{A}_G} = \tilde{\mathbf{C}}_{\mathbf{A}_{g1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g3}} \otimes \ldots \otimes \tilde{\mathbf{C}}_{\mathbf{A}_{g4}}$$

The number of rows of each contrast matrix coincides with the number of contrasts that define the various main effects and interactions. The number of columns of these matrices equals $M_1M_2...M_G$.

These contrast matrices are inserted in (2.13) to define the various hypotheses about the main effects and interactions between the G treatment factors. The sampling units in the initial sample are randomly divided over all possible treatment combinations according to a CRD or an RBD, resulting in $M_1M_2...M_G$ different subsamples. Let $n_{a_1...a_G}$ denote the number of sampling units assigned to treatment combination $a_1...a_G$ in subsample $s_{a_1...a_G}$ and $n_{+...+}$ the size of the initial sample. In the case of a CRD, the first order inclusion probabilities for the units in subsample $s_{a_1...a_G}$ are now given by $\pi_i^* = \pi_i(n_{a_1...a_G} / n_{+...+})$. In the case of an RBD, the first order inclusion probabilities for the units in subsample $s_{a_1...a_G}$ are given by $\pi_i^* = \pi_i(n_{ba_1...a_G} / n_{b+...+})$ where $n_{ba_1...a_G}$ denotes the number of sampling units assigned to treatment combination $a_1...a_G$ in block b and $n_{b+...+}$ the total number of sampling units in block b.

Now $\hat{\bar{Y}}_{a_1...a_G;greg}$ denotes the GREG estimator for $\bar{Y}_{a_1...a_G}$ based on the observations obtained in subsample $s_{a_1...a_G}$ and is defined analogously to expression (2.18). These $M_1M_2...M_G$ GREG estimators are collected in the vector $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}} = (\hat{\bar{Y}}_{1...1;greg},...,\hat{\bar{Y}}_{M_1...M_G;greg})^t$ and is an approximately design-unbiased estimator for $\bar{\mathbf{Y}}$ and $\mathrm{E}_m\bar{\mathbf{Y}}$. Design-based estimators for the covariance matrices of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ are defined by (2.25), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (2.26) in the case of a CRD or (2.27) in the case of an RBD.

Finally hypotheses about main effects and interactions are tested with the Wald statistic (2.28), which is asymptotically distributed as a chi-squared random variable where the number of degrees of freedom equals the number of contrasts specified in the various hypotheses. As an example, the contrast matrices of the main effects and interactions in a factorial design with four factors are given in Table 3.1.

Table 3.1
Contrasts in a $M_1 \times M_2 \times M_3 \times M_4$ factorial design

Contrast matrix	Number of contrasts (degrees of freedom)
$\mathbf{C}_{\mathbf{A}_1} = 1/\left(M_2M_3M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_1 - 1$
$\mathbf{C}_{\mathbf{A}_2} = 1/\left(M_1M_3M_4\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_2 - 1$
$\mathbf{C}_{\mathbf{A}_3} = 1/\left(M_1M_2M_4\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$M_3 - 1$
$\mathbf{C}_{\mathbf{A}_4} = 1/\left(M_1M_2M_3\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$M_4 - 1$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2} = 1/\left(M_3M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_2 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_3} = 1/\left(M_2M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_4} = 1/\left(M_2M_3\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_3} = 1/\left(M_1M_4\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_2 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_4} = 1/\left(M_1M_3\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_2 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_1M_2\right)\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3} = 1/\left(M_4\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \mathbf{j}_{M_4}^t$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_3 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_4} = 1/\left(M_3\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_2\right)\tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4} = 1/\left(M_1\right)\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_2 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$
$\mathbf{C}_{\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4} = \tilde{\mathbf{C}}_{\mathbf{A}_1} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_2} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_3} \otimes \tilde{\mathbf{C}}_{\mathbf{A}_4}$	$\left(M_1 - 1\right)\left(M_2 - 1\right)\left(M_3 - 1\right)\left(M_4 - 1\right)$

4 Further extensions

So far, experimental designs are considered where the ultimate sampling units of the sampling design are randomized over the treatments. Owing to restrictions in the field work there might be practical reasons to randomize clusters of sampling units over the different treatments, at the cost of reduced power for testing hypotheses about treatment effects. It might for example be attractive to assign the sampling units that belong to the same household or are assigned to the same interviewer to the same treatment combination. In van den Brakel (2008) a design-based analysis procedure is developed for single-factor experiments designed as CRDs and RBDs where clusters of sampling units are randomized over the

treatments. These methods directly extend to the analysis of the factorial designs that are considered in this paper.

Consider the general case of a $M_1 \times M_2 \times ... \times M_G$ factorial design. The clusters of sampling units in the initial sample are randomized over the different treatment combinations. The conditional probability that a sampling unit is assigned to a subsample is now derived from the fractions of clusters that are assigned to the different treatment combinations within the sample or within each block. See van den Brakel (2008) for details. The GREG estimator for $\bar{Y}_{a_1 \ldots a_G}$ is defined analogously to expression (2.18). Design-based estimators for the covariance matrices of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$ are defined by (2.25), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (4.6) in van den Brakel (2008), which is based on the variance between the estimated cluster totals.

The target parameters of a survey are often defined as a ratio of two population totals. In van den Brakel (2008) a design-based analysis procedure is developed to test hypotheses about ratios in single-factor experiments designed as a CRD or an RBD. These results can be extended to the analysis factorial designs treated in this paper. Based on each subsample a ratio of two GREG estimators can be constructed for each treatment combination. Design-based estimators for the covariance matrices of the contrasts between the ratios are defined by (2.25), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (4.11) in van den Brakel (2008), which is an estimator for the variance of the ratio of two GREG estimators. Hypotheses about main effects and interactions are tested with the Wald statistic (2.28).

5 Testing new advance letters for the Dutch Labor Force Survey

In this section an experiment with different advance letters embedded in the Dutch Labor Force Survey (LFS) is described, which serves as a numerical example to illustrate the methodology developed in this paper.

5.1 Survey design

The LFS is based on a rotating panel survey. Each month a stratified two-stage cluster sample of about 6,000 addresses is drawn from a register of all known addresses in the Netherlands. Strata are formed by geographical regions, municipalities are considered as primary sampling units, and addresses as secondary sampling units. All households residing at an address, with a maximum of three, are included in the sample. In the first wave, data are collected by means of computer assisted personal interviewing. The respondents are re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing.

The weighting procedure of the LFS is based on the GREG estimator of Särndal *et al.* (1992). The inclusion probabilities reflect the sample design used to select households as well as the different response rates between geographical regions. The weighting scheme is based on a combination of different socio-demographic categorical variables. One of the most important parameters of the LFS is the unemployed labor force, which is defined as the ratio of the total unemployment and the total labor force.

5.2 Experimental design

Advance letters are one of the design parameters of a survey that affect response rates and cooperation of respondents (De Leeuw, Callegaro, Hox, Korendijk and Lensveit-Mulders (2007)). The standard advance letter of the LFS is addressed to the occupants of the accommodation and the tone is formal and high-handed. As a result, this letter does not conform to social psychological theories regarding survey participation proposed by Groves, Cialdini and Couper (1992) and Groves and Couper (1998). In an attempt to improve the LFS response rates, Luiten, Campanelli, Klaasen and Beukenharst (2008) proposed different advance letters for the LFS that better meet these principles about survey participation. The effects of these alternative letters are investigated empirically by means of a large-scale field experiment embedded in the LFS.

The first factor considered in this experiment, say A, concerns the salutation of the respondent on two levels, *i.e.*, the standard approach where the letter is addressed to the occupants of the accommodation (A_1) versus a named letter (A_2). It is anticipated that named letters are more likely to be read and therefore increase response rates and survey participation. The second factor, say B, concerns the content of the letter on three levels, *i.e.*, the standard formal letter (B_1) versus two alternative letters (B_2 and B_3). In the first alternative, the content of the standard letter is adapted by explaining why the survey is conducted, what the respondent gains by participating and why it is important for Statistics Netherlands that the respondent participates in the survey. The second alternative attempts to improve the formal tone of the standard letter. The three versions of the advance letters can be found in van den Brakel (2010).

A new letter is only considered for implementation as a standard in the LFS, if its positive effect on response behavior has been demonstrated and if its effect on the main parameter estimates is quantified in a randomized experiment. Both factors are tested in a 2×3 factorial design resulting in six treatment combinations. This experiment is embedded in the first wave of the LFS for a period of five months (December 2007 through April 2008). During this period the monthly gross sample size is randomized over six subsamples according to an RBD with interviewers as the block variables. About 220 interviewers were available for the field work. In the analysis, adjacent interviewer regions were collapsed into 13 blocks. A fraction of 0.8 of the sample is assigned to the regular advance letter, *i.e.*, treatment combination $A_1 \times B_1$. A fraction of 0.04 of the sample is assigned to each of the other five alternative treatment combinations.

The allocation of the sampling units over the treatments is predominantly based on practical arguments. Embedding experiments in ongoing sample surveys serves two competing purposes. To estimate official figures as precisely as possible it is beneficial to allocate as many sampling units as possible to the control group, since this subsample is also used for regular publication purposes. To estimate the contrasts in the experiment as precisely as possible it is, on the other hand, beneficial to divide the total sample equally over the different treatment combinations. In this application it was decided that a loss of at most 20% of the sample size for regular publication purposes could be tolerated. This led to the aforementioned allocation over the treatment combinations. Under a response rate of 56% and a monthly sample size of 6,000 households it is expected that about 13,440 households are observed in the control group $A_1 \times B_1$ and 670 households in each of the alternative treatment combinations.

Although the allocation is based on practical considerations, it is important to have a notion of the power of the planned experiment. The target variable analyzed in this paper is the unemployed labor force,

expressed as a percentage. Ignoring the block design of this experiment, it follows that the variance of the treatments equals to $\hat{d}_{kl} = \hat{S}_{kl}^2 / n_{kl}$, where $\hat{S}_{kl}^2$ is implicitly defined by (2.26). It is assumed that $\hat{S}_{kl}^2$ is equal to say $\hat{S}^2$ for each treatment combination. With available sample data it follows for the unemployed labor force that $\hat{S}^2 = 285$. Now the minimal observable difference for a contrast that would reject the null hypothesis under a pre-specified significance and power level equals

$$\Delta = \sqrt{\text{var}(\Delta)}(Z_{(1-\alpha/2)} + Z_{(1-\beta)}), \tag{5.1}$$

where $Z_{(\gamma)}$ denotes the γ^{th} percentile point of the standard normal distribution, α the significance level of the test and $(1-\beta)$ the power. The main effect of factor A concerns one contrast $\hat{\Delta}_A = (\hat{\bar{Y}}_{1.;greg} - \hat{\bar{Y}}_{2.;greg})$. From (2.29) it follows that the variance of this contrast equals $\hat{\text{var}}(\hat{\Delta}_A) = (\hat{S}^2 / 9)\sum_{l=1}^{3}(1/n_{1l} + 1/n_{2l})$. The main effect of factor B concern two contrasts $\hat{\Delta}_{B_l} = (\hat{\bar{Y}}_{.1;greg} - \hat{\bar{Y}}_{.l;greg})$, $l = 2,3$ with variances $\hat{\text{var}}(\hat{\Delta}_{B_l}) = (\hat{S}^2 / 4)\sum_{k=1}^{2}(1/n_{k1} + 1/n_{kl})$, $l = 2,3$. The interactions between factors A and B concern the two contrasts $\hat{\Delta}_{AB_l} = (\hat{\bar{Y}}_{11;greg} - \hat{\bar{Y}}_{1l.;greg} - \hat{\bar{Y}}_{21;greg} + \hat{\bar{Y}}_{2l.;greg})$ with variances $\hat{\text{var}}(\hat{\Delta}_{AB_l}) = \hat{S}^2(1/n_{11} + 1/n_{1l} + 1/n_{21} + 1/n_{2l})$, $l = 2,3$.

Inserting the variances of the different contrasts in (5.1), gives minimum values of differences that would reject the null hypothesis for main effects and interactions for pre-specified sample sizes, significance levels and power levels. In Table 5.1 these differences for the unemployed labor force are calculated for the aforementioned applied allocation, and a balanced design where the sample size for each treatment combination is equal to 2,800. Values are given for unspecified alternative hypotheses at a 5% significance level and a power of 50%, 80% and 90%. In experimental design theory, 80% is a widely accepted power level by sample size determination. In survey sampling minimum sample size requirements are generally based on significance level requirements only, which corresponds to a power level of 50%. Differences are specified for separate tests of the contrasts. The main effect of factor B and the interaction effects both contain two contrasts. To preserve an overall significance level of 5%, differences for both tests are also calculated using Bonferroni's simultaneous comparison procedure.

Table 5.1 illustrates different aspects of embedded experiments and factorial designs. First it illustrates the cost-benefits of a factorial setup. Twice as many experimental units are required if the main effects of both factors are tested at the same precision in two separate single factor experiments. Table 5.1 also shows that the power for the test of interactions is much smaller than for the tests of the two main effects. The more treatment factors that are combined in one experiment, the smaller the sample size allocated to each treatment combination and the smaller the power for the tests of interactions. This puts the often cited advantage that factorial designs also allow testing of interactions between the different treatment factors into perspective. In practice, sample sizes are based on power calculations for the tests on the main effects. Consequently, only large interactions can be detected with sufficient power. A factorial design still has the advantage that the validity of observed main effects increases, since they are tested over a wider range of conditions.

Table 5.1
Observable difference for the unemployed labor force in percentages at 5% significance levels and different power levels

Contrast	Number of contrasts	Power separate t-test			Power Bonferroni t-test		
		50%	80%	90%	50%	80%	90%
Applied design							
Main effect A	1	0.96	1.36	1.58	0.96	1.36	1.58
Main effect B	2	1.12	1.59	1.85	1.27	1.75	2.00
Interaction	2	2.23	3.19	3.69	2.55	3.51	4.00
$A_1 \times B_1 - A_k \times B_l$	5	1.31	1.87	2.17	1.72	2.28	2.57
Balanced design							
Main effect A	1	0.51	0.73	0.84	0.51	0.73	0.84
Main effect B	2	0.63	0.89	1.03	0.71	0.98	1.12
Interaction	2	1.25	1.79	2.07	1.43	1.97	2.25
$A_1 \times B_1 - A_k \times B_l$	5	0.88	1.26	1.46	1.16	1.54	1.74

If the null hypothesis of no interactions is rejected, then main effects are difficult to interpret. In that situation it is more useful to compare the control group, *i.e.*, $A_1 \times B_1$, with the five alternative treatment combinations. The minimum observable differences of these five contrasts that reject the null hypothesis at a 5% significance level and different power levels are also included in Table 5.1.

Comparing minimum values for the differences under the applied design and the balanced design, illustrates the loss of power if an extreme skew allocation over the treatment combinations is chosen. Minimizing the risk of losing too much precision for the regular publication is the motivation behind the choice for this allocation. It clearly illustrates the duality of combining two competing purposes in an embedded experiment; estimation for the regular publication purposes versus testing contrasts of different treatment combinations.

To assess the value of the results that can be obtained with this experiment, the minimum observable differences with this experiment are related to the standard errors of the regular survey estimates. Standard errors for the survey estimates at the national level will generally be much smaller than the minimum observable differences with an experiment since the sample size allocated to the alternative treatments is generally much smaller than the regular sample size. If, however, the assumption is adopted that differences observed with an experiment at the national level also apply to the survey estimates for important domains, then the differences observable with the experiment might become comparable with the standard errors of these domain estimates. This assumes no interaction between domains and treatment effects. The standard errors for the monthly unemployed labor force figures at the national level equals 0.15 percent points. The standard errors for the domains vary between 0.3 and 1.0 percent points. Comparing these standard errors with the differences in Table 5.1 shows that the main effects are still larger than the standard errors at the national level but become comparable with the precision of the regular monthly domain estimates.

5.3 Results

Table 5.2 contains an overview of the response rates of the households in the six subsamples of the experiment. It follows that the different advance letters result in relatively small differences in the response rates. Factor A results in an increase of the response of 2.4 percent points by using a personalized letter (after correcting proportions for the unbalanced allocation of the sample over the treatment combinations). The alternative letters considered in factor B resulted in a decrease of 1.5 percent points (alternative B_2) and 1.9 percent points (alternative B_3).

Table 5.2
Response rates experiment with advance letters

Treatment	Response		Refusal		Rest		Total
$A_1 \times B_1$	13,234	56.69%	5,127	21.96%	4,985	21.35%	23,346
$A_1 \times B_2$	604	53.59%	271	24.05%	252	22.36%	1,127
$A_1 \times B_3$	635	56.34%	254	22.54%	238	21.12%	1,127
$A_2 \times B_1$	662	59.00%	256	22.82%	204	18.18%	1,122
$A_2 \times B_2$	663	59.09%	236	21.03%	223	19.88%	1,122
$A_2 \times B_3$	627	55.64%	259	22.98%	241	21.38%	1,127

Response behavior is modeled in a logistic regression model to test hypotheses about the effect of the two treatment factors. This is a typical conditional analysis that does not account for sample design features like unequal selection probabilities and clustering of households within municipalities. Clustering induced by the two-stage sample design is ignored, since households are randomized over the treatments in the experiment. In this logistic regression analysis interest is focussed on differences in the observed sample, in this case due to differences in selective non-response. This gives additional information on whether the factors increase the response across the entire target population or that specific groups react differently to the treatments. Second and higher order interactions between the two treatment factors and socio-demographic categorical variables in the logistic regression model indicate that the variation in response between different subpopulations increases and that they react differently to the treatments.

In the logistic regression model, the dependent binary variable indicates whether a household completely responded versus the remaining response categories. The response behavior is assumed to depend upon:

- a general mean,
- treatment factors A (name) and B (content),
- a block variable in 13 categories,
- auxiliary variables:
  - urbanization level at five categories,

- gender in three categories, specifying whether a household consists of men only, women only, or a mixture of men and women,
- age as a quantitative variable containing the average age of the household members,
- ethnicity in seven categories, specifying household compositions of native, western background, non-western background, and all possible mixtures,
- family composition in four categories: partners, single-parent family, single, and a remainder category.

All third order interactions between the variables are initially considered for backward model selection. The final selected model contains the terms that are given in the first column of Table 5.3. For brevity, the regression coefficients with their standard errors and test statistics for separate categories are only expressed for the treatment factors. The hypothesis that there are no interactions between the two treatment factors cannot be rejected (p -value Wald statistic equals 0.121). From Table 5.3 it follows that factor A, *i.e.*, using a letter addressed to a named individual, has a positive but non-significant effect on the response rate. Factor B, *i.e.*, two alternative letters with an improved content, has even a slightly negative but non-significant effect on the response rates. This is a remarkable result, since the two alternative letters attempt to improve the formal tone of the standard letter, but in line with the results of an earlier experiment where the response to a more informal advance letter for the LFS also resulted in significantly smaller response rates (van den Brakel 2008). Since there are no interactions between the treatment factors and the auxiliary variables, there are also no indications that the treatment factors induce the response of specific subpopulations.

Table 5.3
Logistic regression analysis for response rates

Parameter	Coefficient	Standard error	Wald statistic	D.f.	p -value
Mean	0.287	0.078	13.604	1	0.000
Block			212.425	12	0.000
Treatment A (name, A_2)	0.083	0.045	3.394	1	0.065
Treatment B (content)			2.965	2	0.227
Alternative 1 (B_2)	-0.046	0.051	0.816	1	0.366
Alternative 2 (B_3)	-0.083	0.051	2.678	1	0.102
Urbanization			16.589	4	0.002
Ethnic			127.734	6	0.000
Gender			48.076	2	0.000
Family composition			27.339	3	0.000

In the second step of this analysis it is tested whether the estimates for the unemployed labor force obtained with the six subsamples under the different advance letters are significantly different. The design-based analysis procedure developed in this paper is used to account for the sampling design, the

experimental design and the estimation procedure of the LFS. The GREG estimator is applied to obtain estimates for the unemployed labor force under the six different treatment combinations. With this unconditional analysis it is tested whether the different advance letters introduce differences in selection bias, after correcting for the differences in response rates using the design-based estimation procedure applied in the regular LFS.

With this analysis, the linear measurement error model (2.1) is applied to a binary response variable. This might appear to be ridged, since logistic models are more natural in this case. Under the model-assisted approach linear regression models, however, are frequently applied to derive a GREG estimator for binary response variables. Also in the Dutch LFS a linear regression model is assumed to derive a GREG estimator for official labor force figures. To develop a design-based analysis procedure for embedded experiments that also account for the GREG estimator used in the regular survey, a linear measurement error model is assumed in a similar way. A detailed discussion about the use and interpretation of a linear measurement error model applied to binary response variables is given by van den Brakel (2008).

The inclusion probabilities in the GREG estimator (2.18) reflect the sampling design of the LFS and the experimental design used to divide the initial sample into six subsamples. The following weighting scheme was applied to calibrate the design weights: *age* + *region* + *marital status* + *gender* + *urbanization level*, where the five variables are categorical. This is a reduced version of the regular weighting scheme of the LFS.

The estimation results for the six subsamples are summarized in Table 5.4, where the unemployed labor force is expressed in percentages. It appears that there are no systematic patterns between subsample estimates. The subsample estimates and their variance estimates indicate that there are no significant differences between the control group and the five alternative treatment combinations. Finally the main effects and the interaction effects of the two treatment factors are tested, taking into account that the experiment was designed as an RBD where adjacent interviewer regions are collapsed in 13 blocks. The analysis results are summarized in Table 5.5.

Table 5.4
Point estimates and standard errors unemployed labor force (expressed in percentages)

Treatment combination		Estimate $\hat{\bar{Y}}_{kl;greg}$	Standard error $\sqrt{\hat{d}_{kl}}$
$k\,(A_k)$	$l\,(B_l)$		
1	1	4.100%	0.145%
1	2	3.761%	0.646%
1	3	5.264%	0.753%
2	1	3.609%	0.608%
2	2	4.546%	0.666%
2	3	3.385%	0.664%

Table 5.5
Analysis main effects and interactions unemployed labor force (expressed in percentages)

Source	Estimate $C\hat{\bar{Y}}_{greg}$	Wald statistic	D.f.	p -value
Treatment A (name) $A_1 - A_2$	0.528	1.109	1	0.292
Treatment B (content)		0.732	2	0.694
$B_1 - B_2$	-0.300			
$B_1 - B_3$	-0.471			
Interaction		3.801	2	0.150
$AB_{11} - AB_{12} - AB_{21} + AB_{22}$	1.276			
$AB_{11} - AB_{13} - AB_{21} + AB_{23}$	-1.388			

From the analysis results, summarized in Table 5.5, it can be concluded that there are no indications that the different advance letters result in different parameter estimates. This is in line with the analysis results of the response rates. Since there is no empirical evidence that the different advance letters affect response rates of the entire population or a subpopulation, it might be expected that no significant differences between the parameter estimates occur.

There are no indications that the alternative letters, considered in this experiment, improve response behavior or result in systematic effects in the estimates for target variables like the unemployed labor force. Therefore it was decided not to adapt the standard advance letter of the LFS.

6 Discussion

In factorial designs the levels of two or more treatment factors are varied and all possible treatment combinations are considered simultaneously. These designs are widely used in scientific experimentation for several reasons. The main effects of the factors are averaged over the levels of the other factors. Conclusions about the various effects are therefore based on a wider range of conditions, which increases the validity of the results. Furthermore, interaction between the different treatment factors can be analyzed, although the power of these tests decreases as the number of factors that are combined in one experiment increases. Finally factorial designs are more efficient compared to single-factor experiments, since fewer experimental units are required to estimate the main effects with the same precision.

In this paper a design-based theory is developed for the analysis of factorial designs that are embedded in probability samples. This approach is particularly appropriate to quantify the effects of the different design parameters of a survey process on the parameter estimates of a sample survey. Applications can be found in total survey design, empirical research into survey practice and quantifying discontinuities in series of repeatedly conducted surveys. Design-based analysis procedures are developed to test hypotheses about population means for factorial designs where the ultimate sampling units are randomized over the different treatment combinations through a CRD or an RBD. Procedures for factorial designs where clusters of sampling units are randomized over the treatment combinations or to test hypotheses about

ratios of population totals are obtained analogously to the methods developed in van den Brakel (2008) for single-factor experiments.

The design-based variance estimator that is developed for the various treatment effects does not require joint inclusion probabilities nor design-covariances between the different subsamples. As a result a design-based analysis procedure for factorial designs embedded in complex probability samples is obtained with the attractive relatively simple structure as if the sampling units are drawn with unequal selection probabilities with replacement. The traditional advantages of factorial designs, summarized in the first paragraph of the discussion, still apply under this design-based approach. As illustrated with variance expression (2.29) fewer experimental units are required to estimate the main effects with the same precision in a factorial setup compared to separate single-factor designs.

The advantage of an RBD over a CRD is that the between block variance is removed from the estimated treatment effects. In the standard model-based theory for the analysis of randomized experiments, an F-test for the blocks as well as the treatment factors is available. Under restricted randomization of an RBD, however, it is generally argued that a F-test for the block effects is not valid. In these cases alternative measures to evaluate the efficiency of an RBD are available; see for example Montgomery (2001). In the design-based theory developed for RBDs in this paper there is an asymmetry between the block and treatment factors, as in the case of the randomization approach followed by Hinkelmann and Kempthorne (1994). Due to the restricted randomization within the blocks there is no meaningful test for the main effect of the block factor available.

Acknowledgements

The author wishes to thank the Associate Editor and the unknown referees for giving constructive comments on a former draft of this paper. The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

References

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*, Chichester: John Wiley.

Chipperfield, J., and Bell, P. (2010). Embedded experiments in repeated and overlapping surveys. *Journal of the Royal Statistical Society,* Series A, 173, 51-66.

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E. and Lensvelt-Mulders, G. (2007). The influence of advance letters in response in telephone surveys. *Public Opinion Quarterly*, 71, 413-443.

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

Fienberg, S.E., and Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *The Canadian Journal of Statistics*, 16, 135-151.

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

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

Groves, R.M., Cialdini R.B. and Couper, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56, 475-495.

Groves, R.M., and Couper, M.P. (1998). *Nonresponse in household interview surveys*, New York: John 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.

Hidiroglou, M.A., and Lavallée, P. (2005). Indirect two-phase sampling: Applying it to questionnaire field-testing. *Proceedings of Statistics Canada Symposium 2005: Methodological challenges for future information needs*.

Hinkelmann, K., and Kempthorne, O. (1994). *Design and Analysis of Experiments, Volume 1: Introduction to experimental design*, New York: John Wiley.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Jäckle, A., Roberts, C. and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78, 3-20.

Kempthorne, O. (1955). The Randomization Theory of Experimental Inference. *Journal of the American Statistical Association*, 50, 946-967.

Luiten, A., Campanelli, P., Klaasen, D. and Beukenhorst, D. (2008). Advance letters and the language and behaviour profile, paper presented at the 19th International Workshop on Household Survey Nonresponse.

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

Montgomery, D.C. (2001). *Design and Analysis of Experiments*, New York: John Wiley.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

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

Scheffé, H. (1959). *The Analysis of Variance*, New York: John Wiley.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*, Chichester: John Wiley.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society*, Series A, 171, 581-613.

van den Brakel, J.A. (2010). Design-based analysis of factorial designs embedded in probability samples. Discussion paper 201014, Statistics Netherlands, Heerlen.

van den Brakel, J.A., and Binder, D. (2000). Variance estimation for experiments embedded in complex sampling schemes. *Proceedings of the section on Survey Research Methods*, American Statistical Association, 805-810.

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. An Application in the Dutch Labor Force Survey. *Journal of Official Statistics*, 18, 217-231.

van den Brakel, J.A., and Renssen, R.H. (1998). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14, 277-295.

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

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. American Mathematical Society*, 54, 426-482.