Design-based analysis of factorial designs embedded in probability samples

Jan ven den Brakel

Discussion paper (201014)

Statistics Netherlands

The Hague/Heerlen, 2010

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2008–2009	= 2008 to 2009 inclusive
2008/2009	= average of 2008 up to and including 2009
2008/'09	= crop year, financial year, school year etc. beginning in 2008 and ending in 2009
2006/'07–2008/'09	= crop year, financial year, etc. 2006/'07 to 2008/'09 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201014 X-10

Design-based analysis of factorial designs embedded in probability samples

Jan van den Brakel

Summary: 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 in 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.

Keywords: completely randomized designs, design-based inference, embedded experiments, measurement error models, model-assisted inference, randomized block designs.

1 Introduction

The purpose of survey sampling is to obtain statistical information about a finite population, by selecting a probability sample from this population, measuring the required information about the units in this sample and estimating finite population parameters such as means, totals and ratios. The statistical inference in the traditional design-based and model-assisted approach is predominantly based on the stochastic structure induced by the sampling design. Parameter and variance estimators are derived under the concept of repeatedly drawing samples from a finite population according to the same sampling design with the population values held fixed. Statistical models only play a minor role under this type of inference. This is the traditional approach of survey sampling theory, developed by notable authors like Hansen, et al. (1953), Kish (1965), Cochran (1977) and Särndal et al., (1992). A well known design-based estimator is the Horvitz-Thompson estimator, developed by Narain (1951) and Horvitz and Thompson (1952) for unequal probability sampling from finite populations without replacement. The accuracy of the Horvitz-Thompson estimator can be improved by making advantage of available auxiliary information about the complete target population, resulting in the model-assisted approach developed by Särndal et al. (1992). Many national statistical institutes rely on this design-based and model-assisted approach to compile official statistics.

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 and to obtain more quantitative insight in the various sources of non-sampling errors. At national statistical offices such experiments are particularly useful to quantify discontinuities in the series of repeated surveys due to adjustments into the survey process. A continuously or repeatedly conducted survey makes up series that describe the development of target parameters. Comparability over time is a key aspect of the relevance of these figures. Embedded experiments can be used to avoid that one or more modifications in the survey process result in unexplained differences in the series of a survey.

In an embedded experiment, the sample is randomly divided into two (or more) subsamples according to an experimental design. In survey literature, such experiments are also referred to as split-ballot designs or interpenetrating subsampling, and date back to Mahalanobis (1946). See also Fellegi (1964), Cochran (1977 section 13.15) and Hartley and Rao (1978) for examples of embedded experiments which are aimed to estimate the different variance components of non-sampling errors. Fienberg and Tanur (1987, 1988, 1989 and 1996) reviewed the parallels and differences between the methodology of random sampling and randomized experiments and discussed how the methodology of both fields can be applied in the design and analysis of embedded experiments. In their 1988 article they give a comprehensive overview of applications of embedded experiments that can be found in the literature. A series of applications of embedded experiments conducted at Statistics Netherlands can be found in Van den Brakel and Renssen (1998) and Van den Brakel (2001, 2008).

The statistical inference that is traditionally employed in the theory of design and analysis of randomized experiments is predominantly model-based, see for example Searle (1971), Scheffé (1959) or Cochran and Cox (1957). 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 developed 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 like 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 significantly different. This requires an analysis procedure that accounts for 1) the sample design that is used to select a probability sample from the finite target population, 2) the experimental design which is used to randomly assign the sampling units to the different treatments in the experiment, and 3) the estimation procedure that is used in the regular survey for the estimation of target parameters.

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 (2001, 2008), Van den Brakel and Renssen (1998, 2005) and Van den Brakel and Van Berkel (2002). In their approach the Horvitz-Thompson estimator and the generalized regression estimator are applied to derive approximately design-unbiased estimators for the population parameters observed under the different treatments of the experiment. Furthermore, an approximately design-unbiased estimator for the covariance matrix of the contrasts between the parameter estimates is derived. This gives rise to a design-based Wald- or t-statistic to test whether the differences between finite population parameter estimates observed under the different survey implementations are significantly different. 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) or Cochran and Cox, 1957). Therefore the design-based theory for the analysis of embedded experiments is extended to factorial designs in this paper.

A real life example of an experiment embedded in the Dutch Labor Force Survey (LFS) with advance letters is used to illustrate the methodology developed in this paper. The purpose of this experiment is to investigate whether the standard advance letter can be improved. Before a new advance letter is implemented as a standard in the LFS, its effect on response behavior as well as the possible effects on the main parameter estimates of the LFS must be quantified. To this end, six different advance letters are considered in a 2×3 factorial setup. The paper starts with developing the theory for factorial designs where the effect of two factors is tested simultaneously in sections 2 through 7. Subsequently the methodology is extended to higher order factorial designs in section 8. In these sections we confined ourselves to experimental designs where the ultimate sampling units of the sampling design are randomized over the different treatment combinations to test hypotheses about parameters that are defined as population means. In section 9, 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 10, these methods are applied to the factorial experiment with advance letters in the Dutch LFS. The paper concludes with a discussion in section 11.

2 Design of embedded experiments

Consider a $K \times L$ factorial design to test the effect of two factors at $K \geq 2$ and $L \geq 2$ levels simultaneously. The most straightforward approach for an embedded factorial design is to apply unrestricted randomization of the ultimate sampling units over the $KL = K \times L$ different treatment combinations, resulting in a factorial CRD. Unrestricted randomization is, however, not very efficient from a statistical point of view. The power of an experiment might be improved by using sampling structures such as strata, clusters or interviewers as block variables in a randomized block design (RBD) since restricted randomization removes the variance between the blocks from the analysis of the experiment, Fienberg and Tanur (1987, 1988). Unrestricted randomization by

means of a CRD might also result in practical complications, like overly long traveling distances for interviewers. This can be avoided by using small geographical regions as a block variable, see e.g. Van den Brakel and Van Berkel (2002). Restricted randomization results in factorial RBDs with strata, clusters or interviewers as the block variable.

The field staff requires special attention in the planning and design stage of an experiment. From a statistical point of view it is attractive to use interviewers as the block variable in an RBD, since this removes the interviewer variance component from the analysis of the experiment. A major drawback is that this implies that each interviewer has to collect data under the KL different treatment combinations, which might give rise to confusion. If it is decided that interviewers are assigned to one treatment only, then this must be done randomly to avoid one of the treatments being systematically favored with experienced interviewers or handicapped with newly recruited staff. See Van den Brakel and Renssen (1998) and Van den Brakel (2008) for more details about issues concerning the field staff in embedded experiments.

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 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, see for example Hinkelmann and Kempthorne (2005) or Montgomery (2001). These designs, however, are highly balanced and therefore generally hard to combine with the fieldwork restrictions encountered in the daily practice of survey sampling. An example of a factorial RBD with interviewers as the block variable is given by Van den Brakel et al. (2006).

In the remaining part of this section, the inclusion probabilities for the sampling units are derived for factorial CRDs and RBDs. These are the probabilities that a unit in the finite population is selected in the initial sample and subsequently is assigned to one of the subsamples according to the experimental design. These probabilities play a central role in the design-based inference that is developed in the next sections, since they are used to incorporate the information about both the sampling design and the experimental design in the estimation procedure.

Consider a finite population U containing N units. Estimates for unknown parameters are obtained through probability survey sampling. We allow for generally complex sampling schemes to draw a sample s from the finite population U. Consider an experiment embedded in a sample survey, aimed to investigate the effect of two different factors. 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 as well as the interaction effects between both factors on the main parameter estimates of the ongoing survey. To this end a probability sample s is drawn from a finite target population according to the sample design of the regular survey. Let π_i denote the first order inclusion probabilities for unit i and $\pi_{ii'}$ the second order inclusion probabilities for units i and i'. Subsequently, this sample is randomly divided in KL subsamples according to a randomized experiment.

In the case of a CRD, the sample is randomly divided in 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. Let $n_{+l} = \sum_{k=1}^{K} n_{kl}$ be the total number of sampling units assigned to treatment l, $n_{k+} = \sum_{l=1}^{L} n_{kl}$ the total number of sampling units assigned to treatment k, and $n_{++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}$ 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 $\pi_i^* = \pi_i(n_{kl}/n_{++})$.

In the case of an RBD, the sampling units are deterministically grouped in B more or less homogeneous blocks s_b. Potential block variables are sampling structures like strata, primary sampling units, clusters or interviewers. 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, $n_{b++} = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{bkl}$ the number of sampling units in block b and $n_{+kl} = \sum_{b=1}^{B} n_{bkl}$ the number of sampling units that is assigned to subsample s_{kl}. 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 $\pi_i^* = \pi_i(n_{bkl}/n_{b++})$.

Each subsample can be considered as the realization of a two-phase sample. The first phase sample coincides with the sampling design of the regular survey that is used to draw the initial sample from the finite target population. The second phase sample is simple random sampling without replacement from the first phase sample in the case of a CRD or stratified simple random sampling without replacement in the case of an RBD where the strata are the block variable of the experimental design. In many practical applications, one of the KL subsamples is assigned to the regular survey, and serves besides the regular publication purposes of the ongoing sample survey also as the control group in the experiment. In such situations, the size of this subsample will be substantially larger compared with the other subsamples.

3 Measurement error models

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. Design-based sampling theory is largely based on the traditional notion that the observations that are obtained from the sampling units are true fixed values observed without error (e.g. Cochran, 1977). This approach is not tenable if experiments are conducted to test systematic differences between finite population parameter estimates that are obtained under different survey implementations due to non-sampling errors. Therefore a measurement error model is required to link systematic differences between finite population parameters due to different survey implementations or treatments. 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 realiza-

tion of the measurement error model $y_{iqkl} = u_i + \beta_{kl} + \gamma_q + \epsilon_{ikl}$. 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. 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:

$$\mathbf{y}_{iq} = \mathbf{j}_{KL} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL} \gamma_q + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbf{y}_{iq} = (y_{iq11}, \ldots, y_{iqkl}, \ldots, y_{iqKL})^t$, $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{kl}, \ldots, \beta_{KL})^t$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i11}, \ldots, \varepsilon_{ikl}, \ldots, \varepsilon_{iKL})^t$ and $\mathbf{j}_{KL}$ a vector of order KL with each element equal to one. 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. Let E_m and Cov_m denote the expectation and the covariance with respect to the measurement error model. The following model assumptions are made:

$$\mathrm{E}_m(\boldsymbol{\varepsilon}_i) = \mathbf{0}, \tag{2}$$

$$\mathrm{Cov}_m(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'}^t) = \begin{cases} \boldsymbol{\Sigma}_i & : \quad i = i' \\ \mathbf{O} & : \quad i \neq i' \end{cases}, \tag{3}$$

$$\mathrm{E}_m(\xi_q) = 0, \tag{4}$$

$$\mathrm{Cov}_m(\xi_q, \xi_{q'}) = \begin{cases} \tau_q^2 & : \quad q = q' \\ 0 & : \quad q \neq q' \end{cases}, \tag{5}$$

$$\mathrm{Cov}_m(\varepsilon_{ikl}, \xi_q) = 0, \tag{6}$$

where $\mathbf{0}$ is a vector of order KL with each element zero and $\mathbf{O}$ a matrix of order $KL \times KL$ with each element zero.

Note that the treatment effect can be decomposed in the traditional way of an ANOVA for a two-way layout:

$$\beta_{kl} = u + A_k + B_l + AB_{kl}. \tag{7}$$

If the treatment effects are defined as fixed deviations from the individuals' intrinsic value u_i, then the overall mean u equals zero. The following restrictions are required to identify the model:

$$\sum_{k=1}^{K} A_k = 0, \quad \sum_{l=1}^{L} B_l = 0, \tag{8}$$

$$\sum_{k=1}^{K} AB_{kl} = 0, \quad l = 1, 2, \ldots, L, \quad \sum_{l=1}^{L} AB_{kl} = 0, \quad k = 1, 2, \ldots, K. \tag{9}$$

4 Testing hypotheses about finite population parameters

Suppose that there are Q interviewers available for the data collection and that the finite population U can conceptually be divided in Q subpopulations U_q of size N_q, $q = 1, \ldots, Q$ such that all units within each U_q are interviewed by the same interviewer if these units are included in the sample. Under measurement error model (1), KL values for the same target parameter in the

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

$$\bar{\mathbf{Y}} = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}_{KL} \psi + \mathbf{j}_{KL} \sum_{q=1}^{Q} \frac{N_q}{N} \xi_q + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\varepsilon}_i. \tag{10}$$

The objective of the experiment is to test hypotheses about the main effects and the interaction between the two treatment factors on the population parameters. 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. Since random deviations due to measurement errors and interviewer effects should not lead to significant differences between the estimated population parameters, hypotheses are formulated about $\bar{\mathbf{Y}}$ in expectation over the measurement error model, i.e.

$$\mathrm{E}_m(\bar{\mathbf{Y}}) = \mathbf{j}_{KL} \frac{1}{N} \sum_{i=1}^{N} u_i + \mathbf{j}_{KL} \psi + \boldsymbol{\beta}. \tag{11}$$

The hypothesis about the main effects of factor A is formulated as

$$\begin{aligned} H_0: &\quad \mathbf{C_A} \mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{0}, \\ H_1: &\quad \mathbf{C_A} \mathrm{E}_m \bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{12}$$

where

$$\mathbf{C_A} = \frac{1}{L} \left(\mathbf{j}_{(K-1)} | - \mathbf{I}_{(K-1)} \right) \otimes \mathbf{j}_L^t \equiv \frac{1}{L} \tilde{\mathbf{C}}_\mathbf{A} \otimes \mathbf{j}_L^t, \tag{13}$$

with $\mathbf{I}_{(K-1)}$ the identity matrix of order $K-1$. Matrix $\mathbf{C_A}$ defines the $K-1$ contrasts between the K levels of factor A, averaged over the L levels of factor B. From (11) and due to restrictions (8) and (9) it follows that the contrasts between the population parameters exactly correspond to the main effects of the first factor: $\mathbf{C_A} \mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{C_A} \boldsymbol{\beta} = (A_1 - A_2, \ldots, A_1 - A_K)^t$. Note that the contrasts in hypothesis (12) also can be expressed as $\tilde{\mathbf{C}}_\mathbf{A} \mathrm{E}_m \bar{\mathbf{Y}}_\mathbf{A}$ where $\bar{\mathbf{Y}}_\mathbf{A} = (\bar{Y}_{1.}, \ldots \bar{Y}_{K.})^T$ and $\bar{Y}_{k.} = \frac{1}{L} \sum_{l=1}^{L} \bar{Y}_{kl}$ the population mean observed under the $k-$th level of factor A, averaged over the L levels of factor B.

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

$$\begin{aligned} H_0: &\quad \mathbf{C_B} \mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{0}, \\ H_1: &\quad \mathbf{C_B} \mathrm{E}_m \bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{14}$$

where

$$\mathbf{C_B} = \frac{1}{K} \mathbf{j}_K^t \otimes \left(\mathbf{j}_{(L-1)} | - \mathbf{I}_{(L-1)} \right) \equiv \frac{1}{K} \mathbf{j}_K^t \otimes \tilde{\mathbf{C}}_\mathbf{B}, \tag{15}$$

which is the matrix that defines the $L-1$ contrasts between the L levels of factor B, averaged over the K levels of factor A. From (11) and due to restrictions (8) and (9) it follows that the contrasts between the population parameters exactly correspond to the main effects of the second factor: $\mathbf{C_B} \mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{C_B} \boldsymbol{\beta} = (B_1 - B_2, \ldots, B_1 - B_L)^t$. The contrasts in hypothesis (14) can be expressed as

$\tilde{\mathbf{C}}_{\mathbf{B}}\mathrm{E}_m\bar{\mathbf{Y}}_{\mathbf{B}}$, where $\bar{\mathbf{Y}}_{\mathbf{B}} = (\bar{Y}_{.1}, \ldots \bar{Y}_{.L})^T$ and $\bar{Y}_{.l} = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_{kl}$ the population mean observed under the $l-$th level of factor B, averaged over the K levels of factor A.

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, see e.g. Hinkelmann and Kempthorne (1994, ch. 11). Therefore the hypothesis about the interactions between factor A and B can be defined as

$$\begin{aligned} H_0: &\quad \mathbf{C}_{\mathbf{AB}}\mathrm{E}_m\bar{\mathbf{Y}} = \mathbf{0}, \\ H_1: &\quad \mathbf{C}_{\mathbf{AB}}\mathrm{E}_m\bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{16}$$

where

$$\mathbf{C}_{\mathbf{AB}} = \left(\mathbf{j}_{(K-1)}| - \mathbf{I}_{(K-1)}\right) \otimes \left(\mathbf{j}_{(L-1)}| - \mathbf{I}_{(L-1)}\right) = \tilde{\mathbf{C}}_{\mathbf{A}} \otimes \tilde{\mathbf{C}}_{\mathbf{B}} \tag{17}$$

denotes the $(K-1)(L-1)\times KL$ matrix with 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 $\mathbf{C}_{\mathbf{AB}}\mathrm{E}_m\bar{\mathbf{Y}} = \mathbf{C}_{\mathbf{AB}}\boldsymbol{\beta} = (AB_{11} - AB_{12} - AB_{21} + AB_{22}, \ldots, AB_{11} - AB_{1L} - AB_{21} + AB_{2L}, \ldots, AB_{11} - AB_{12} - AB_{K1} + AB_{K2}, \ldots, AB_{11} - AB_{1L} - AB_{K1} + AB_{KL})^t$ defines a $(K-1)(L-1)$ vectors containing the interactions between the two treatment factors.

The observations obtained from the sampling units in the subsamples s_{kl}, are used to obtain design-based estimates for the elements of $\bar{\mathbf{Y}}$ as well as the covariance matrix of the contrasts between these subsample estimates. This gives rise to a Wald statistic to test hypotheses (12), (14) and (16), which accounts for the sampling design, the experimental design, and the weighting procedure applied in the regular sample survey.

5 Parameter estimation under different treatments

Based on the observations obtained in the subsamples s_{kl}, design-unbiased estimates of the population treatment means in $\bar{\mathbf{Y}}$ can be derived. In section 2, the inclusion probabilities π_i^* for the sampling units in the subsamples are derived under the sampling design that is used to draw the initial sample from the finite population and the experimental design that is used to divide this sample randomly into KL subsamples. These probabilities are used in the Horvitz-Thompson estimator, developed by Narain (1951) and Horvitz and Thompson (1952) for unequal probability sampling without replacement from finite populations, to obtain design-unbiased estimators for the unknown population parameters.

For notational convenience, the subsript q will be omitted in y_{iqkl} if possible, since there is no need to sum explicitly over the interviewer subscript in most of the formulas developed in the rest of this paper. The Horvitz-Thompson estimator for $\bar{Y}_{kl}$ is given by

$$\hat{\bar{Y}}_{kl} = \frac{1}{N}\sum_{i\in s_{kl}}\frac{y_{ikl}}{\pi_i^*} \tag{18}$$

Since each subsample can be considered as a two-phase sample, it follows directly that (18) is design unbiased, since $\mathrm{E}_s\mathrm{E}_\varepsilon(\hat{\bar{Y}}_{kl} \mid s) = \bar{Y}_{kl}$ where E_s and E_ε denote the expectation with respect to the

sampling design and the experimental design, respectively. This estimator is also design unbiased for $\mathrm{E}_m \bar{Y}_{kl}$ by definition.

In survey sampling the model-assisted approach developed by Särndal et al. (1992) is widely applied to improve the accuracy of the Horvitz-Thompson estimator by making advantage of available auxiliary information about the target population. This estimator is derived from a linear regression model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables for which the totals in the finite target population are known. If the underlying linear model explains the variation of the target parameter in the finite population reasonably well, then this might result in a reduction of the design variance of the Horvitz-Thompson estimator as well as the bias due to selective nonresponse. If the model is misspecified, then this might result in an increase of the design variance but the property that the generalized regression estimator is approximately design unbiased remains. The use of auxiliary information in the analysis of an embedded experiment by means of the generalized regression estimator might improve the accuracy of the analysis and has a direct design-based analogy with covariance analysis that is used in experimental design theory.

For each unit in the population a $H-$vector $\mathbf{x}_i$ with auxiliary information is available and it is assumed that the $\mathbf{x}$ variables are observed without measurement errors and thus not affected by the treatments of the experiment. The finite population means of these variables are known and are denoted by $\bar{\mathbf{X}}$. 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 values u_i the measurement error model (1) of section 3 are an independent realization of the following linear regression model:

$$u_i = \mathcal{B}^t \mathbf{x}_i + e_i,$$

where $\mathcal{B}^t$ is 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 regression coefficients for the intrinsic variable in the finite population are defined as

$$\mathbf{b} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i=1}^{N} \frac{\mathbf{x}_i u_i}{\omega_i^2}. \tag{19}$$

As follows from the measurement error model (1), the variables u_i cannot be observed without measurement bias. Therefore (19) is not observable, even in the case of a complete enumeration of the finite population. In the context of embedded experiments a separate set of regression coefficients can be defined for each treatment combination. In the case of a complete enumeration under the $kl-$th treatment combination,

$$\tilde{\mathbf{b}}_{kl} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i=1}^{N} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2}, \tag{20}$$

denotes the finite population regression coefficients of the regression of y_{ikl} on $\mathbf{x}_i$. A Horvitz-Thompson type estimator for the finite population regression coefficients (20) based on the obser-

vations obtained under the n_{kl} sampling units in subsample s_{kl} is given by

$$\hat{\mathbf{b}}_{kl} = \left(\sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i^*} \right)^{-1} \sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i^*}. \tag{21}$$

The generalized regression estimator for $\bar{Y}_{kl}$, based on the n_{kl} observations of subsample s_{kl}, is defined as (Särndal et al., 1992)

$$\hat{\bar{Y}}_{kl;greg} \;=\; \hat{\bar{Y}}_{kl} + \hat{\mathbf{b}}_{kl}^t (\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}), \quad k = 1, 2, \ldots, K, \;\text{ and }\; l = 1, 2, \ldots, L, \tag{22}$$

where $\hat{\bar{\mathbf{X}}}$ denotes the Horvitz-Thompson estimator for the population means of the auxiliary variables $\bar{\mathbf{X}}$ based on the n_{kl} sample units of subsample s_{kl}. Now

$$\hat{\bar{\mathbf{Y}}}_{GREG} = (\hat{\bar{Y}}_{11;greg}, \ldots, \hat{\bar{Y}}_{kl;greg}, \ldots, \hat{\bar{Y}}_{KL;greg})^t$$

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 $\hat{\mathbf{b}}_{kl} = \hat{\mathbf{b}}_{k'l'}$. In that case, it might be efficient to substitute for $\hat{\mathbf{b}}_{kl}$ in the generalized regression estimator (22) the pooled estimator

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2 \pi_i} \right)^{-1} \left(\sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{kl}} \frac{\mathbf{x}_i y_{ikl}}{\omega_i^2 \pi_i} \right). \tag{23}$$

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 generalized regression 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 completed included in the regression coefficients. In case of the pooled estimator (23), the KL generalized regression estimators are exactly equal by definition, since $\hat{\bar{Y}}_{kl;greg} = \hat{\mathbf{b}}^t \bar{\mathbf{X}}$ for all k and l.

6 Variance estimation

Since, the generalized regression estimator defined by (22) is not linear, an approximation of the variance of this estimator is usually obtained from a linearized approximation of (22), see for example Särndal et al., (1992), ch. 6. Expressing (22) as a function of $(\hat{\bar{Y}}_{kl}, \hat{\mathbf{b}}_{kl}, \hat{\bar{\mathbf{X}}})$, the generalized regression estimator can be approximated by means of a Taylor linearization about $(\mathrm{E}_m(\bar{Y}_{kl}), \mathrm{E}_m(\tilde{\mathbf{b}}_{kl}), \bar{\mathbf{X}})$ that is truncated at the first order term, i.e.

$$\hat{\bar{Y}}_{kl;greg} \doteq \hat{\bar{Y}}_{kl} + \mathbf{b}_{kl}^t (\bar{\mathbf{X}} - \hat{\bar{\mathbf{X}}}), \tag{24}$$

with $\mathbf{b}_{kl}$ the expectation of the finite population regression coefficients with respect to the measurement error model observed under the $kl-$th treatment combination:

$$\mathbf{b}_{kl} = \mathrm{E}_m \tilde{\mathbf{b}}_{kl} = \left(\sum_{i=1}^{N} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i=1}^{N} \frac{\mathbf{x}_i (u_i + \beta_{kl} + \psi)}{\omega_i^2}. \tag{25}$$

Since the KL subsamples are drawn without replacement from a finite population, there is a nonzero design covariance between elements of $\hat{\bar{\mathbf{Y}}}_{GREG}$. Design-based estimators for these covariance terms require that for each unit in the sample observations under each of the KL treatments are obtained. These paired observations are, however, not available since each sampling unit is assigned to one the KL treatments only. Van den Brakel and Binder (2000) and Hidiroglou and Lavellée (2005) approximated this covariance matrix using an imputation technique for the missing paired observations. Van den Brakel (2001, 2008) and Van den Brakel and Renssen (2005) developped a design-based estimator for the covariance matrix of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{GREG}$ for single-factor experiments that only requires one observation for each sampling unit instead of the repeated measurements of all treatment combinations within each sampling unit. This estimator is extended to factorial designs in this section.

Let $\mathbf{C}$ denote the contrast matrix $\mathbf{C_A}$, $\mathbf{C_B}$ or $\mathbf{C_{AB}}$. Under the condition that a constant $H-$vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$, an expression for the covariance matrix of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{GREG}$ is derived in the appendix. It is also proved that a design-based estimator for this covariance matrix is given by

$$\widehat{\mathrm{Cov}}(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{GREG}) = \mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t, \tag{26}$$

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

$$\hat{d}_{kl} = \frac{1}{n_{kl}}\frac{1}{n_{kl}-1}\sum_{i=1}^{n_{kl}}\left(\frac{n_{++}(y_{ikl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{kl}}\sum_{i'=1}^{n_{kl}}\frac{n_{++}(y_{i'kl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{27}$$

in the case of a CRD and

$$\hat{d}_{kl} = \sum_{b=1}^{B}\frac{1}{n_{bkl}}\frac{1}{n_{bkl}-1}\sum_{i=1}^{n_{bkl}}\left(\frac{n_{b++}(y_{ikl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bkl}}\sum_{i'=1}^{n_{bkl}}\frac{n_{b++}(y_{i'kl}-\hat{\mathbf{b}}_{kl}^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{28}$$

in the case of an RBD.

The condition that a constant $H-$vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$, implies that a weighting model is used that at least uses the size of the finite population as a priori information. This condition holds for weighting models that contain an intercept or at least one categorical variable that poststratifies the finite population in two or more subpopulations. This is a rather weak condition that is met by many weighting model used in practice. An exception, however, is the ratio model that generates the ratio estimator, since this model only uses one continuous variable as auxiliary information, Särndal et al. (1992), section 7.3.

Although (26) is an approximately design-unbiased estimator for the covariance matrix of $(\mathbf{C}\hat{\bar{\mathbf{Y}}}_{GREG})$, it has a structure as if the subsamples are drawn independently through sampling with replacement using unequal selection probabilities. This is a remarkable result and gives rise to an attractive variance estimation procedure for embedded experiments, since no design covariances between the subsample estimates appear in (26) and no second order inclusion probabilities are required in the variance estimators (27) and (28). This result is obtained by making advantage of several factors. The covariance matrix is derived for contrasts between the subsample estimates. This covariance matrix is decomposed in a covariance component with respect to the measurement

error model, the sampling design and the experimental design, see (47) in the appendix. For the (co)variance of the contrasts between GREG estimators that uses a weighting model that meets the condition that a constant vector $\mathbf{a}$ exists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$, it follows that the residuals of the GREG estimator only contain measurement errors, see formula (48) in the appendix. Under the assumption that the measurement errors between the units are independent, second order inclusion probabilities cancel out in the covariance components with respect to the measurement error model (49) and the sampling design (50). The covariance components with respect to the experimental design (51) have the structure of simple random sampling without replacement in the case of CRD, or stratified simple random sampling without replacement in the case of an RBD. In the variance of the contrasts under (stratified) simple random sampling without replacement, the covariance terms between the subsample estimates cancel out against the finite population corrections in the variance terms.

The minimum use of auxiliary information used in the generalized regression estimator is obtained with a weighting scheme that only uses the size of the finite population as a priory 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

$$\hat{\bar{Y}}_{kl;greg} = \left(\sum_{i\in s_{kl}} \frac{1}{\pi_i^*}\right)^{-1} \left(\sum_{i\in s_{kl}} \frac{y_{ikl}}{\pi_i^*}\right) \equiv \tilde{y}_{kl}, \tag{29}$$

and $(\hat{\mathbf{b}}_{kl}) = \tilde{y}_{kl}$. Expression (29) 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 (27) and (28) for a CRD and an RBD respectively, where $\hat{\mathbf{b}}_{kl}^t\mathbf{x}_i = \tilde{y}_{kl}$. Estimator (29) is preferable above the Horvitz-Thompson estimator (18), since (29) is more stable and the covariance matrix of the contrasts between (29) always has the relatively simple form of (26). The covariance matrix of contrasts between the Horvitz-Thompson estimators (18) is more complex for designs where $\sum_{i\in s_{kl}} \frac{1}{\pi_i^*} \neq N$, see Van den Brakel (2001).

To have more stable variance estimators, pooled variance estimators for the diagonal elements of $\hat{\mathbf{D}}$ can be used as an alternative for (27) or (28). Under the assumption that the variances of the measurement errors under the different treatments are equal, i.e. $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$ in (3), a pooled variance estimator for a CRD is given by

$$\hat{d}_{kl}^p = \frac{1}{n_{kl}} \frac{1}{n_{++} - KL} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{k'l'}} \left(\frac{n_{++}(y_{ik'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{k'l'}} \sum_{i'=1}^{n_{k'l'}} \frac{n_{++}(y_{i'k'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{30}$$

and for an RBD by

$$\hat{d}_{kl}^p = \sum_{b=1}^{B} \frac{1}{n_{bkl}} \frac{1}{n_{b++} - KL} \sum_{k'=1}^{K} \sum_{l'=1}^{L} \sum_{i=1}^{n_{bk'l'}} \left(\frac{n_{b++}(y_{ik'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bk'l'}} \sum_{i'=1}^{n_{bk'l'}} \frac{n_{b++}(y_{i'k'l'} - \hat{\mathbf{b}}_{k'l'}^t\mathbf{x}_{i'})}{N\pi_{i'}} \right)^2. \tag{31}$$

7 The Wald test

The Wald test (Wald, 1943) is frequently applied in design-based testing procedures, see e.g. Skinner et al. (1989) or Chambers and Skinner (2003). The design-based estimators that are derived for the population parameter that is observed under the different treatment combinations (section 5) and the covariance matrix of the contrasts between these estimates (section 6) can be used to construct a design-based Wald statistic to test the hypotheses described in section 4:

$$W = \hat{\tilde{\mathbf{Y}}}^t_{GREG} \mathbf{C}^t \left(\mathbf{C} \hat{\mathbf{D}} \mathbf{C}^t \right)^{-1} \mathbf{C} \hat{\tilde{\mathbf{Y}}}_{GREG}. \tag{32}$$

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. The limit distribution for $\hat{\tilde{\mathbf{Y}}}_{GREG}$ is unknown for generally complex sampling schemes, but it is conjectured that $\hat{\tilde{\mathbf{Y}}}_{GREG}$ is asymptotically multivariate normally distributed. Then 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 (Searle 1971, theorem 2, ch. 2). The validity of the conjecture mentioned, has been confirmed by simulation studies for single-factor experiments, see Van den Brakel and Renssen (2005) and Van den Brakel (2008).

The Wald statistic for the hypotheses about the main effects (12) and (14) are given by (32) using the contrast matrix $\mathbf{C_A}$ or $\mathbf{C_B}$, which are specified by (13) and (15) respectively. Under the null hypothesis, it follows that $W \rightarrow \chi^2_{[K-1]}$ for the test about the main effects of factor A or $W \rightarrow \chi^2_{[L-1]}$ for the test about the main effects of factor B. The Wald statistic for the hypothesis about the interaction effects between factor A and B (16) is given by (32) using the contrast matrix $\mathbf{C_{AB}}$ that is specified by (17). Under the null hypothesis, it follows that $W \rightarrow \chi^2_{[(K-1)(L-1)]}$.

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

$$\hat{\tilde{\mathbf{Y}}}_{A;GREG} = (\hat{\tilde{Y}}_{1.;greg}, \ldots, \hat{\tilde{Y}}_{K.;greg})^t, \quad \text{with} \quad \hat{\tilde{Y}}_{k.;greg} = \frac{1}{L} \sum_{l=1}^{L} \hat{\tilde{Y}}_{kl;greg},$$

$$\hat{\mathbf{D}}_{\mathbf{A}} = \text{Diag}(\hat{d}_{1.}, \ldots, \hat{d}_{K.}), \quad \text{with} \quad \hat{d}_{k.} = \frac{1}{L^2} \sum_{l=1}^{L} \hat{d}_{kl}.$$

It follows that

$$\mathbf{C_A} \hat{\tilde{\mathbf{Y}}}_{GREG} = \frac{1}{L} (\tilde{\mathbf{C}}_{\mathbf{A}} \otimes \mathbf{j}^t_L) \hat{\tilde{\mathbf{Y}}}_{GREG} = \tilde{\mathbf{C}}_{\mathbf{A}} \hat{\tilde{\mathbf{Y}}}_{A;GREG}, \tag{33}$$

$$\mathbf{C_A} \hat{\mathbf{D}} \mathbf{C}^t_{\mathbf{A}} = \frac{1}{L^2} (\tilde{\mathbf{C}}_{\mathbf{A}} \otimes \mathbf{j}^t_L) \hat{\mathbf{D}} (\tilde{\mathbf{C}}_{\mathbf{A}} \otimes \mathbf{j}^t_L)^t = \tilde{\mathbf{C}}_{\mathbf{A}} \hat{\mathbf{D}}_{\mathbf{A}} \tilde{\mathbf{C}}^t_{\mathbf{A}}. \tag{34}$$

With result (33), (34) and the matrix inversion lemma, also known as Bartlett's identity (Morisson, 1990, ch.2), the Wald statistic for the main effects of factor A can be simplified to:

$$\begin{aligned} W &= \hat{\tilde{\mathbf{Y}}}^t_{A;GREG} \tilde{\mathbf{C}}^t_{\mathbf{A}} (\tilde{\mathbf{C}}_{\mathbf{A}} \hat{\mathbf{D}}_{\mathbf{A}} \tilde{\mathbf{C}}^t_{\mathbf{A}})^{-1} \tilde{\mathbf{C}}_{\mathbf{A}} \hat{\tilde{\mathbf{Y}}}_{A;GREG} \\ &= \hat{\tilde{\mathbf{Y}}}^t_{A;GREG} \left(\hat{\mathbf{D}}^{-1}_{\mathbf{A}} - \frac{1}{\text{Trace}(\hat{\mathbf{D}}^{-1}_{\mathbf{A}})} \hat{\mathbf{D}}^{-1}_{\mathbf{A}} \mathbf{j}_{K-1} \mathbf{j}^t_{K-1} \hat{\mathbf{D}}^{-1}_{\mathbf{A}} \right) \hat{\tilde{\mathbf{Y}}}_{A;GREG} \end{aligned}$$

$$= \sum_{k=1}^{K} \frac{\hat{\bar{Y}}^2_{k.;greg}}{\hat{d}_{k.}} - \left(\sum_{k=1}^{K} \frac{1}{\hat{d}_{k.}}\right)^{-1} \left(\sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k.;greg}}{\hat{d}_{k.}}\right)^2 \tag{35}$$

See the proof of formula (32) in Van den Brakel and Renssen (2005) for more details of this derivation. It follows that the main advantages of factorial designs still hold under this design-based approach. Since the main effects of a factor are averaged over the levels of the other factor, it is more efficient to conduct one factorial design instead of two separate single-factor experiments. As illustrated with variance expression (34) less experimental units are required to estimate the main effects with the same precision in a factorial setup compared to two separated single-factor designs. Moreover, a factorial design offers the possibility to analyze interactions between the two treatment factors.

In the remainder of this section, it will be shown for two special cases that the design-based Wald statistic is equal to the F-test of a standard ANOVA. 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 $\bar{y}_{kl} = \frac{1}{n_s}\sum_{i=1}^{n_s} y_{ikl}$. Under Hájek's ratio estimator (29) and the pooled variance estimator (30) it follows that $\hat{\bar{Y}}_{kl;greg} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, and

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

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

$$\bar{y}_{k.} = \frac{1}{L}\sum_{l=1}^{L}\bar{y}_{kl} = \frac{1}{n_{k+}}\sum_{l=1}^{L}\sum_{i=1}^{n_{kl}} y_{ikl}, \quad k = 1, \ldots, K. \tag{36}$$

The diagonal elements of $\hat{\mathbf{D}}_{\mathbf{A}}$ are now given by

$$\hat{d}^p_{k.} = \frac{1}{L^2}\sum_{l=1}^{L}\hat{d}^p_{kl} = \frac{1}{L^2}\sum_{l=1}^{L}\frac{\hat{S}^2_{p;CRD}}{n_s} = \frac{\hat{S}^2_{p;CRD}}{Ln_s} = \frac{\hat{S}^2_{p;CRD}}{n_{k+}}, \quad k = 1, \ldots, K. \tag{37}$$

Let $\bar{y}_{..} = \frac{1}{n_{++}}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{n_s} y_{ikl}$. Inserting (36) and (37) into (35), gives rise to the following expression for the Wald statistic of the main effects of factor A

$$W = \frac{1}{\hat{S}^2_{p;CRD}}\left[\sum_{k=1}^{K} n_{k+}\bar{y}^2_{k.} - n_{++}\bar{y}^2_{..}\right]. \tag{38}$$

Note that $W/(K-1)$ in (38) corresponds with the $F-$statistic for the main effects of an ANOVA for the two-way layout with interactions, Scheffé (1959), ch. 4. Under the null hypothesis and the assumption of normally and independently distributed observations, 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 $\pi = n_{+++}/N$ and $n_{kl} = n_{k'l'} = n_s$. 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 (29) and the pooled variance estimator (31) it follows that $\hat{\bar{Y}}_{kl;greg} = \bar{y}_{kl}$, $\hat{\mathbf{b}}_{kl} = \bar{y}_{kl}$, and

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

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

$$\bar{y}_{.k.} = \frac{1}{L} \sum_{l=1}^{L} y_{kl} = \frac{1}{n_{+k+}} \sum_{b=1}^{B} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}, \quad k = 1, \ldots, K, \tag{39}$$

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

$$\hat{d}^p_{k.} = \frac{1}{L^2} \sum_{l=1}^{L} \hat{d}^p_{kl} = \frac{\hat{S}^2_{p;RBD}}{n_{+k+}}, \quad k = 1, \ldots, K. \tag{40}$$

Let $\bar{y}_{...} = \frac{1}{n_{+++}} \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{i=1}^{n_{bkl}} y_{ikl}$. If these results are inserted into (35), then the expression for the Wald statistic of the main effects of factor A can be simplified to

$$W = \frac{1}{\hat{S}^2_{p;RBD}} \left[\sum_{k=1}^{K} n_{+k+} \bar{y}^2_{.k.} - n_{+++} \bar{y}^2_{...} \right]. \tag{41}$$

It can be recognized that $W/(K-1)$ in (41) corresponds with the $F-$statistic for the main effects of an ANOVA for the three-way layout with interactions, Scheffé (1959), ch. 4. As in the case of a CRD, this Wald and F-statistic have the same limit distribution.

8 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 appropriate notation for the treatment factors is introduced first. Let A_g denote the $g-$ th treatment factor in the experiment with levels $a_g = 1, \ldots, M_g$. In the general case there are $g = 1, \ldots, G$ factors included in the experiment. The population parameters observed under the $M_1 M_2 \ldots M_G$ treatment combinations are collected in the vector $\bar{\mathbf{Y}} = (\bar{Y}_{11\ldots1}, \ldots, \bar{Y}_{a_1 a_2 \ldots a_G}, \ldots, \bar{Y}_{M_1 M_2 \ldots M_G})^t$. The index for the levels of a factor runs within each level of its preceding factor. Thus index a_g runs from $a_g = 1, \ldots, M_g$ within each level of $a_{(g-1)}$. Hypotheses about the main effects and interactions are, as motivated in section (4) formulated about $\bar{\mathbf{Y}}$ in expectation over the measurement error model, i.e.

$$\begin{aligned} H_0 &: \quad \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} = \mathbf{0}, \\ H_1 &: \quad \mathbf{C}\mathrm{E}_m \bar{\mathbf{Y}} \neq \mathbf{0}, \end{aligned} \tag{42}$$

The contrast matrices for the main effects and interactions in (42) are developed for the general case of a $M_1 \times M_2 \times \ldots \times M_G$ factorial design. Let $\mathcal{A} = \{1, \ldots, G\}$ denote the set of labels for the factors and $\tilde{\mathbf{C}}_{\mathbf{A_g}} = (\mathbf{j}_{(M_g-1)} | -\mathbf{I}_{(M_g-1)})$. The following three functions are defined first;

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

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

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

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:

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

Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A_{g_1}}}$ by $\mathbf{J}_{2_{g_1}}$ sums over the levels of the factors $A_{(g_1+1)} \ldots A_G$ 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 $A_1 \ldots A_{(g_1-1)}$. 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 $A_1 \ldots A_{(g_1-1)}$.

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:

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

Postmultiplication of $\tilde{\mathbf{C}}_{\mathbf{A_{g_2}}}$ by $\mathbf{J}_{2_{g_1}}$ adds the levels of the factors $A_{(g_2+1)} \ldots A_G$ 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 $A_1 \ldots A_{(g_2-1)}$. 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 $A_{(g_1+1)} \ldots A_{(g_2-1)}$ that are nested within each combination of the levels of $A_1 \ldots A_{(g_1)}$. 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 $A_1 \ldots A_{(g_1-1)}$. 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 $A_1 \ldots A_{(g_1-1)}$.

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:

$$\mathbf{C}_{\mathbf{A_{g_1} A_{g_2} A_{g_3}}} = \left(\textstyle\prod_{g \in \mathcal{A} \setminus \{g_1, g_2, g_3\}} M_g\right)^{-1} \mathbf{J}_{1_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_1}}} \otimes \mathbf{J}_{3_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_2}}} \otimes \mathbf{J}_{3_{g_2,g_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_3}}} \otimes \mathbf{J}_{2_{g_3}},$$
$$g_1 = 1, \ldots, G-2, \ g_2 = 2, \ldots, G-1, \ g_3 = 3, \ldots G, \ g_1 < g_2 < g_3,$$

$$\mathbf{C}_{\mathbf{A_{g_1}A_{g_2}A_{g_3}A_{g_4}}} = \left(\prod_{g\in\mathcal{A}\setminus\{g_1,g_2,g_3,g_4\}} M_g\right)^{-1} \mathbf{J}_{1_{g_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_1}}} \otimes \mathbf{J}_{3_{g_1,g_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_2}}} \otimes \mathbf{J}_{3_{g_2,g_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_3}}} \otimes \mathbf{J}_{3_{g_3,g_4}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_{g_4}}} \otimes \mathbf{J}_{2_{g_4}},$$

$$g_1 = 1,\ldots,G-3,\ g_2 = 2,\ldots,G-2,\ g_3 = 3,\ldots G-1\ g_4 = 4,\ldots G,\ g_1 < g_2 < g_3 < g_4,$$

$$\vdots$$

$$\mathbf{C}_{\mathbf{A_1A_2A_3\ldots A_G}} = \tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \ldots \otimes \tilde{\mathbf{C}}_{\mathbf{A_G}}.$$

The number of rows of each contrast matrix coincides with the number of contrasts that define the various main effects and interactions and are specified in Table 1. The number of columns of these matrices equals $M_1 M_2 \ldots M_G$.

Table 1: *Number of constrasts for the various contrast matrices*

Contrast matrix	Number of contrasts (rows of the contrast matrix)
$\mathbf{C}_{\mathbf{A_{g_1}}}$	$M_{g_1} - 1$
$\mathbf{C}_{\mathbf{A_{g_1}A_{g_2}}}$	$(M_{g_1} - 1)(M_{g_1} - 1)$
$\mathbf{C}_{\mathbf{A_{g_1}A_{g_2}A_{g_3}}}$	$(M_{g_1} - 1)(M_{g_2} - 1)(M_{g_3} - 1)$
$\vdots$	$\vdots$
$\mathbf{C}_{\mathbf{A_1A_2A_3\ldots A_G}}$	$(M_1 - 1)(M_2 - 1)\ldots(M_G - 1)$

These contrast matrices are inserted in (42) 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 $M_1 M_2 \ldots M_G$ different subsamples. Let $n_{a_1\ldots a_G}$ denote the number of sampling units assigned to treatment combination $a_1 \ldots a_G$ in subsample $s_{a_1\ldots a_G}$ and $n_{+\ldots+}$ the size of the initial sample. In the case of a CRD, the first order inclusion probabilities for the units in subsample $s_{a_1\ldots a_G}$ are now given by $\pi_i^* = \pi_i(n_{a_1\ldots a_G}/n_{+\ldots+})$. In the case of an RBD, the first order inclusion probabilities for the units in subsample $s_{a_1\ldots a_G}$ are given by $\pi_i^* = \pi_i(n_{ba_1\ldots a_G}/n_{b+\ldots+})$ where $n_{ba_1\ldots a_G}$ denotes the number of sampling units assigned to treatment combination $a_1 \ldots a_G$ in block b and $n_{b+\ldots+}$ the total number of sampling units in block b.

Now $\hat{\bar{Y}}_{a_1\ldots a_G;greg}$ denotes the generalised regression estimator for $\bar{Y}_{a_1\ldots a_G}$ based on the observations obtained in subsample $s_{a_1\ldots a_G}$ and is defined analogously to expression (22). These $M_1 M_2 \ldots M_G$ generalised regression estimators are collected in the vector $\hat{\bar{\mathbf{Y}}}_{GREG} = (\hat{\bar{Y}}_{1\ldots1;greg}, \ldots, \hat{\bar{Y}}_{a_1\ldots a_G;greg}, \ldots, \hat{\bar{Y}}_{M_1\ldots M_G;greg})^t$ 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 $\hat{\bar{\mathbf{Y}}}_{GREG}$ are defined by (26), where the diagonal elements of $\hat{\mathbf{D}}$ are defined analogously to expression (27) in the case of a CRD or (28) in the case of an RBD. Finally hypotheses about main effects and interactions are tested with the Wald statistic (32), 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 defined above, see Table 1. As an example, the contrast matrices of the main effects and interactions in a factorial design with four factors are given in Table 2.

Table 2: *Constrasts in an $M_1 \times M_2 \times M_3 \times M_4$ factorial design*

Contrast matrix	Number of contrasts (degrees of freedom)
$\mathbf{C_{A_1}} = \frac{1}{M_2 M_3 M_4}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_1 - 1$
$\mathbf{C_{A_2}} = \frac{1}{M_1 M_3 M_4}\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$M_2 - 1$
$\mathbf{C_{A_3}} = \frac{1}{M_1 M_2 M_4}\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \mathbf{j}_{M_4}^t$	$M_3 - 1$
$\mathbf{C_{A_4}} = \frac{1}{M_1 M_2 M_3}\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$M_4 - 1$
$\mathbf{C_{A_1A_2}} = \frac{1}{M_3 M_4}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \mathbf{j}_{M_3}^t \otimes \mathbf{j}_{M_4}^t$	$(M_1 - 1)(M_2 - 1)$
$\mathbf{C_{A_1A_3}} = \frac{1}{M_2 M_4}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \mathbf{j}_{M_4}^t$	$(M_1 - 1)(M_3 - 1)$
$\mathbf{C_{A_1A_4}} = \frac{1}{M_2 M_3}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \mathbf{j}_{M_2}^t \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_1 - 1)(M_4 - 1)$
$\mathbf{C_{A_2A_3}} = \frac{1}{M_1 M_4}\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \mathbf{j}_{M_4}^t$	$(M_2 - 1)(M_3 - 1)$
$\mathbf{C_{A_2A_4}} = \frac{1}{M_1 M_3}\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_2 - 1)(M_4 - 1)$
$\mathbf{C_{A_3A_4}} = \frac{1}{M_1 M_2}\mathbf{j}_{M_1}^t \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_3 - 1)(M_4 - 1)$
$\mathbf{C_{A_1A_2A_3}} = \frac{1}{M_4}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \mathbf{j}_{M_4}^t$	$(M_1 - 1)(M_2 - 1)(M_3 - 1)$
$\mathbf{C_{A_1A_2A_4}} = \frac{1}{M_3}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \mathbf{j}_{M_3}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_1 - 1)(M_2 - 1)(M_4 - 1)$
$\mathbf{C_{A_1A_3A_4}} = \frac{1}{M_2}\tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \mathbf{j}_{M_2}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_1 - 1)(M_3 - 1)(M_4 - 1)$
$\mathbf{C_{A_2A_3A_4}} = \frac{1}{M_1}\mathbf{j}_{M_1}^t \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_2 - 1)(M_3 - 1)(M_4 - 1)$
$\mathbf{C_{A_1A_2A_3A_4}} = \tilde{\mathbf{C}}_{\mathbf{A_1}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_2}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_3}} \otimes \tilde{\mathbf{C}}_{\mathbf{A_4}}$	$(M_1 - 1)(M_2 - 1)(M_3 - 1)(M_4 - 1)$

9 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 CRD's and RBD's 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 $M_1 \times M_2 \times \ldots \times M_G$ 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 are 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 generalised regression estimator for $\bar{Y}_{a_1 \ldots a_G}$ is defined analogously to expression (22). Design-based estimators for the covariance matrices of the contrasts between the elements of $\hat{\bar{\mathbf{Y}}}_{GREG}$ are defined by (26), 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 generalised regression estimators can be constructed for each treatment combination. Design-based estimators for the covariance matrices of the contrasts between the ratios are defined by (26), 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 regression estimators. Hypotheses about main effects and interactions are tested with the Wald statistic (32).

10 Application to the Dutch LFS

In this section an experiment with different advance letters embedded in the Dutch Labor Force Survey is described. The survey design of the LFS is summarized in section 10.1. The purpose of this experiment is explained in section 10.2. The experimental design and the analysis results are described in sections 10.3 and 10.4 respectively.

10.1 Survey design

The LFS is based on a rotating panel survey. Each month a stratified two-stage cluster sample of about 6.500 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. Addresses of people aged 65 and over are under-sampled, since the target parameters of the LFS concern people aged 15 through 64. Finally the sample size on a regional level is adapted to the available field staff capacity. All households, with a maximum of three, residing on an address, are included in the sample. In the first wave, data are collected by means of computer assisted personal interviewing (CAPI) using laptops. Interviewers collect data for the LFS in areas close to where they live. The respondents are re-interviewed four times at quarterly intervals. In these four subsequent waves, data are collected by means of computer assisted telephone interviewing (CATI).

The weighting procedure of the LFS is based on the generalized regression estimator of Särndal et al. (1992). The inclusion probabilities reflect the under-sampling of addresses described above as well as the different response rates between geographical regions. The weighting scheme is based on a combination of different socio-demographic categorical variables. The integrated method for weighting persons and families of Lemaître and Dufour (1987) is applied to obtain equal weights for persons belonging to the same household.

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

10.2 Purpose

Advance letters are one of the design parameters of a survey that affect response rates and cooperation of respondents, De Leeuw et al. (2007). The standard advance letter of the Dutch Labour Force Survey (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, for example, Dillman (2007), Groves, Cialdini and Couper (2001) and Groves and Couper (1998). In an attempt to improve the LFS response rates, Luiten et al. (2008) conducted an experiment to test the effect of different advance letters for the LFS that better meet these principles about survey participation.

The first objective of this experiment was to test the effect of personalisation of the advance letter by addressing the letter to a named individual instead of the standard approach where the letter is addressed to "the occupants of the accommodation". It is anticipated that named letters are more likely to be read and therefore increase response rates and survey participation. The second objective was to compare the effect of two alternative variations on the content of the letter with the standard letter. The first alternative is based on the previously mentioned theories regarding survey participation. More specifically, 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. Call-back surveys of nonrespondents indicate that the autocratic tone of the letter is an important reason to refuse to participate in the LFS. In an earlier experiment, however, it was concluded that a more informal letter resulted in significantly smaller response rates, Van den Brakel (2008), section 2.4. This indicates that groups of individuals react differently on formal and informal letters. The second alternative further improves the content of the letter in first alternative using the theory of Language and Behavior Profile, Charvet, (1997) and attempts to be appealing to different kinds of people. The wording of the letter was adapted using influencing language so that it would attract people that react positively to formal as well as informal letters. See Luiten et al. (2008) for more details and the three versions of the advance letters.

Before a new advance letter is implemented as a standard in the Dutch LFS, its effect on response behavior and response bias must be quantified. Therefore the effects of the alternative letters on response behavior are analyzed using logistic regression analysis. The design-based procedures developed in the preceding sections are applied to analyze possible effects in the estimated unemployed labor force.

10.3 Experimental design

The two objectives described in subsection 10.2 result in two treatment factors. The first factor, 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 versus a named letter. The second factor, say B, concerns the content of the letter on three levels, i.e. the standard formal letter versus the two alternative letters described in the second last paragraph of subsection 10.2. Both

factors are combined in a 2×3 factorial design resulting in six treatment combinations, which are summarized in Table 3.

Table 3: *Treatment combinations experiment with advance letters*

Treatment nr.	Factor A (salutation)	Factor B (content)	Subsample fraction
1	A_1 unnamed (standard)	B_1 formal (standard)	5/6
2	A_1 unnamed (standard)	B_2 first alternative	1/30
3	A_1 unnamed (standard)	B_3 second alternative	1/30
4	A_2 named	B_1 formal (standard)	1/30
5	A_2 named	B_2 first alternative	1/30
6	A_2 named	B_3 second alternative	1/30

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 83 percent of the sample is assigned to the regular advance letter. The remaining 17 percent is assigned to the other five alternative treatment combinations. The fractions that were used to split the sample in six subsamples are specified in Table 3.

10.4 Results

The purpose of this experiment was to test hypotheses about the different versions of the advance letters on response rates and response bias. Luiten et al. (2008) analyzed differences in response rates using a chi-squared test for association in contingency tables. This approach doesn't account for the factorial setup and the block design of the experiment. Therefore a logistic regression analysis will be applied to test hypotheses about effects on response rates. To investigate the effect on response bias, it is tested whether the parameter estimates for the unemployed labor force obtained with the subsamples assigned to the six treatment combinations are significantly different using the design-based procedures for factorial designs developed in this paper.

Table 4 contains an overview of the response account of the six subsamples in the experiment. The households are classified in completely responding households (response), partially responding households, refusals, no contact, frame error and a rest category. Frame errors contain unable to locate addresses, in construction, no housing units, or vacant housing units. The rest category contains nonresponding households due to language problems, and no opportunity. Since the experiment is based on an unbalanced design, crude and adjusted proportions are reported for the margins in Table 4. Crude proportions are not corrected for the unequal distribution of observations over the different treatment combinations. As a result the crude marginal proportions are dominated by the large sample size assigned to the regular advance letter. Adjusted proportions are corrected for unbalanced allocation of the sample over the treatment combinations.

It follows from the results in Table 4 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. The alternative letters considered in factor B did not increase the

response rate. The differences in response rates between the six treatment combinations are larger. Treatment combinations 2 and 5, e.g. differ about five percent points, which might indicate the presence of interactions between the two treatment factors.

Table 4: *Response account experiment with advance letters*

Treatment	response	part. resp.	refusal	no contact	frame error	rest	total
1 ($A_1 \times B_1$)	13234	183	5127	1843	1226	1733	23346
	56.69%	0.78%	21.96%	7.89%	5.25%	7.42%	
2 ($A_1 \times B_2$)	604	5	271	89	73	85	1127
	53.59%	0.44%	24.05%	7.90%	6.48%	7.54%	
3 ($A_1 \times B_3$)	635	10	254	93	62	73	1127
	56.34%	0.89%	22.54%	8.25%	5.50%	6.48%	
4 ($A_2 \times B_1$)	662	3	256	84	49	68	1122
	59.00%	0.27%	22.82%	7.49%	4.37%	6.06%	
5 ($A_2 \times B_2$)	663	9	236	80	47	87	1122
	59.09%	0.80%	21.03%	7.13%	4.19%	7.75%	
6 ($A_2 \times B_3$)	627	9	259	85	56	91	1127
	55.63%	0.80%	22.98%	7.54%	4.97%	8.07%	
No name (A_1)	14473	198	5652	2025	1361	1891	25600
crude prop.	56.54%	0.77%	22.08%	7.91%	5.32%	7.39%	
adjusted prop.	55.54%	0.70%	22.85%	8.01%	5.74%	7.15%	
Name (A_2)	1952	21	751	249	152	246	3371
crude/adj. prop.	57.91%	0.62%	22.28%	7.39%	4.51%	7.30%	
Stand. cont. (B_1)	13896	186	5383	1927	1275	1801	24468
crude prop.	56.79%	0.76%	22.00%	7.88%	5.21%	7.36%	
adjusted prop.	57.85%	0.53%	22.39%	7.69%	4.81%	6.74%	
Alt. 1 (B_2)	1267	14	507	169	120	172	2249
crude/adj. prop.	56.34%	0.62%	22.54%	7.51%	5.34%	7.65%	
Alt. 2 (B_3)	1262	19	513	178	118	164	2254
crude/adj. prop.	55.99%	0.84%	22.76%	7.90%	5.24%	7.28%	
Total	16425	219	6403	2274	1513	2137	28971
crude prop.	56.69%	0.76%	22.10%	7.85%	5.22%	7.38%	
adjusted prop.	56.73%	0.66%	22.57%	7.70%	5.13%	7.23%	

Response behavior is modeled in a logistic regression model. This analysis serves two purposes. Firstly, hypotheses about the effect of the two treatment factors on response behavior may be tested. Secondly, additional information may be obtained on whether the factors increase the response across the entire target population or that specific groups react differently on 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 five response categories. The response behavior is assumed to depend upon:

- a general mean,
- treatment factor A (name) in two categories specifying the standard letter addressed to the occupants of the accommodation versus the letter addressed to a named individual,
- treatment factor B (content) in three categories specifying the standard advance letter versus two alternative letters,
- a block variable in 13 categories (interviewers are the block variable, but adjacent interviewer regions are collapsed in 13 blocks),
- 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 rest category,
  - accommodation type in nine categories.

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 of estimation results for response rates. For brevity, the regression coefficients with their standard errors and test statistics for separate categories are only expressed for the treatment factors.

The logistic regression analysis shows that the hypothesis that there are no interactions between the two treatment factors cannot be rejected (p-value Wald statistic equals 0.121). Therefore this interaction term is excluded from the finally selected model. From Table 5 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, and using influencing language. This negative effect on the response rates is, on the other hand, in line with the results of an earlier experiment where the response of a more informal advance letter of the LFS resulted in significant 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: *Logistic regression analysis for response rates*

Parameter	Coefficient	Standard error	Wald statistic	Degrees of freedom	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

The second step in the analysis of this field experiment is to test 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 in this analysis to account for the sampling design and the estimation procedure of the LFS. The generalized regression estimator is applied to obtain estimates for the unemployed labor force under the six different treatment combinations in the first wave. The inclusion probabilities reflect the sampling design of the LFS and the experimental design that is 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 6. 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 7.

Table 6: *Point and variance estimates unemployed labor force*

Treatment nr.	Estimate	Variance
1	$\hat{\bar{Y}}_{11;greg} = 4.100$	$\hat{d}_{11} = 0.021$
2	$\hat{\bar{Y}}_{12;greg} = 3.761$	$\hat{d}_{12} = 0.417$
3	$\hat{\bar{Y}}_{13;greg} = 5.264$	$\hat{d}_{13} = 0.567$
4	$\hat{\bar{Y}}_{21;greg} = 3.609$	$\hat{d}_{21} = 0.370$
5	$\hat{\bar{Y}}_{22;greg} = 4.546$	$\hat{d}_{22} = 0.443$
6	$\hat{\bar{Y}}_{23;greg} = 3.385$	$\hat{d}_{23} = 0.441$

Table 7: *Analysis main effects and interactions unemployed labor force*

Source	Estimate $\mathbf{C}\hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}$	Variance matrix $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$	Wald statistic	Degrees of freedom	p-value
Treatment A (name) $A_1 - A_2$	0.528	[1,1]=0.251	1.109	1	0.292
Treatment B (content)			0.732	2	0.694
$B_1 - B_2$	-0.300	[1,1]= 0.313			
$B_1 - B_3$	-0.471	[2,2]= 0.350			
		[1,2]= 0.098			
Interaction			3.801	2	0.150
$AB_{11} - AB_{12} - AB_{21} + AB_{22}$	1.276	[1,1]=1.252			
$AB_{11} - AB_{13} - AB_{21} + AB_{23}$	-1.388	[2,2]=1.400			
		[1,2]=0.392			

From the analysis results, summarized in Table 7, 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 is no empirical evidence that the alternative letters, considered in this experiment, improve response behavior or the accuracy of the estimates for target variables like the unemployed labor force. Therefore it was decided not to adapt the standard advance letter of the LFS.

11 Discusssion

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. Finally factorial designs are more efficient compared to single-factor experiments, since less 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. Design-based analyses 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 and 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.

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 an F-test for the block effects is not valid. In these cases alternative measures to evaluate the efficiency of an RBD are available; see e.g. Montgomery (2001). In the design-based theory developed for RBD's 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.

As a numerical example, the methods developed in this paper are applied to a 2×3 factorial design embedded in the Dutch LFS to test the effect of six different versions of an advance letter. The parameter and variance estimators developed in this paper are implemented in the software package X-tool, which is available as a component of the Blaise survey processing software package, Statistics Netherlands (2002). This component supports the analysis of single-factor experiments, see Van den Brakel (2008) for details. The Wald statistics for the different hypotheses about main effects and interactions for factorial designs are not implemented yet and require a separate matrix programming package for the moment.

References

[1] Charvet, S.R. (1997), *Words that change minds: Mastering the language of influence*, Dubuque: Kendal Hunt.

[2] Cochran, W.G. (1977), *Sampling Techniques* (3rd Ed.), New York: John Wiley.

[3] Cochran, W.G., and Cox, G.M. (1957), *Experimental Designs*, New York: John Wiley.

[4] Dillman, D.A. (2007), *Mail and internet surveys: The tailored design method*, New York, John Wiley.

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

[6] Fellegi, I.P. (1964), "Response Variance and its Estimation", *Journal of the American Statistical Association*, 59, 1016-1041.

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

[8] 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.

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

[10] 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.

[11] 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.

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

[13] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), *Sample Survey Methods and Theory*, Vol I and II. New York: John Wiley.

[14] Hájek, J. (1971), Comment on "An essay on the logical foundations of survey sampling" by Basu, D., in *Foundations of Statistical Inference* (Eds. Godambe, V.P., and Sprott D.A.), Toronto: Holt, Rinehart, and Winston.

[15] 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.*

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

[17] Hinkelmann, K., and Kempthorne, O. (2005), *Design and Analysis of Experiments, Volume 2: Advanced experimental design*, New York: John Wiley.

[18] 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.

[19] Hartley, H.O., and Rao, J.N.K. (1978), "Estimation of nonsampling variance components in sample surveys" (Eds. Namboodiri, N.K.), *Survey Sampling and Measurement*, New York: Academic Press, 35-43.

[20] Kempthorne, O. (1955), "The Randomization Theory of Experimental Inference", *Journal of the American Statistical Assoication*, 50, 946-967.

[21] Kish, L. (1965), *Survey Sampling*, New York: John Wiley.

[22] 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.

[23] Lemaître, G., and Dufour, J. (1987), "An Integrated Method for Weighting Persons and Families", *Survey Methodology*, 13, 199-207.

[24] 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.

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

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

[27] Morisson, D.F. (1990), *Multivariate Statistical Methods*, Singapore: McGraw-Hill.

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

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

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

[31] Searle, S.R. (1971). Linear Models. Wiley, New-York.

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

[33] Statistics Netherlands (2002), *Blaise Developer's Guide*, Heerlen: Statistics Netherlands. (Available from http://www.Blaise.com)

[34] Van den Brakel, J.A. (2001), *Design and Analysis of Experiments Embedded in Complex Sample Surveys*, Ph.D. Thesis, Rotterdam: Erasmus University of Rotterdam.

[35] 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.

[36] 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.

[37] 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.

[38] Van den Brakel. J.A., Vis-Visschers, R., and Schmeets, J.J.G. (2006), "An Experiment with Data Collection Modes and Incentives in the Dutch Family and Fertility Survey for Young Moroccans and Turks", *Field Methods*, 18, 321-334.

[39] 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.

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

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

Appendix: Variances of treatment effects

Generalized regression estimator

The first order Taylor approximation of the generalized regression estimator (22) can be expressed as

$$\hat{\bar{Y}}_{kl;greg} \doteq \hat{\bar{E}}_{kl} + \mathbf{b}_{kl}^t \bar{\mathbf{X}}, \tag{43}$$

according to (24) with

$$\hat{\bar{E}}_{kl} = \hat{\bar{Y}}_{kl} - \mathbf{b}_{kl}^t \hat{\bar{\mathbf{X}}} = \sum_{i \in s} \left(\frac{\mathbf{p}_{ikl}^t (\mathbf{y}_i - \mathbf{B}^t \mathbf{x}_i)}{\pi_i \; N} \right). \tag{44}$$

Here $\mathbf{B}$ denotes a $H \times KL$ matrix whose columns are the $H-$vectors $\mathbf{b}_{kl}$ and $\mathbf{p}_{ikl}$ are $KL-$vectors that describe the randomization mechanism of the experimental design. For a CRD, it follows that

$$\mathbf{p}_{ikl} \equiv \begin{cases} \frac{n_{++}}{n_{kl}} \mathbf{r}_{kl} & \text{if} \quad i \in s_{kl} \\ \mathbf{0} & \text{if} \quad i \notin s_{kl} \end{cases}, \tag{45}$$

and for an RBD

$$\mathbf{p}_{ikl} \equiv \begin{cases} \frac{n_{b++}}{n_{bkl}} \mathbf{r}_{kl} & \text{if} \quad i \in s_{bkl} \\ \mathbf{0} & \text{if} \quad i \notin s_{bkl} \end{cases}, \tag{46}$$

where $\mathbf{r}_{kl}$ denotes the unit vector of order KL with the $kl-$th element equal to one and the other elements equal to zero and $\mathbf{0}$ a vector of order KL with each element equal to zero.

Properties of the randomization vectors $\mathbf{p}_{ikl}$

In this subsection the statistical properties of the randomization vectors $\mathbf{p}_{ikl}$ are derived. They are required to derive the covariance matrix of the contrasts between the generalized regression estimators for the different treatment combinations. From the randomization mechanism of a CRD, the following probability mass function for $\mathbf{p}_{ikl}$ can be derived:

$$P\left(\mathbf{p}_{ikl} = \frac{n_{++}}{n_{kl}} \mathbf{r}_{kl} \mid s\right) = \frac{n_{kl}}{n_{++}}, \quad \text{and} \quad P\left(\mathbf{p}_{ikl} = \mathbf{0} \mid s\right) = 1 - \frac{n_{kl}}{n_{++}}.$$

The $\mathbf{p}_{ikl}$ for an RBD have the following probability mass function

$$P\left(\mathbf{p}_{ikl} = \frac{n_{b++}}{n_{bkl}} \mathbf{r}_{kl} \mid s_b\right) = \frac{n_{bkl}}{n_{b++}}, \quad \text{and} \quad P\left(\mathbf{p}_{ikl} = \mathbf{0} \mid s_b\right) = 1 - \frac{n_{bkl}}{n_{b++}}.$$

The expectation of $\mathbf{p}_{ikl}$ with respect to the experimental design for CRD is given by:

$$\mathrm{E}_{\varepsilon}(\mathbf{p}_{ikl}) = P\left(\mathbf{p}_{ikl} = \frac{n_{++}}{n_{kl}}\mathbf{r}_{kl}\right)\frac{n_{++}}{n_{kl}}\mathbf{r}_{kl} + P\left(\mathbf{p}_{ikl} = \mathbf{0}\right)\mathbf{0} = \mathbf{r}_{kl}.$$

For an RBD, the expectation of $\mathbf{p}_{ikl}$ with respect to the experimental design is given by:

$$\mathrm{E}_{\varepsilon}(\mathbf{p}_{ikl}) = P\left(\mathbf{p}_{ikl} = \frac{n_{b++}}{n_{bkl}}\mathbf{r}_{kl}\right)\frac{n_{b++}}{n_{bkl}}\mathbf{r}_{kl} + P\left(\mathbf{p}_{ikl} = \mathbf{0}\right)\mathbf{0} = \mathbf{r}_{kl}.$$

Equivalent to the derivation for these randomization vectors given by Van den Brakel and Renssen (2005) it follows that the following covariances with respect to the experimental design of a factorial CRD can be derived:

$$\begin{aligned}
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{ikl}^{t}) = \frac{(n_{++} - n_{kl})}{n_{kl}}\mathbf{r}_{kl}\mathbf{r}_{kl}^{t},\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{i'kl}^{t}) = -\frac{(n_{++} - n_{kl})}{n_{kl}}\frac{1}{(n_{++} - 1)}\mathbf{r}_{kl}\mathbf{r}_{kl}^{t},\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{ik'l'}^{t}) = -\mathbf{r}_{kl}\mathbf{r}_{k'l'}^{t}, \quad \text{for} \quad k \neq k' \quad \text{or} \quad l \neq l',\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{i'k'l'}^{t}) = \frac{1}{(n_{++} - 1)}\mathbf{r}_{kl}\mathbf{r}_{k'l'}^{t}, \quad \text{for} \quad k \neq k' \quad \text{or} \quad l \neq l'.
\end{aligned}$$

Finally the following covariances with respect to the experimental design for a factorial RBD can be derived:

$$\begin{aligned}
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{ikl}^{t}) = \frac{(n_{b++} - n_{bkl})}{n_{bkl}}\mathbf{r}_{kl}\mathbf{r}_{kl}^{t},\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{i'kl}^{t}) = -\frac{(n_{b++} - n_{bkl})}{n_{bkl}}\frac{1}{(n_{b++} - 1)}\mathbf{r}_{kl}\mathbf{r}_{kl}^{t}, \quad \text{if} \quad i \in s_b \quad \text{and} \quad i' \in s_b\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{i'kl}^{t}) = \mathbf{O}, \quad \text{if} \quad i \in s_b \quad \text{and} \quad i' \in s_{b'}\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{ik'l'}^{t}) = -\mathbf{r}_{kl}\mathbf{r}_{k'l'}^{t}, \quad \text{for} \quad k \neq k' \quad \text{or} \quad l \neq l',\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{i'k'l'}^{t}) = \frac{1}{(n_{b++} - 1)}\mathbf{r}_{kl}\mathbf{r}_{k'l'}^{t}, \quad \text{if} \quad i \in s_b \quad \text{and} \quad i' \in s_b; \quad \text{for} \quad k \neq k' \quad \text{or} \quad l \neq l'.\\
&\mathrm{Cov}_{\varepsilon}(\mathbf{p}_{ikl}, \mathbf{p}_{i'k'l'}^{t}) = \mathbf{O}, \quad \text{if} \quad i \in s_b \quad \text{and} \quad i' \in s_{b'}; \quad \text{for} \quad k \neq k' \quad \text{or} \quad l \neq l'.
\end{aligned}$$

Covariance matrix of contrasts between GREG estimates

The covariance matrix of the contrasts between the elements of $\hat{\hat{\mathbf{Y}}}_{GREG}$ is obtained by deriving the covariance matrix of $\mathbf{C}\hat{\hat{\mathbf{E}}}$, where $\hat{\hat{\mathbf{E}}}$ is the $KL-$ vector with elements $\hat{\hat{E}}_{kl}$, defined by (44). Conditioning on the realization of the measurement error model (1) and the sample s results into the following covariance decomposition:

$$\mathrm{Cov}(\mathbf{C}\hat{\hat{\mathbf{Y}}}_{GREG}) = \mathrm{Cov}_m\mathrm{E}_s\mathrm{E}_{\varepsilon}(\mathbf{C}\hat{\hat{\mathbf{E}}} \mid m, s) + \mathrm{E}_m\mathrm{Cov}_s\mathrm{E}_{\varepsilon}(\mathbf{C}\hat{\hat{\mathbf{E}}} \mid m, s) + \mathrm{E}_m\mathrm{E}_s\mathrm{Cov}_{\varepsilon}(\mathbf{C}\hat{\hat{\mathbf{E}}} \mid m, s). \quad (47)$$

Under the condition that a constant $H-$vector $\mathbf{a}$ exsists such that $\mathbf{a}^t\mathbf{x}_i = 1$ for all $i \in U$, it follows for $\mathbf{b}_{kl}$ in (25) that $\mathbf{b}_{kl} = \mathbf{b} + \mathbf{a}(\psi + \beta_{kl})$, where $\mathbf{b}$ is defined by (19). Subsequently it follows that $\mathbf{B}^t\mathbf{x}_i = \mathbf{j}_{KL}(\mathbf{b}^t\mathbf{x}_i + \psi) + \boldsymbol{\beta}$. Since $\mathbf{C}\mathbf{j}_{KL} = \mathbf{0}$, it follows from measurement error model (1) that

$$\mathbf{C}\left(\mathbf{y}_i - \mathbf{B}^t\mathbf{x}_i\right) = \mathbf{C}\left(\mathbf{j}_{KL}(u_i + \gamma_q) + \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i - \mathbf{j}_{KL}(\mathbf{b}^t\mathbf{x}_i + \psi) - \boldsymbol{\beta}\right) = \mathbf{C}\boldsymbol{\varepsilon}_i. \quad (48)$$

Taking covariances and expectations over the three components in (47) and taking advantage of result (48), the expectations and covariances of the randomization vectors $\mathbf{p}_{ikl}$ and the assumptions of the measurement error model (2), (3), (4), (5), and (6) gives:

$$\mathrm{Cov}_m \mathrm{E}_s \mathrm{E}_\varepsilon(\mathbf{C}\hat{\tilde{\mathbf{E}}} \mid m, s) = \frac{1}{N^2} \sum_{i=1}^{N} \mathbf{C}\boldsymbol{\Sigma}_i \mathbf{C}^t, \tag{49}$$

$$\mathrm{E}_m \mathrm{Cov}_s \mathrm{E}_\varepsilon(\mathbf{C}\hat{\tilde{\mathbf{E}}} \mid m, s) = \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1 \right) \mathbf{C}\boldsymbol{\Sigma}_i \mathbf{C}^t, \tag{50}$$

$$\mathrm{E}_m \mathrm{E}_s \mathrm{Cov}_\varepsilon(\mathbf{C}\hat{\tilde{\mathbf{E}}} \mid m, s) = \mathrm{E}_m \mathrm{E}_s(\mathbf{C}\mathbf{D}\mathbf{C}^t) - \frac{1}{N^2} \sum_{i=1}^{N} \frac{\mathbf{C}\boldsymbol{\Sigma}_i \mathbf{C}^t}{\pi_i}. \tag{51}$$

In (51), $\mathbf{D}$ denotes a $KL \times KL$ diagonal matrix with diagonal elements

$$d_{kl} = \frac{1}{n_{kl}} \frac{1}{n_{++} - 1} \sum_{i=1}^{n_{++}} \left(\frac{n_{++}(y_{ikl} - \mathbf{b}_{kl}^t \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{++}} \sum_{i'=1}^{n_{++}} \frac{n_{++}(y_{i'kl} - \mathbf{b}_{kl}^t \mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{52}$$

in the case of a CRD and

$$d_{kl} = \sum_{b=1}^{B} \frac{1}{n_{bkl}} \frac{1}{n_{b++} - 1} \sum_{i=1}^{n_{b++}} \left(\frac{n_{b++}(y_{ikl} - \mathbf{b}_{kl}^t \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{b++}} \sum_{i'=1}^{n_{b++}} \frac{n_{b++}(y_{i'kl} - \mathbf{b}_{kl}^t \mathbf{x}_{i'})}{N\pi_{i'}} \right)^2, \tag{53}$$

in the case of an RBD. Insersting (49), (50), and (51) in (47) gives

$$\mathrm{Cov}(\mathbf{C}\hat{\tilde{\mathbf{Y}}}_{GREG}) = \mathrm{E}_m \mathrm{E}_s \mathbf{C}\mathbf{D}\mathbf{C}^t. \tag{54}$$

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 (54) is conveniently stated implicitly as the expectation over the measurement error model and the sampling design. In the case of a CRD, the allocation of the experimental units to subsample s_{kl} can be considered as simple random sampling without replacement from s. Consequently, an unbiased estimator for d_{kl} in the case of a CRD is given by (27). Equivalently, for an RBD the allocation of the experimental units to subsample s_{kl} can be considered as stratified simple random sampling without replacement from s, where the strata are the blocks and an unbiased estimator for d_{kl} is given by (28).