



**Discussion Paper**

# **A Bayesian analysis of design parameters in survey data collection**

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

**2017 | 14**

**Schouten, B.,  
Mushkudiani, N.,  
Shlomo, N., Durrant, G.,  
Lundquist, P. and  
Wagner, J.**

# Content

<b>1. Introduction</b>	<b>4</b>
<b>2. Adaptive strategies and design parameters</b>	<b>6</b>
2.1 Types of strategies and notation	6
2.2 Survey design parameters	7
<b>3. Modelling survey design parameters</b>	<b>8</b>
3.1 Decomposition of survey design parameters	8
3.2 Models for survey design parameter components	10
<b>4. A Bayesian analysis of survey design parameters</b>	<b>12</b>
4.1 Prior distributions	12
4.2 Posterior distributions	13
4.3 A Bayesian analysis of functions of survey design parameters	15
<b>5. A simulation study to investigate the utility of Bayesian analysis</b>	<b>16</b>
5.1 Design of the simulation study	16
5.2 Convergence properties	18
5.3 Simulation results	20
<b>6. Discussion</b>	<b>25</b>
<b>Acknowledgements</b>	<b>27</b>
<b>References</b>	<b>27</b>
<b>Appendix A: The full conditionals of the regression parameters in the Gibbs sampler</b>	<b>29</b>
A.1 Full conditionals for regression parameters in response propensity models	29
A.2. Full conditionals for regression parameters in costs models	31
<b>Appendix B: Elicitation of hyperparameters in the Health Survey simulation study</b>	<b>32</b>
<b>Appendix C: Simulation study propensities and costs</b>	<b>33</b>

### **Summary**

In the design of surveys a number of input parameters, such as contact propensities, participation propensities and costs per sample unit, play a decisive role. In on-going surveys, these survey design parameters are usually estimated from previous experience and updated gradually with new experience. In new surveys, these parameters are estimated from expert opinion and experience with similar surveys. Although survey institutes have a fair expertise and experience, the postulation, estimation and updating of survey design parameters is rarely done in a systematic way. This paper presents a Bayesian framework to include and update prior knowledge and expert opinion about the parameters. This framework is set in the context of adaptive survey designs in which different population units may receive different treatment given quality and cost objectives. For this type of survey, the accuracy of design parameters becomes even more crucial to effective design decisions. The framework allows for a Bayesian analysis of the performance of a survey during data collection and in between waves of a survey. We demonstrate the utility of the Bayesian analysis using a simulation study based on the Dutch Health Survey.

### **Keywords**

Survey monitoring, Survey quality, Survey costs, Nonresponse, Adaptive survey design;

# 1. Introduction

Over the last two decades, there has been an increasing interest in survey data collection monitoring, analysis, and intervention or adaptation. The main drivers for this are the diversification of data collection that followed the emergence of online communication, the lack of predictability of survey response propensities despite years of research into survey design, the gradual increase in costs per respondent when response rates are kept at traditional levels, and the availability of a wide range of data collection process data (termed *paradata*). See, for instance, Kreuter (2013) to get a view on research aiming at a deeper understanding of data collection processes. In a changing survey data collection environment with unpredictable, and, hence, only partially controllable outcomes, a close watch at the progress of data collection is imperative. For this purpose, timely and accurate estimates of survey design parameters, such as contact propensities, participation propensities, costs per contact and costs per interview, are needed.

These developments paved the way for survey designs that adapt strategies to known and relevant characteristics of sampled units from the target population. Such designs we term *adaptive*, see Groves and Heeringa (2006), Wagner (2008) and Schouten, Calinescu and Luiten (2013). In order to adapt, accurate estimates of survey design parameters are needed at the deeper level of population subgroups. This higher resolution puts a pressure on the accuracy of such parameter estimates. Burger, Perryck and Schouten (2016) analysed the performance of adaptive survey designs under inaccurate design parameter estimates and concluded, not surprisingly, that biased parameter estimates may lead to suboptimal, and, consequently, inefficient designs.

A natural approach to evaluate inaccuracy of survey design parameters and to account for the uncertainty in the optimization of survey design is through Bayesian analysis (Gelman et al 2014). In such an analysis, survey design parameters are treated as random variables and are assigned prior distributions, which are then updated and transformed to posterior distributions during data collection. These posterior distributions may be used as prior distributions in new waves of the same survey. The added benefit is that prior distributions may also be elicited from expert judgment, so that historic survey data in one survey may also be re-used in other surveys.

Despite these advantages, Bayesian analysis of survey data collection is rarely done and literature is very thin. An exception is Schafer (2013). Reasons for this absence may be that a Bayesian analysis is not straightforward, conceptually and/or computationally, that the added value may be unclear to survey designers and that the elicitation of prior distributions is complex and cumbersome in practice.

Therefore, we have the following objectives:

1. To set up (sufficiently general) models for survey design parameters;
2. To introduce a Bayesian analysis of survey design parameters;
3. To introduce a Bayesian analysis of quality and cost indicators based on survey design parameters, and;
4. To evaluate under what conditions a Bayesian analysis has added value;

We show that a Bayesian analysis can be set up and can be applied to a variety of survey designs. Our focus on overall quality and cost indicators is motivated by the desire to ultimately adapt data collection strategies to different population subgroups. To reach this goal, we include linked auxiliary data and paradata in models for survey design parameters. The main objective here, however, is to show that Bayesian analysis can be effective in monitoring and analysis of survey data collection.

A Bayesian analysis of data collection is by itself not novel. There is a vast literature in biostatistics and medical statistics that presents methodology to monitor and optimize treatments using prior knowledge or beliefs. There is a close resemblance to dynamic treatment regimes and continual re-assessment methods in clinical trials (e.g. Heyd and Carlin 1999, O’Quigley and Shen 1996, Murphy 2003 and 2005, Scharfstein, Daniels and Robins 2003, and Schulte et al 2014). An application to survey data collection is, however, novel, and introduces three specific elements: a potential large and/or diverse set of survey target variables, a very explicit focus on data collection costs, and a multitude of quality indicators describing different survey errors. In the analysis, we assume that survey data collection consists of a series of phases for which costs and quality are evaluated separately and cumulatively, including all outcomes up to a current phase. Per phase, survey design parameters are defined and are assigned prior distributions. A Gibbs sampler with data augmentation is applied to derive posterior distributions.

The explicit presentation of uncertainty in survey design parameter values is an advantage of Bayesian analysis by itself. However, a Bayesian analysis has even more added value when the inclusion of prior knowledge from historic survey data and expert knowledge provides stronger guidance to design decisions than a non-Bayesian analysis. In order to prove this, we take two approaches in a simulation study linked to the Dutch Health Survey: We gradually misspecify the locations of the prior distributions and gradually increase the variance of the prior distributions. We do this to analyse if, and when, the prior information is too weak to be of use, and may just as well be replaced by a fully non-informative prior distribution, which essentially amounts to a non-Bayesian analysis.

A natural subsequent step is to adapt survey design within the Bayesian analysis framework. In order to be able to do so, a range of available strategies needs to be randomized. Below we shortly discuss randomization of strategies in the Bayesian analysis and we distinguish contact and participation in obtaining response, but we leave actual optimization to future papers.

We focus in this paper on adaptive survey design with the objective to minimize nonresponse error. Survey design parameters associated with measurement error, e.g. the adjusted mode effect, are out of scope in this paper. We refer to Calinescu (2013) and Calinescu and Schouten (2015) for designs incorporating both types of survey errors.

This paper has three main sections: In section 2, we describe the various adaptive survey design strategies and link them to response propensities and costs, the survey design parameters of interest. In section 3, we break down the response propensities and costs into their basic components, e.g. contact propensities and cost components per call and present models for these components. In section 4, we assign prior distributions to the regression coefficients and derive expressions for the

corresponding posterior distributions. Apart from the survey design parameters themselves, we also consider a number of functions of these parameters like the response rate, overall costs and coefficient of variation of the response propensities. In section 5, we investigate the utility of Bayesian analysis through a simulation study on the Dutch Health Survey.

## 2. Adaptive strategies and design parameters

In this section, we provide the necessary background for the models of section 3; we discuss the various adaptive strategies and the corresponding survey design parameters.

In order to clarify notation and illustrate ideas, we use the 2015 Dutch Health Survey (HS), see CBS (2016). The HS has a sequential mixed-mode survey design with Web followed by face-to-face interviewing, i.e. non-respondents to a Web survey invitation are re-allocated to interviewers. We consider three data collection phases: Web, short face-to-face, and extended face-to-face. The extended face-to-face corresponds to an additional round of face-to-face visits for those sample units that have not been contacted after three face-to-face visits or that opted for another time because they were busy. The first three visits are termed short face-to-face. We label the three phases as Web, F2F-short, F2F-extended. During the online first design phase it is recorded whether a sample person breaks off the online interview and does not submit survey data. Furthermore, the age and gender of sample units are linked from the sampling frame.

### 2.1 Types of strategies and notation

The design of each survey has a range of features, e.g. sample design, advance letter, contact protocol, screener interview, number of phases, reminder protocol, use of incentive, mode of administration, interviewer, refusal conversion procedure and type of questionnaire. The total of choices made for the design features is called a data collection strategy or simply strategy. In non-adaptive survey designs, these features are implemented uniformly over the whole sample, i.e. there is one strategy. In adaptive survey designs, part of the design features may be implemented differently for different sample units, i.e. there is a set of strategies. See for example Groves and Heeringa (2006), Wagner (2008), Coffey, Reist and White (2013) and Schouten, Calinescu and Luiten (2013). In the HS example, the design features of interest are the set of survey modes and the number of interviewer visits. Different sample units may be distinguished based on linked auxiliary data from the sampling frame or administrative data, from paradata obtained during data collection, and/or from survey data from previous waves in longitudinal or panel settings. In

analogy to clinical trials, e.g. Scharfstein, Daniels and Robins (2003) and Murphy (2003 and 2005), an adaptive strategy based on auxiliary data available at the start of data collection is called static and an adaptive strategy based (also) on auxiliary data collected during data collection is called dynamic. For a subject  $i$  in the sample, we let  $x_i$  be the vector of auxiliary variables. We suppose the auxiliary vector of subject  $i$  consists of the following entries

$$x_i = (x_{0,1,i}, \dots, x_{0,m_0,i}, \dots, x_{T,1,i}, \dots, x_{T,m_T,i})',$$

where  $x_{0,i} = (x_{0,1,i}, \dots, x_{0,m_0,i})'$  contains the  $m_0$  auxiliary variables available at the start of data collection, and  $x_{t,i} = (x_{t,1,i}, \dots, x_{t,m_t,i})'$  are the auxiliary variables that are observed for the sample units in phase  $t$ . In the HS example, the auxiliary data consists of age, gender and break-off. Break-off is a binary indicator for a person starting the Web survey but aborting it without submitting his/her answers. Since break-off is measured during data collection, any adaptation is dynamic.

Two archetypes of adaptive strategies may be distinguished: concurrent and sequential. A concurrent strategy has a single phase in which multiple actions are possible. A sequential strategy has two phases and after the first phase a follow-up is possible for those sample units that did not respond. These two strategies can be combined to build any adaptive strategy that has two or more phases. Let the survey design consist of a maximum of  $T$  phases, labelled  $t = 1, 2, \dots, T$ . (The use of the character  $t$  for the phase does not necessarily indicate it is related to time.) We define  $\mathcal{S}_t$  as the collection of all possible actions in phase  $t$  and let  $s_t$  represent the action in phase  $t$ . We define the total collection of possible actions:  $\mathcal{S} := \bigcup_{t=1}^T \mathcal{S}_t$ . The action sets may contain  $s_0$ , which, if selected, implies that no attempt is made to obtain a response. We include this action, since it can occur that in a certain phase some groups of respondents are not approached. We define the collection of survey strategies from phase 1 to  $T$

$$\mathcal{S}_{1,T} := \{(s_1, \dots, s_T) : s_t \in \mathcal{S}_t, t = 1, 2, \dots, T\}$$

and let  $s_{1,T} \in \mathcal{S}_{1,T}$  denote one possible strategy, i.e. sequence of actions. Observe that we have different notations for the action in phase  $t$ ,  $s_t$ , and the strategies up to phase  $t$ ,  $s_{1,t}$ . In the HS example, there are three phases and there are two sequential decisions: to use an interviewer follow-up and, subsequently, to use an extended interviewer follow-up.

## 2.2 Survey design parameters

Adaptive survey designs either maximize a quality objective subject to cost constraints and other quality constraints or minimize a cost objective subject to quality constraints. The quality and cost constraints depend on the setting in which the survey is conducted, but may concern any survey error. Three sets of survey design parameters suffice to compute most of the quality and cost constraints:

1. Response propensities,  $\rho_i(s_{1,T})$ , per unit  $i$  and strategy  $s_{1,T}$ ;
2. Costs,  $C_i(s_{1,T})$ , per unit  $i$  and strategy  $s_{1,T}$ ;
3. Method effects on specified survey statistics,  $D_i(s_{1,T})$ , per unit  $i$  and strategy  $s_{1,T}$ ;

In this paper, we concentrate on nonresponse error and costs and, to keep the scope manageable, we do not model key survey variables. For this reason, we do not consider the method effects on key statistics, like means or totals of the main survey variables, as these design parameters require models for survey variables. We leave this to future research. In the HS example, we will consider response rates, coefficients of variation of response propensities and costs, see section 3.4 for details. There are two options in defining and modelling survey design parameters: the stratum or subgroup level, and the individual sample unit level. The first option implies that the average response and costs in a stratum are modelled, i.e. addressing variation within such strata, whereas the second option implies that models for individual units are created. The two options, essentially, represent two main approaches in adaptive design, stratum allocation and case prioritization, e.g. Peytchev et al (2010), Wagner et al (2013), Rosen et al (2014), Luiten and Schouten (2013), Särndal and Lundquist (2014) and Schouten and Shlomo (2015). In this paper, we model individual design parameters, since it offers more flexibility. Any stratification may still be applied afterwards and stratum design parameters may be derived from the individual propensities and cost functions by taking averages over all individuals in the stratum.

### 3. Modelling survey design parameters

We construct hierarchical Bayes models for response propensities and costs per sample unit by first decomposing these parameters into their main components, linking these components through general linear models to the available auxiliary variables, and assigning prior distributions to the parameters in these models.

#### 3.1 Decomposition of survey design parameters

We will give basic models for response propensities and costs. We break down these parameters into their basic components: contact propensities participation propensities and costs. By introducing the contact and response propensities we serve one of our main goals, ensuring that the model is sufficiently general. Other outcomes than contact, non-contact, and refusal/participation are possible and can be included in a relatively straightforward way. However, the number of parameters to be estimated increases with each outcome that is included.

We make two general assumptions: First, we assume that making contact, obtaining participation and the costs associated with an individual sample unit are independent of contact, participation and the costs of any other individual sample unit. We, thus, ignore any effect that clustering of sample units may have on response or costs. These simplifications imply that there may be more variability in response rates and costs over realizations of the survey than is accounted for. Second, we assume that



there is a stable workload, i.e. we ignore any impact of scale, i.e. sample size, on data collection.

However, we do allow for associations between contact propensities over phases, between participation propensities over phases, and between cost functions over phases.

We introduce some further notation. We let  $\kappa_{t,i}(s_{1,t})$  be the propensity of a contact in phase  $t$  of subject  $i$  under strategy  $s_{1,t}$  given that the unit did not respond in earlier phases and is eligible for follow-up. We let  $\lambda_{t,i}(s_{1,t})$  be the propensity of a participation in phase  $t$  of subject  $i$  under strategy  $s_{1,t}$  given contact. The response propensity in phase  $t$  of a subject  $i$  under strategy  $s_{1,t}$ , is denoted by  $\rho_{t,i}(s_{1,t})$ . We have that  $\rho_{t,i}(s_{1,t}) = \kappa_{t,i}(s_{1,t})\lambda_{t,i}(s_{1,t})$ . When in subsequent phases all nonresponse receives a follow-up, then

$$\rho_i(s_{1,T}) = \kappa_{1,i}(s_1)\lambda_{1,i}(s_1) + \sum_{t=2}^T \left( \prod_{l=1}^{t-1} (1 - \kappa_{l,i}(s_{1,l})\lambda_{l,i}(s_{1,l})) \right) \kappa_{t,i}(s_{1,t})\lambda_{t,i}(s_{1,t}). \quad (1)$$

When in subsequent phases only noncontacts receive a follow-up, i.e. attempting to convert a refuser to participate is not allowed, then

$$\rho_i(s_{1,T}) = \kappa_{1,i}(s_1)\lambda_{1,i}(s_1) + \sum_{t=2}^T \left( \prod_{l=1}^{t-1} (1 - \kappa_{l,i}(s_{1,l})) \right) \kappa_{t,i}(s_{1,t})\lambda_{t,i}(s_{1,t}) \quad (2)$$

In general, the costs per sample depend on the phase, the sample unit and the strategy. For notational convenience here we drop the subscript  $t$  for phase. We define:

- $C_{0,i}(s_{1,t})$  as the cost to make a contact attempt (visit or call) with a sample unit  $i$  in phase  $t$ , following strategy  $s_{1,t} \in \mathcal{S}_{1,t}$ ;
- $C_{R,i}(s_{1,t})$  as the cost for the response of a sample unit  $i$  in phase  $t$ , following strategy  $s_{1,t} \in \mathcal{S}_{1,t}$ , given contact in phase  $t$ ;
- $C_{NR,i}(s_{1,t})$  as the cost for a nonresponse of a sample unit  $i$  in phase  $t$ , following strategy  $s_{1,t} \in \mathcal{S}_{1,t}$ , given contact in phase  $t$ .

For some actions, these functions may be identical to zero, e.g. a response or nonresponse to a web survey. In this paper, we often make the simplification that cost functions do not depend on the phase and history of actions but only on the current action, i.e.

$$C_{0,t,i}(s_{1,t}) = C_{0,i}(s_t), \quad C_{R,t,i}(s_{1,t}) = C_{R,i}(s_t), \quad C_{NR,t,i}(s_{1,t}) = C_{NR,i}(s_t). \quad (3)$$

The cost parameters  $C_i(s_{1,T})$  can be written using these components and the contact and participation propensities. Under (3), when all nonresponse receives follow-up, we get

$$\begin{aligned} C_i(s_{1,T}) = & C_{0,i}(s_1) + \kappa_{1,i}(s_1)(1 - \lambda_{1,i}(s_1))C_{NR,i}(s_1) + \kappa_{1,i}(s_1)\lambda_{1,i}(s_1) C_{R,i}(s_1) + \\ & + \sum_{t=2}^T \left( \left( \prod_{l=1}^{t-1} (1 - \kappa_{l,i}(s_{1,l})\lambda_{l,i}(s_{1,l})) \right) \left( C_{0,i}(s_{1,t}) + \kappa_{t,i}(s_{1,t}) (1 - \right. \right. \\ & \left. \left. \lambda_{t,i}(s_{1,t})) C_{NR,i}(s_{1,t}) + \kappa_{t,i}(s_{1,t}) \lambda_{t,i}(s_{1,t}) C_{R,i}(s_{1,t}) \right) \right) \end{aligned} \quad (4)$$

For self-administered modes, like Web or mail, nonresponse costs may be negligible and we adopt the convention that participation propensities are one,  $\lambda_{t,i}(s) = 1$ , so that

$$C_i(s_{1,T}) = C_{0,i}(s_1) + \rho_{1,i}(s_1) C_{R,i}(s_1) + \sum_{t=2}^T \left( \left( \prod_{l=1}^{t-1} (1 - \rho_{l,i}(s_{1,l})) \right) \left( C_{0,i}(s_{1,t}) + \rho_{t,i}(s_{1,t}) C_{R,i}(s_{1,t}) \right) \right) \quad (5a)$$

We can rewrite (1) as follows:

$$\rho_i(s_{1,T}) = \rho_{1,i}(s_1) + \sum_{t=2}^T \left( \left( \prod_{l=1}^{t-1} (1 - \rho_{l,i}(s_{1,l})) \right) \rho_{t,i}(s_{1,t}) \right). \quad (5b)$$

### 3.2 Models for survey design parameter components

General models for  $\kappa_{t,i}(s_{1,t})$  and  $\lambda_{t,i}(s_{1,t})$  accounting for all possible associations with the full set of actions  $s_{1,t}$ , the auxiliary vector  $x_i$  and the phase  $t$  would be very complicated and cumbersome to write down and may lead to confusion rather than to clarity in the context of this paper. We refer to Durrant et al (2011, 2013, 2015) who describe multi-level models and hazard rate models for such general settings. We, therefore, make a number of simplifications to present the model specification. The auxiliary vector  $x_i$  is treated as fixed and given, i.e. we do not model the underlying distribution. We assume that propensities at phase  $t$  do not depend on the future actions after phase  $t$ , i.e.

$$\begin{aligned} \kappa_{t,i}(s_{1,T}) &= \kappa_{t,i}(s_{1,t}), \\ \lambda_{t,i}(s_{1,T}) &= \lambda_{t,i}(s_{1,t}). \end{aligned}$$

We consider the model for contact propensities, which we model using a probit model, i.e. using a binomial link function. Each sample unit has a contactability represented as a continuous latent variable  $Z_{t,i}^C(s_{1,t})$  and contact is obtained when this latent variable is larger than zero

$$u_{t,i}^C(s_{1,t}) = \begin{cases} 1, & Z_{t,i}^C(s_{1,t}) > 0, \\ 0, & Z_{t,i}^C(s_{1,t}) \leq 0, \end{cases}$$

where  $u_{t,i}^C(s_{1,t})$  is the indicator of contact of subject  $i$  in phase  $t$  following strategy  $s_{1,t}$  and  $Z_{t,i}^C(s_{1,t}) \sim N(\mu(s_{1,t}), \sigma(s_{1,t}))$ , for some  $\mu_{t,i}(s_{1,t}), \sigma_{t,i}(s_{1,t})$  so that

$$\kappa_{t,i}(s_{1,t}) = P(Z_{t,i}^C(s_{1,t}) > 0).$$

For computational convenience, we assume that  $Z_{t,i}^C(s_{1,t})$  has a normal distribution. Recall that unit  $i$  has a corresponding auxiliary vector  $x_i$ . This vector consists of auxiliary variables,  $x_{t,i}$ ,  $t = 1, \dots, T$ . Here  $x_{t,i}$  are variables observed in phase  $t$ . For  $m \leq m_k$  and  $k \leq t$ , let  $\alpha_{t,k,m}^C(s_{1,t})$  be the regression coefficient in phase  $t$  corresponding to the  $m$ -th entry in the auxiliary vector  $x_{k,i}$  given that  $s_{1,t}$  is applied to a unit. Obviously,  $\alpha_{t,k,m}^C(s_{1,t}) = 0$ , when  $k > t$ . Let  $\alpha_{t,k}^C(s_{1,t}) = (\alpha_{t,k,1}^C(s_{1,t}), \dots, \alpha_{t,k,m_k}^C(s_{1,t}))'$  be the coefficients corresponding to  $x_{k,i}$  in phase  $t$ . The model can be written as

$$Z_{t,i}^C(s_{1,t}) = \sum_{k=0}^t \alpha_{t,k}^C(s_{1,t}) x_{k,i} + \varepsilon_{t,i}^C,$$

where  $\varepsilon_{t,i}^C \sim N(0,1)$  is an error term for the uncertainty of contact of the subject. Again, we choose a normal distribution for computational convenience. In case of a skewed distribution, the log(costs) may be modelled. Other distributions than normal may obviously be chosen as well and will require modifications in later sections. Now, the number of coefficients in all contact propensity models is  $\sum_{k=0}^T (T-k)m_k$  for one specific strategy  $s_{1,T}$ . The total number of coefficients depends on the sizes  $|\mathcal{S}_t|$  of the action sets. It is clear that this number can become very large; too large to be feasible in estimation. Hence, in practice, usually, models are simplified by lowering the number of coefficients.

We evaluate a model that has all important features for adaptive survey designs, but that is as simple as possible. First, to be able to include dynamic adaptive survey designs, we need to include paradata. To keep the model simple, we assume, however, there is just one phase, say  $t_1$ , in which paradata is collected. Up to phase  $t_1$  only the auxiliary variables in  $x_{0,i}$  can be used to model the propensities. After phase  $t_1$ , the auxiliary variables obtained in phase  $t_1$  can also be included in the model. Second, we account for dependence of success in a certain phase on past actions, which can be included by introducing a fixed or random effect per possible history. We add the history as a random effect here. Since we add a dependence on the history of actions, the regression coefficients become necessarily dependent on the phase. The model becomes

$$Z_{t,i}^C(s_{1,t}) = \begin{cases} \alpha_{t,0}^C(s_t) x_{0,i} + \varepsilon_{t,i}^C + \delta_t^C(s_{1,t}), & t \leq t_1, \\ \alpha_{t,0}^C(s_t) x_{0,i} + \alpha_{t,1}^C(s_t) x_{t_1,i} + \varepsilon_{t,i}^C + \delta_t^C(s_{1,t}), & t > t_1, \end{cases} \quad (6)$$

where  $\alpha_{t,1}^C$  are the coefficients of the paradata obtained in phase  $t_1$  and  $\delta_t^C(s_{1,t})$  is a random effect.

The model for the participation propensity can be derived analogously and is not given. The  $\alpha_{t,k}^C$ ,  $\varepsilon_{t,i}^C$  and  $\delta_t^C(s_{1,t})$  are replaced by  $\alpha_{t,k}^P$ ,  $\varepsilon_{t,i}^P$  and  $\delta_t^P(s_{1,t})$ .

We make the simplification that cost functions do not depend on the phase and design features in previous phases, but only on the current phase and design features. We can make this simplification since we can assume that these dependences are negligible. For example we assume that breaking off in CAWI and the traveling costs of an interviewer in CAPI are independent. Costs are continuous variables and we use linear models for the costs functions

$$C_{0,i}(s) = \gamma_0(s) x_i + \varepsilon_{0,i}(s), \quad s \in \mathcal{S} \quad (7a)$$

$$C_{R,i}(s) = \gamma_R(s) x_i + \varepsilon_{R,i}(s), \quad s \in \mathcal{S} \quad (7b)$$

$$C_{NR,i}(s) = \gamma_{NR}(s) x_i + \varepsilon_{NR,i}(s), \quad s \in \mathcal{S} \quad (7c)$$

where  $\gamma_0(s)$ ,  $\gamma_R(s)$  and  $\gamma_{NR}(s)$  are regression parameters allowing for interaction between the current action  $s$  and the auxiliary vector  $x_i$  and  $\varepsilon_{0,i}(s)$ ,  $\varepsilon_{R,i}(s)$  and  $\varepsilon_{NR,i}(s)$  are error terms that again allow for an interaction with the current action. The error terms are modelled as independent normal

$$\varepsilon_{0,i}(s) \sim N(0, \sigma_0^2(s)), \quad (8a)$$

$$\varepsilon_{R,i}(s) \sim N(0, \sigma_R^2(s)), \quad (8b)$$

$$\varepsilon_{NR,i}(s) \sim N(0, \sigma_{NR}^2(s)). \quad (8c)$$

## 4. A Bayesian analysis of survey design parameters

We make the analysis Bayesian by assigning prior distributions to the regression coefficients and random effects of section 3.2. Our aim is the derivation of the posterior distributions of the individual response propensities  $\rho_i(s_{1,T})$  and the individual cost parameters  $C_i(s_{1,T})$  per strategy given observed data. These propensities and costs are, in general, complex functions of the underlying survey design parameters per phase. Consequently, their distributions involve multidimensional integration over the distributions of these parameters. We derive expressions for the posterior distributions of the regression coefficients and random effects, but we propose to rely on numerical approximations and Markov Chain Monte Carlo methods to generate draws from the posterior distributions.

### 4.1 Prior distributions

We assign prior distributions to the model parameters in 6(a-b), 7(a-c) and 8(a-c), i.e. the slope parameters, the random effects and the dispersion parameters of the error terms. We assume that regression slope parameters and dispersion parameters are independent over different data collection phases, but they may be dependent within a phase.

For the regression slope parameters and random effects in contact and participation models, we choose normal prior distributions. Despite being based on normal distributions themselves, probit models do not allow for conjugate pairs of prior and posterior distributions for the regression parameters, e.g. Albert and Chib (1993). The normal distributions are an obvious choice, see Gelman et al (2014), but may also be replaced by other distributions. The contact and participation regression slope parameters are modelled as

$$\alpha_t^C(s) \sim N\left(\mu_t^C(s), \Sigma_t^C(s)\right) \quad (9a)$$

$$\alpha_t^P(s) \sim N(\mu_t^P(s), \Sigma_t^P(s)). \quad (9b)$$

The contact and participation random effects are modelled as

$$\delta_t^C \sim N(0, (\tau_t^C)^2), \quad (10a)$$

$$\delta_t^P \sim N(0, (\tau_t^P)^2), \quad (10b)$$

where  $(\tau_t^C)^2$  and  $(\tau_t^P)^2$  are specified covariance matrices.

The models for the cost functions are linear. Here, conjugate prior-posterior pairs are possible. We choose normal distributions for the regression slope parameters and inverse Gamma for the regression dispersion parameters. Inverse Gamma distributions are suggested for random effect variance parameters, see Gelman (2006), as they lead to conditionally conjugate prior-posterior pairs. The cost regression slope parameters are modelled as

$$\gamma_0(s) \sim N(\mu_0(s), \Sigma_0(s)), \quad (11a)$$

$$\gamma_R(s) \sim N(\mu_R(s), \Sigma_R(s)), \quad (11b)$$

$$\gamma_{NR}(s) \sim N(\mu_{NR}(s), \Sigma_{NR}(s)), \quad (11c)$$

and the cost error term variances are modelled as

$$\sigma_0^2(s) \sim \Gamma^{-1}(a_0(s), b_0(s)), \quad (12a)$$

$$\sigma_R^2(s) \sim \Gamma^{-1}(a_R(s), b_R(s)), \quad (12b)$$

$$\sigma_{NR}^2(s) \sim \Gamma^{-1}(a_{NR}(s), b_{NR}(s)). \quad (12c)$$

The contact, participation and costs models are hierarchical Bayes, because different individuals share parameters and because random effects spread out over phases  $t$  and actions  $s_t$ . The normal and inverse Gamma probability distributions for the regression parameters and random effects are then called hyperpriors. The hyperparameters are  $\mu_t^C, \Sigma_t^C, (\tau_t^C)^2, \mu_t^P, \Sigma_t^P, (\tau_t^P)^2, \mu_0, \Sigma_0, \mu_R, \Sigma_R, \mu_{NR}, \Sigma_{NR}, a_0, b_0, a_R, b_R, a_{NR}, b_{NR}$ ; they need to be elicited from historic survey data and/or expert knowledge.

## 4.2 Posterior distributions

The aim is to derive the posterior distributions of response propensities and cost parameters given the observed data. In standard data collection, the observed data consist of

- The response outcome per phase per sample unit, denoted as  $u_{t,i}$ ;
- The realized costs per phase per sample unit, denoted as  $c_{0,t,i}, c_{R,t,i}$  and  $c_{NR,t,i}$ . Per phase  $c_{R,t,i}$  or  $c_{NR,t,i}$  is observed only when contact is made. Contact costs are always observed in every phase. Since we do not model variation in costs for contact, response and refusal when the same action is applied in multiple phases, we need to average realized costs over all phases that employed the same actions. For example, if we make ten telephone calls within a certain time window, then we take the average duration of the ten calls times the interviewer hourly pay rate;
- The complete auxiliary vector  $x_i$ ;
- The survey variable vector  $y_i$  for respondents (not used here)

Additionally, data collection may apply randomization of data collection strategies in order to learn about multiple strategies simultaneously. There is a vast literature on efficient randomization in adaptive or dynamic treatment regimes, e.g. Murphy 2003,

Chakraborty and Murphy 2014 and Laber et al (2014). In general, designs are called SMART (sequential multiple assignment randomized trial), e.g. Lei et al (2012), when the randomization is independent of future outcomes and, hence, allows for disentangling the outcomes for different strategies. Here, we assume that the observed data may contain randomization over strategies but that randomization is only at the outset. Hence, strategy allocation probabilities may depend on auxiliary information known at the start of data collection, but not on paradata coming in during data collection. So in addition to the outcomes, costs and auxiliary vectors, we observe

- The series of actions  $s_{1,T}^i$ , or simply strategy, that were applied per sample unit. Since we assume that future actions do not impact the outcomes or costs of earlier phases, we do not need to include the planned actions for phases not actually conducted. The latter occurs when there is a response or a nonresponse not eligible for follow-up;

In the following, we use  $\rho(s_{1,T})$  and  $C(s_{1,T})$  as shorthand for the vector of response propensities and cost parameters over all sample units for a particular strategy. In the same fashion, we use  $u_t$ ,  $c_0$ ,  $c_R$ ,  $c_{NR}$  and  $x$  to denote the vectors of outcomes, realized costs components and auxiliary variables over sample units. Note that  $x$  may in fact be a matrix, when the auxiliary variables are a vector by themselves. With  $\{s_{1,T,i}\}$  we denote the vector of used strategies for all sample units. To shorten expressions, we use  $\alpha, \delta, \gamma, \sigma^2$  for the vectors of regression slope parameters, random effects and regression dispersion parameters over phases and actions, but elaborate when needed. For convenience, we use  $p$  to express joint and marginal distribution functions; we omit the reference to the random variables to which they apply and ignore differences between discrete and continuous probability distributions. Finally, in the density functions, we omit the dependence on the hyperparameters.

The joint posterior distribution of interest is:

$$p(\rho(s_{1,T}), C(s_{1,T}) | u_t, c_0, c_R, c_{NR}, x, \{s_{1,T}^i\}). \quad (13)$$

This joint distribution follows from integration over all possible combinations of regression parameters  $\alpha, \beta, \delta, \gamma, \sigma^2$  and cannot be written in closed form. A straightforward solution is to perform a Gibbs sampler to the joint density of the regression parameters  $\alpha, \beta, \delta, \gamma, \sigma^2$

$$p(\alpha, \delta, \gamma, \sigma^2 | u_t, c_0, c_R, c_{NR}, x, \{s_{1,T}^i\}). \quad (14)$$

An approximation to the joint density in (13) comes as an important by-product of a Gibbs sampler applied to (14); per draw the response propensities and cost parameters can be computed using (1) or (2), (4) or (5a), (6) and (7a-c).

A Gibbs sampler for (14) requires repeated draws from the conditional densities of each regression parameter given the observed data and the other regression parameters, the so-called full conditionals. Appendix A contains expressions for the

full conditionals per regression parameter. Except for the cost model parameters, we cannot derive a closed form expression for these full conditionals, because of the non-linear probit link functions in (6). The likelihood functions are relatively straightforward to derive.

In this paper, we include draws of the latent variables for contact and participation in the Gibbs sampler.

The Gibbs sampler has the following steps:

1. Set the random effects for the contact and participation equations to zero,  $\delta_t^C = 0$  and  $\delta_t^P = 0$ , and fit regression models to all contact, participation and cost equations and use the resulting estimated parameter values as starting values for the regression parameters  $\alpha^C, \alpha^P, \gamma$ ;
  2. For each unit ( $i$ ) in each phase ( $t$ ), sample the latent variables  $Z_{t,i}^C$  and  $Z_{t,i}^P$  from  $p(Z_{t,i}^C | \alpha^C, \delta_t^C, u_{t,i}, x_i, \{s_{1,T}^i\})$  and  $p(Z_{t,i}^P | \alpha^P, \delta_t^P, u_{t,i}, x_i, \{s_{1,T}^i\})$ ;
  3. For each phase, sample the contact slope parameters  $\alpha_t^C$  from  $p(\alpha_t^C | Z_t^C, \delta_t^C, x_i, \{s_{1,T}^i\})$ ;
  4. Sample the random effects  $\delta_t^C$  from  $p(\delta_t^C | Z_t^C, \alpha_t^C, x_i, \{s_{1,T}^i\})$ ;
  5. For each phase, sample the participation slope parameters  $\alpha_t^P$  from  $p(\alpha_t^P | Z_t^P, \delta_t^P, x_i, \{s_{1,T}^i\})$ ;
  6. Sample the random effects  $\delta_t^P$  from  $p(\delta_t^P | Z_t^P, \alpha_t^P, x_i, \{s_{1,T}^i\})$ ;
  7. For the three cost components, sample the variance parameters  $\sigma^2$  from  $p(\sigma^2 | \gamma, c_0, c_R, c_{NR}, x, \{s_{1,T}^i\})$ ;
  8. For the three cost components sample the slope parameters  $\gamma$  from  $p(\gamma | \sigma^2, c_0, c_R, c_{NR}, x, \{s_{1,T}^i\})$ ;
  9. Return to step 2 for another round of defining posteriors based on a new prior ;
- In order to carry out the data augmentation, we did not make use of standard libraries in R (e.g. mcmc or gibbs.met) or SAS (e.g. PROC MCMC), but programmed the Gibbs sampler from scratch. Code is available upon request.

### 4.3 A Bayesian analysis of functions of survey design parameters

For monitoring and optimization of data collection, the focus is on functions of the design parameters that correspond to overall quality or cost objectives. We consider three such functions here for brevity: the response rate, the total costs and the coefficient of variation of the response propensities; the analysis of other functions can often be done in an analogous way. See Nishimura, Wagner and Elliott (2016) for a discussion of indicators.

Let  $d_i$  represent the design or inclusion weight for sample unit  $i, i = 1, 2, \dots, n$ . The response rate, RR, for strategy  $s_{1,T}$  can be written as

$$RR(s_{1,T}) = \frac{1}{\sum_{i=1}^n d_i} \sum_{i=1}^n d_i \rho_i(s_{1,T}), \quad (15)$$

the coefficient of variation, CV, is

$$CV(X, s_{1,T}) = \frac{\sqrt{\frac{1}{\sum_{i=1}^n d_i} \sum_{i=1}^n d_i (\rho_i(s_{1,T}) - RR(s_{1,T}))^2}}{RR(s_{1,T})} \quad (16)$$

where  $\sum_{i=1}^n d_i = N$  for many customary sampling designs and the total costs, or required budget,  $B$ , associated with  $s_1^T$  are

$$B(s_{1,T}) = \sum_{i=1}^n c_i(s_{1,T}). \quad (17)$$

For the *CV* (Schouten, Cobben and Bethlehem 2009 and De Heij, Schouten and Shlomo 2015), we explicitly denote the dependence on the covariate vector  $X$ ; for any other choice of auxiliary variables it will, generally, attain a different value. The response rate and total costs do not depend on the choice of  $X$ .

Obviously, the prior and posterior distributions for these three functions are determined by the prior and posterior distributions of the components of the response propensities and cost functions. The posteriors have even more complex forms than the posteriors for individual response propensities and cost parameters. However, they can again be approximated as a by-product of the Gibbs sampler in section 4.2. For every draw of the individual response propensities and cost parameters, we compute (15) to (17).

## 5. A simulation study to investigate the utility of Bayesian analysis

In the simulation study, we investigate the impact of prior distribution specification and of survey sample size on the location and variance of posterior distributions. Specifically, we look at the added value of the prior information. First, we present the specifics of the simulation study and discuss how we attempt to prove the efficacy of a Bayesian analysis. Next, we briefly discuss convergence properties of the Gibbs sampler. Finally, we show results of the simulation study and discuss the conditions under which the analysis is useful.

### 5.1 Design of the simulation study

To evaluate the utility of a Bayesian analysis, we compare posterior distributions of response rates, coefficients of variation of response propensities and total costs starting from different prior distributions for the survey design parameters, more specifically for the regression slope and dispersion parameters in contact, participation and cost models per data collection phase. The prior distributions that we compare to are fully non-informative priors, which have (arbitrary) large variances and expectations that are the same for all population subgroups. These priors conform to lack of knowledge at the start of data collection and we view this choice as a “non-Bayesian” analysis, despite the use of prior and posterior distributions. Thus, we still make use of the benefit of a Bayesian analysis in that it allows for an



easy display of uncertainty during and after data collection. We make two comparisons that both start from “true” priors. The true priors have expectations that exactly match the simulation model and have variances that correspond to the standard errors for a historic dataset of sample size 10000, i.e. as if we have already observed a fairly large and unbiased realization of the survey. In the first comparison, we gradually misspecify expectations of the true priors in order to mimic bias due to time change and/or a change of survey design. However, the variances of the priors remain the same. In the second comparison, we gradually increase variances, but keep expectations constant, in order to mimic imprecision. The two comparisons allow us to see how much gain comes from the prior knowledge.

We quantify this gain by the root mean square error (RMSE) of the posterior distribution relative to the simulation model values. Let  $p_{\pi}(\Theta|u_t, c_0, c_R, c_{NR}, x, \{s_{1,T}^i\})$  be a posterior for a data collection quality or cost indicator  $\Theta$  of interest, e.g. the response rate,  $CV$  or total costs, using prior  $\pi$ . The RMSE for this indicator and prior is then defined as

$$\text{RMSE}(\Theta; \pi) = \sqrt{(E_{p_{\pi}}(\Theta) - \Theta_0)^2 + \text{var}_{p_{\pi}}(\Theta)}, \quad (18)$$

where  $\Theta_0$  is the simulation model value.

We base our simulation study on the 2015 Dutch Health Survey (HS), that we described in the previous sections. Two auxiliary variables, gender and age, are linked from administrative data, and one variable, web break-off, is added from phase 1 paradata. Gender and age are crossed to form six strata, {0-29 years, 30-59 years, 60 years and older}×{female, male}. Web break-off is a binary indicator for a broken-off Web response; it is not crossed with the gender-age variable but added as a main effect. We refer to the variables as GenderAge and BreakOff. From 2015 HS data, contact propensities, participation propensities and costs per sample unit are derived for the three phases and used to simulate analysis data sets of sample size 1250, 2500, 5000 and 10000. The simulation probabilities and costs are given in appendix C. To model contact and participation, we use a probit regression with GenderAge in phase 1 and GenderAge + BreakOff in phases 2 and 3. For phase 1, online data collection, we set participation propensities equal to response and participation costs are set to zero. We do this, because for online surveys costs are only associated with contact and not with interview. For phases 2 and 3, F2F short and extended, we do distinguish contact and participation propensities. To model costs, we use a linear regression with GenderAge in all phases. Table 1 gives simulation response rates, coefficients of variation and total costs cumulatively for all phases based on the true simulation model values in the top row of each section.

**Table 1: Expected response rates (RR), coefficients of variation (CV) and total costs (B) cumulatively based on the 2015 HS simulation model, and based on the three misspecified priors.**

	Data	Web	F2F short	F2F extended
RR	True	30.2%	57.6%	60.5%
	Misspecified light	32.2%	57.2%	59.7%
	Misspecified medium	35.2%	56.8%	58.8%
	Misspecified strong	40.2%	56.8%	58.2%
CV	True	0.277	0.069	0.102

	Misspecified light	0.260	0.061	0.094
	Misspecified medium	0.238	0.049	0.082
	Misspecified strong	0.208	0.036	0.063
B	True	3.0	15.2	19.4
	Misspecified light	3.0	14.5	19.5
	Misspecified medium	3.0	13.6	19.8
	Misspecified strong	3.0	12.1	20.3

We chose prior distributions as specified in section 4.1 and applied the Gibbs sampler of section 4.2. In order to set true prior hyperparameters for the main effect models GenderAge + BreakOff, we simulated 2000 data sets of sample size 10000 according to the simulation model, fitted main effects models to each and took the mean and variances of the 2000 fits. We derived the true prior hyperparameters for the cost model parameters by approximating outcomes of historic data by normal and chi-square distributions. We refer to the supplemental file for details. Misspecification was introduced by shifting contact and participation propensities for each subgroup in the same direction. For the online phase 1 they were increased by 2%, 5% and 10%. For the F2F phases, they were decreased by 2%, 5% and 10%. Hence, we mimic an overestimation of online response and an underestimation of subsequent F2F response, which essentially leads to an underestimation of required budget. Table 1 also contains the expected response rates, coefficients of variation and costs based on the three sets of misspecified priors.

## 5.2 Convergence properties

The Gibbs sampler produces a draw of a Markov chain that has the posterior distribution of interest as its stationary distribution. The Markov chain is initiated from one or more starting values. Under certain conditions, the Markov chain converges to its stationary distribution but at a certain speed. For this reason, usually a burn-in period is discarded where the Markov chain has not yet reached its stationary distribution. After the burn-in period, the Markov chain moves through its parameter space at a certain “speed”. This speed is termed the mixing property of the chain and determines the required length of the Markov chain, i.e. the number of iterations in the Gibbs sampler. Both the burn-period and the mixing of the Gibbs sampler cannot be determined with certainty, since the stationary distribution is unknown. However, various diagnostics have been developed to make an empirical assessment. We checked the burn-in period and convergence of the Gibbs sampler using the Raftery Lewis convergence diagnostic (Raftery and Lewis 1992) as implemented in the R package CODA. We required that the 2.5%-quantile of the posterior distribution could be approximated with a specified precision. We apply the Raftery and Lewis convergence diagnostic on the first 5000 iterations to determine the number of iterations that are needed for convergence.

CODA

Figure 1: Gibbs sampler runs for the slope parameters in the phase 1 contact model for the non-informative prior.

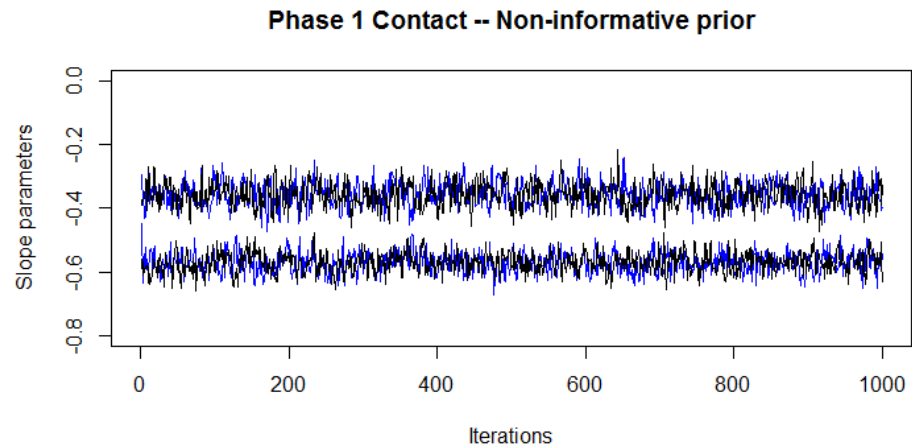
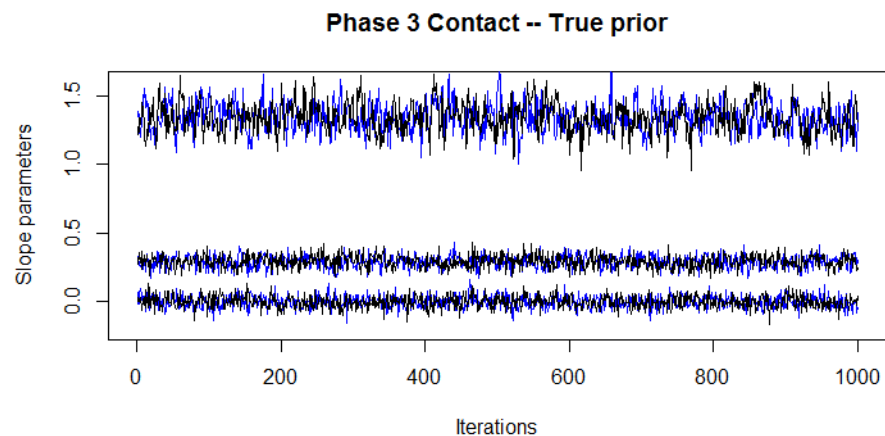


Figure 2: Gibbs sampler runs for the slope parameters in the phase 3 contact model for the true prior.



Somewhat surprisingly, we found for all scenarios, that the burn-in period is very short and below 100 iterations. Nonetheless, we always discard the first 5000 iterations. We then carried out the second part of iterations. The starting values here are the parameters obtained after 5000 iterations in the first run. The number of iterations in the second run is derived from the Raftery and Lewis convergence diagnostic. Convergence was usually reached within 4500 iterations, however we had some cases where up to 10000 iterations were required. Computation times in R were roughly 5000 iterations in 20 minutes. Figures 1, 2 and 3 show Gibbs sampler runs for regression slope parameters in the phase 1 and 3 for contact model and phase 2 for participation model. The plots in blue are the first 1000 runs from the originally discarded 5000 runs and in black the first 1000 runs from the second runs. Figure 4 shows the regression parameters in the phase 2 for contact model for the second run. Here 10000 iterations were required by the Raftery and Lewis convergence diagnostic.

Figure 3: Gibbs sampler runs for the slope parameters in the phase 2 participation model for the misspecified prior.

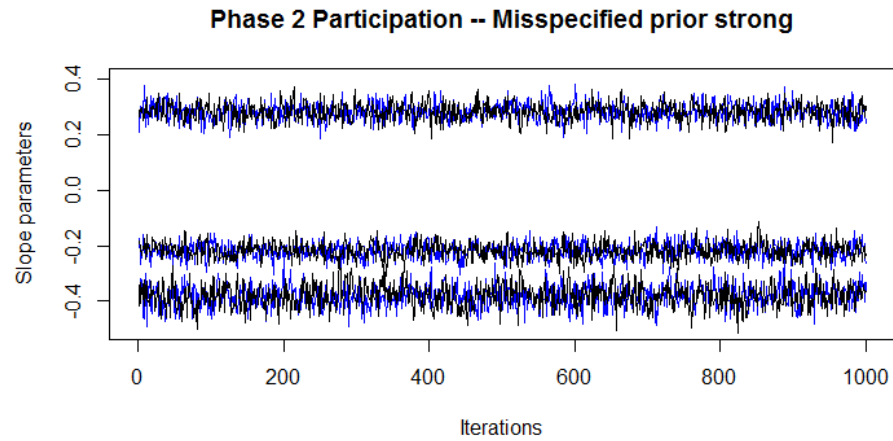
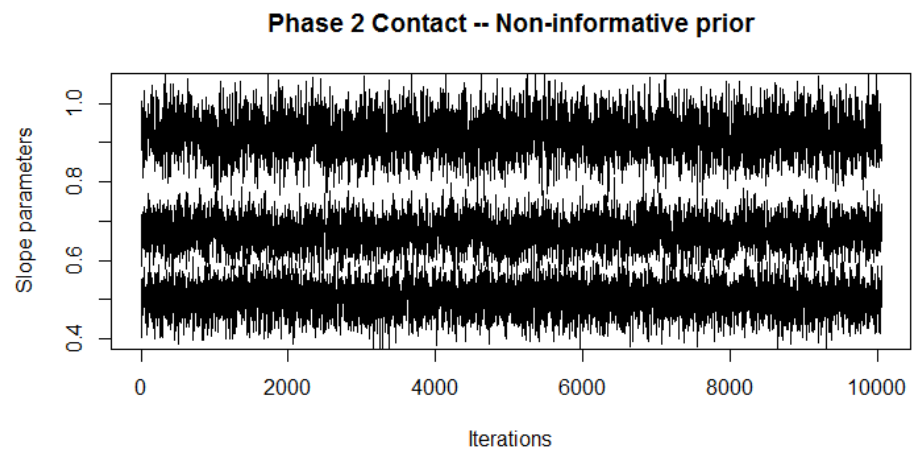


Figure 4: Gibbs sampler runs for the slope parameters in the phase 2 contact model for the Non-informative prior.



## 5.3 Simulation results

In the following two subsections, we discuss the two comparisons to evaluate the utility of the Bayesian analysis: increasing the variances of prior distributions and shifting their expectations.

### 5.3.1 Variance of the prior distributions

In the first evaluation, we focus on the variance term of the RMSE of the posterior distributions and vary the sample size of the observed data. The true prior is compared with the fully non-informative prior. We view the non-informative prior as a non-Bayesian analysis benchmark.

Table 2 shows the RMSE values for the two priors for four sample sizes: 1250, 2500, 5000 and 10000 units. Three variance levels for the misspecified priors are chosen, corresponding to a historic data set of a modest size of 1250 units (V1), a moderate size of 2500 units (V2) and a large size of 10000 units (V3). We note that the RMSE depends on the scale of the population parameters of interest; RMSE values for costs are, therefore, larger.

The RMSE values under the true priors are always lower than for the non-informative prior, as expected. The gap gets larger when the sample size decreases and/or the true prior variance decreases. However, for a sample size of 10000, the added value of prior information is already quite small. For even larger sample sizes, it will not make much difference whether the prior knowledge is added or not.

The most advantageous setting is where the both prior variance and the observed data sample size are smallest. The biggest gap in RMSE is indeed found for a prior with variance V3 and sample size 1250. The RMSE values of this combination are comparable to that of the non-informative prior with sample size 10000.

**Table 2: RMSE for fully non-informative and true priors for response rates (RR), coefficients of variation (CV) and costs (B) cumulatively after each phase and for a dataset of sample sizes 1250, 2500, 5000 and 10000. The true priors have a variance corresponding to 1250 (V1), 2500 (V2) and 10000 (V3) historic sample units.**

Size	Prior	RR			CV			B		
		Web	F2F	F2FE	Web	F2F	F2FE	Web	F2F	F2FE
1250	Non-informative	0.014	0.019	0.015	0.046	0.045	0.037	0.010	0.316	0.374
	True V1	0.010	0.012	0.010	0.021	0.023	0.018	0.010	0.218	0.273
	True V2	0.008	0.009	0.008	0.014	0.015	0.012	0.010	0.178	0.223
	True V3	0.004	0.005	0.005	0.007	0.008	0.008	0.010	0.116	0.142
2500	Non-informative	0.010	0.010	0.010	0.012	0.055	0.041	0.009	0.239	0.298
	True V1	0.008	0.008	0.008	0.010	0.035	0.025	0.009	0.204	0.247
	True V2	0.007	0.007	0.007	0.008	0.025	0.017	0.009	0.181	0.217
	True V3	0.004	0.004	0.004	0.006	0.009	0.007	0.009	0.128	0.148
5000	Non-informative	0.006	0.007	0.007	0.009	0.023	0.016	0.009	0.156	0.183
	True V1	0.006	0.006	0.006	0.008	0.018	0.013	0.008	0.143	0.166
	True V2	0.005	0.006	0.006	0.007	0.015	0.010	0.008	0.134	0.157
	True V3	0.004	0.004	0.004	0.005	0.007	0.006	0.009	0.105	0.121
10000	Non-informative	0.005	0.005	0.005	0.007	0.008	0.009	0.010	0.120	0.135
	True V1	0.004	0.005	0.005	0.006	0.008	0.009	0.009	0.114	0.130
	True V2	0.004	0.005	0.005	0.006	0.008	0.010	0.009	0.111	0.125
	True V3	0.003	0.004	0.004	0.005	0.008	0.009	0.010	0.110	0.108

In the analysis, we consider the population as a whole. However, once subpopulations are of interest and statistics are detailed to such subpopulations, then, obviously, sample sizes get smaller and the prior distributions will still have added value. Table 2 should then be evaluated as the sample sizes of such subpopulations.

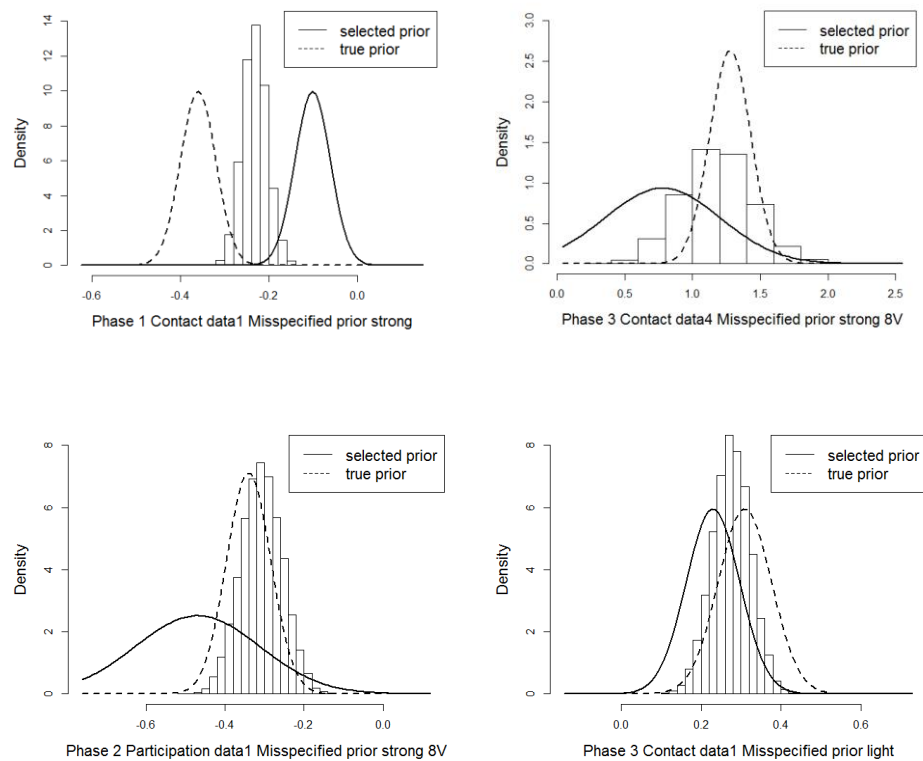
The results of the second evaluation suggest that a Bayesian analysis is advantageous for small to modest size samples of (sub)populations as the historic survey data and expert knowledge lower the variances of the posterior distributions.

### 5.3.2 Misspecification of the prior distributions

In the second evaluation, we gradually misspecify the prior distributions for the contact and participation regression slope parameters, and compare the RMSE to a fully non-informative prior. We view the non-informative prior again as a non-Bayesian analysis benchmark.

To give an impression, figure 5 shows empirical posterior distributions produced by the Gibbs sampler for a selection of the regression slope parameters over the three phases. Next to the empirical posterior densities, also the densities of the selected priors are shown as well as the density of the true prior assuming a historic dataset of 10000 sample units. The empirical posterior densities clearly have different modi and variances, depending on the prior specification. However, in most cases, the posterior densities overlap with the true prior and often have very similar support.

**Figure 5: Empirical posterior densities for the slope parameters after all three phases for the Gibbs sampler runs of the misspecified and non-informative priors for different data sets.**



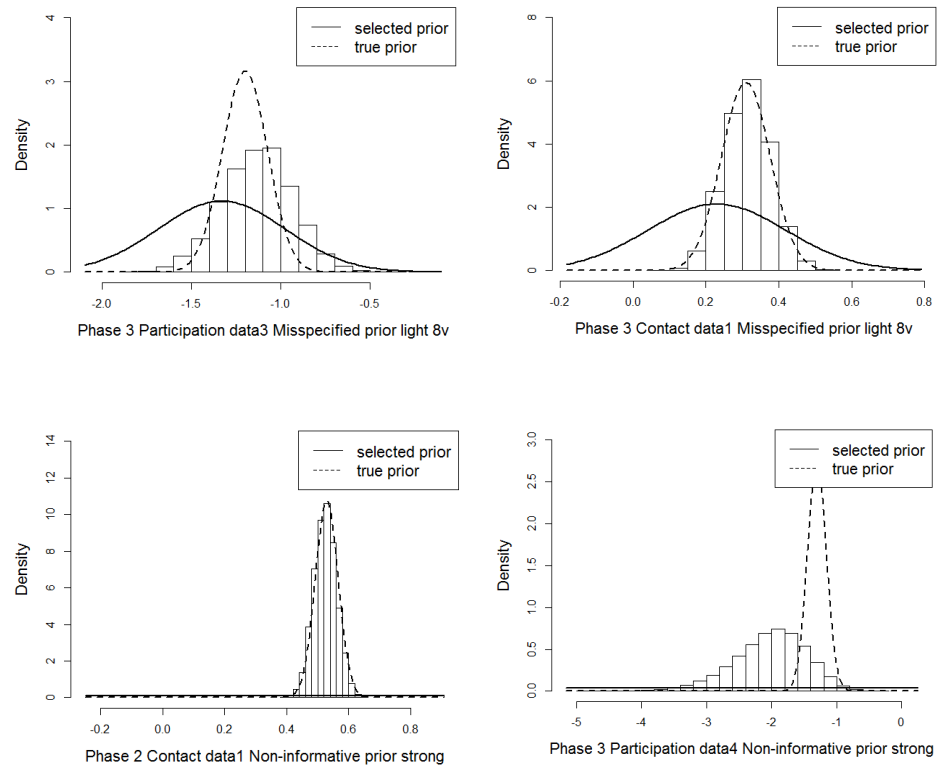


Table 3 contains the RMSE values for non-informative and misspecified priors estimated using the Gibb sampler. Again, we have chosen three variance levels, corresponding to a historic dataset of 1250 (V1), 2500 units (V2) and 10000 units (V3). Furthermore, we evaluate four sample sizes: 1250, 2500, 5000 and 10000. Recall from table 1 that, for phase 1, the misspecification leads to a growing overestimation of the response rate and a growing underestimation of the coefficient of variation, whereas costs are fixed. The cumulative response rates after phases 2 and 3 are affected only little, but the coefficient of variation is underestimated. The cumulative costs after phase two are underestimated, but after phase 3 they are slightly overestimated.

The main observation from the RMSE values is that a misspecified prior can outperform a non-informative prior, but misspecification should be modest and/or the variance of the prior should be relatively large. Furthermore, in close analogy to the results in the previous subsection, it holds that the larger the sample of the observed data, the smaller the misspecification must be to outperform the non-informative prior.

The 2% shift in propensities under misspecified light is small enough for the CV to get RMSE values that are similar or smaller than those for the non-informative prior. This holds also to some extent for the 5% and 10% shifts under misspecified medium and large, when the variance of the prior is large.

For the response rate and costs, RMSE values are almost always larger for the misspecified priors, unless the variance of the prior is relatively large.

Decreasing the sample size of the observed data leads to higher RMSE values, as expected, for all priors. When sample sizes are lowered, in general, the misspecified priors will ultimately perform better than the non-informative prior; misspecified

knowledge beats no knowledge. The (pathological) exception is where the expectation of the non-informative prior happens to be close to the true value, e.g. true contact or participation propensities do not vary between subpopulations and are also close to 50%.

**Table 2 RMSE for fully non-informative and misspecified priors for response rates (RR), coefficients of variation (CV) and costs (B) cumulatively after each phase and for a dataset of sample sizes 1250, 2500, 5000 and 10000. The misspecified priors have a variance corresponding to 1250 (V1), 2500 (V2) and 10000 (V3) historic sample units.**

Size	Prior	RR			CV			B		
		Web	F2F	F2FE	Web	F2F	F2FE	Web	F2F	F2FE
1250	Non-informative	0.014	0.019	0.015	0.046	0.045	0.037	0.010	0.316	0.374
	Missp light V1	0.012	0.014	0.012	0.019	0.020	0.016	0.010	0.398	0.265
	Missp light V2	0.014	0.011	0.011	0.012	0.013	0.012	0.010	0.459	0.284
	Missp light V3	0.018	0.007	0.009	0.006	0.013	0.015	0.010	0.549	0.356
	Missp medium V1	0.023	0.026	0.026	0.016	0.020	0.016	0.010	1.061	0.719
	Missp medium V2	0.032	0.027	0.030	0.010	0.014	0.015	0.010	1.360	0.964
	Missp medium V3	0.044	0.029	0.036	0.006	0.023	0.028	0.010	1.767	1.295
	Missp strong V1	0.046	0.010	0.010	0.013	0.015	0.016	0.010	1.346	0.689
	Missp strong V2	0.063	0.008	0.008	0.008	0.019	0.026	0.010	1.768	0.942
	Missp strong V3	0.087	0.005	0.005	0.008	0.033	0.047	0.010	2.324	1.281
2500	Non-informative	0.010	0.010	0.010	0.012	0.055	0.041	0.009	0.239	0.298
	Missp light V1	0.008	0.009	0.008	0.009	0.032	0.024	0.009	0.213	0.196
	Missp light V2	0.010	0.008	0.008	0.008	0.021	0.015	0.009	0.286	0.203
	Missp light V3	0.030	0.008	0.005	0.006	0.010	0.014	0.009	0.706	0.543
	Missp medium V1	0.014	0.015	0.017	0.009	0.032	0.022	0.009	0.619	0.441
	Missp medium V2	0.022	0.019	0.022	0.008	0.019	0.012	0.009	0.959	0.702
	Missp medium V3	0.053	0.015	0.022	0.005	0.015	0.022	0.009	1.787	1.379
	Missp strong V1	0.029	0.008	0.008	0.009	0.023	0.015	0.009	0.819	0.440
	Missp strong V2	0.045	0.007	0.007	0.009	0.012	0.012	0.009	1.251	0.682
	Missp strong V3	0.077	0.008	0.004	0.009	0.028	0.037	0.009	2.065	1.151
5000	Non-informative	0.006	0.007	0.007	0.009	0.023	0.016	0.009	0.156	0.183
	Missp light V1	0.007	0.006	0.006	0.008	0.017	0.012	0.009	0.164	0.155
	Missp light V2	0.009	0.006	0.006	0.008	0.013	0.009	0.009	0.217	0.168
	Missp light V3	0.014	0.005	0.006	0.006	0.008	0.009	0.009	0.399	0.267
	Missp medium V1	0.011	0.009	0.010	0.009	0.016	0.011	0.009	0.407	0.308
	Missp medium V2	0.017	0.012	0.014	0.009	0.012	0.009	0.009	0.669	0.504
	Missp medium V3	0.033	0.021	0.027	0.008	0.013	0.016	0.009	1.323	0.994
	Missp strong V1	0.020	0.006	0.006	0.009	0.012	0.009	0.009	0.520	0.300
	Missp strong V2	0.032	0.006	0.006	0.010	0.010	0.011	0.009	0.864	0.488
	Missp strong V3	0.065	0.007	0.004	0.010	0.027	0.031	0.009	1.736	0.984
10000	Non-informative	0.005	0.005	0.005	0.007	0.008	0.009	0.010	0.120	0.135
	Missp light V1	0.005	0.005	0.005	0.007	0.009	0.010	0.009	0.106	0.116
	Missp light V2	0.005	0.005	0.005	0.006	0.009	0.011	0.009	0.128	0.116
	Missp light V3	0.010	0.005	0.006	0.006	0.012	0.013	0.009	0.287	0.196
	Missp medium V1	0.006	0.008	0.008	0.007	0.009	0.010	0.009	0.216	0.176
	Missp medium V2	0.010	0.010	0.011	0.007	0.010	0.011	0.010	0.386	0.297



Missp medium V3	0.024	0.019	0.023	0.007	0.016	0.018	0.009	0.988	0.752
Missp strong V1	0.011	0.005	0.005	0.007	0.010	0.013	0.009	0.275	0.169
Missp strong V2	0.019	0.005	0.005	0.007	0.013	0.016	0.009	0.500	0.287
Missp strong V3	0.048	0.004	0.004	0.009	0.026	0.029	0.009	1.288	0.739

The results of this second evaluation suggest turning points for the utility of a Bayesian analysis that depend on the size of the misspecification, the size of the sample and the variance of the prior distributions. This is a complex function that requires further study. However, the results under the current simulation model show that misspecification may be very influential and may quickly reduce the added value of a Bayesian analysis.

## 6. Discussion

We introduced a Bayesian model for survey design parameters related to response and costs. The model is general in that it describes multiple data collection phases, includes both auxiliary variables that are given when data collection starts and auxiliary variables that become available during data collection, acknowledges multiple nonresponse outcomes, accounts for dependence on previous actions and enables the inclusion of randomization over different data collection strategies. Many surveys conducted by statistical institutes can fit into this framework. Furthermore, we constructed an analysis strategy based on a Gibbs sampler in which all model parameters are repeatedly drawn. The Gibbs sampler provides estimates for the posterior distributions of the contact and participation propensities and the costs per sample unit. From the Gibbs sampler, the posterior distributions for overarching quality indicators, such as the response rate or coefficient of variation of the response propensities, and cost indicators can easily be derived as an important by-product. The computation times of the Gibbs sampler are manageable and sufficiently short to run overnight for a range of scenarios. We are, thus, able to meet the first three objectives of the paper to set up a Bayesian analysis for survey data collection monitoring and analysis.

The fourth, and most, important objective is to show the added value of a Bayesian analysis. In the evaluation, we viewed a fully non-informative prior as representing, essentially, a non-Bayesian analysis in which no historic survey data or expert knowledge is incorporated. In order, to be able to compare, we remain in the Bayesian framework of prior and posterior distributions. The evaluation is based on a simulation study using realistic contact propensities and costs, and participation propensities and costs from a multi-mode survey. The evaluation shows that the Bayesian analysis is sensitive to misspecification in the propensities and costs; shifts in propensities and costs should be relatively modest to outperform an analysis with a non-informative prior. The corresponding turning point does depend on the variance of the informative prior, and, consequently, hints at some form of moderation of historic/expert knowledge. The evaluation also shows that without

misspecification the Bayesian analysis is to be favoured to a non-Bayesian analysis, especially, for smaller sample sizes of observed data.

We see three conceptual limitations to our study that deserve future research and extension. First, although our model for monitoring of response and costs has general features, it does not fit all possible data collection designs and analyses, and particular designs and analyses may require adaptations of the model. However, we believe that such changes are relatively straightforward given the exposition in this paper. Second, we have not yet considered the (key) survey variables. Such variables may be modelled and monitored simultaneously, and design decisions may be based on a mix of overall quality and cost indicators and key survey estimates. Such an extension is fairly easy to include, but does introduce new modelling choices because values of survey variables are unknown for nonrespondents. For this reason, we leave this extension to future research. Third, and strongly related to the previous point, we focused on nonresponse and have not yet considered strategy-dependent measurement biases. In multi-mode surveys, such an extension and broader look is inevitable. Extending the Bayesian analysis to measurement error implies modelling survey variables, and, hence, may best be picked up simultaneously with extending the scope to key survey estimates.

The findings of this paper point at a sensitivity of Bayesian analyses to misspecification in prior distributions. Such a sensitivity may be partially overcome by moderating the strength of historic survey data and expert knowledge over time, i.e. the more timely the data and knowledge the more power is attached. Such moderation can be done using so-called power priors (Ibrahim and Chen 2000 and Ibrahim et al 2015). However, moderation may also be achieved by adding a hierarchical level to the Bayesian models representing change in time, which comes at the cost of extra model parameters. In retrospective Bayesian analyses, we are currently investigating the use of moderation in time.

We briefly touched on the elicitation of prior distributions from historic survey data and expert knowledge. In models with many auxiliary variables, such elicitation may be difficult to conduct. Furthermore, data collection experts will, generally, not be able to provide values for slope and dispersion parameters in regression models, but only for propensities and costs at the subgroup level. An effective elicitation of expert knowledge, therefore, likely requires some interpolation or proportional fitting of detailed models to marginal distributions that are given by experts. This trade-off holds, especially, for settings where priors are elicited from different, but similar, surveys. When prior distributions are based on historic data from the same survey, then models may be fitted directly. In order to develop effective prior elicitation procedures, we will apply the Bayesian analysis framework to a broad set of case studies.

Ultimately, the Bayesian analysis framework should support adaptive survey design decisions. Such an application means that historic survey data and expert knowledge should comprise of multiple, possibly randomized, strategies, and that observed data may be used to learn and update strategies for which information is weak or missing.

# Acknowledgements

This paper was produced within the international Bayesian Adaptive survey Design Network (BADEN). BADEN is funded by the Leverhulme Trust International Network Grant (reference number: IN-2014-046) for the period January 2015 to December 2017. The participating institutions are the Universities of Manchester (coordinator), Michigan and Southampton, and RTI, Statistics Sweden, US Census Bureau and Statistics Netherlands.

# References

- Albert, J.H., Chib, S. (1993), Bayesian analysis of binary and polytomous response data, *Journal of the American Statistical Association*, 88, 669 – 680.
- Burger, J., Perryck, K.H., Schouten, B. (2016), Assessing the impact of inaccuracy in design parameters on the performance of adaptive survey designs, under review with *JOS*.
- Calinescu, M. (2013), Optimal Resource Allocation in Adaptive Survey Designs, PhD Thesis, VU University Amsterdam, available at <http://dare.ubvu.nl>.
- Calinescu, M., Schouten, B. (2015). Adaptive survey designs to minimize mode effects. A case study on the Dutch Labour Force Survey, *Survey Methodology*, 41 (2), 403 – 425.
- CBS (2016), Short description of Dutch Health Survey, Statistics Netherlands, Den Haag, The Netherlands, available at <https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/health-survey-from-2010-2013>.
- Chakraborty, B., Murphy, S.A. (2014), Dynamic Treatment Regimes, *Annual Review of Statistics and Its Applications*, 1, 447 – 464.
- Coffey, S., Reist, B., White, M. (2013). Monitoring Methods for Adaptive Design in the National Survey of College Graduates, In *JSM Proceedings, Survey Methods Research Section*, Alexandria, VA: American Statistical Association, 3085-3099.
- Durrant, G.D., D’Arrigo, J., Steele, F. (2011), Using paradata to predict best times of contact, conditioning on household and interviewer influences, *Journal of the Royal Statistical Society A*, 174 (4), 1029 – 1049.
- Durrant, G.D., Maslovskaya, O., Smith, P.W.F. (2015), Modelling final outcome and length of call sequence to improve efficiency in interviewer call scheduling, *Journal of Survey Statistics and Methodology*, 3, 397 – 424.
- Durrant, G.D., Steele, F. (2013), Analysing interviewer call record data by using a multilevel discrete time event history modelling approach, *Journal of the Royal Statistical Society A*, 176 (1), 251 – 269.
- Gelman, A. (2006), Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, 1 (3), 515 – 534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis*. London: Chapman and Hall, second edition.

Groves, R.M., Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs, *Journal of the Royal Statistical Society: Series A*, 169, 439 – 457.

Heyd, J.M., Carlin, B. (1999), Adaptive design improvements in the continual reassessment method for phase I studies, *Statistics in Medicine*, 18, 1307 – 3121.

Ibrahim, J.G., Chen, M.H. (2000), Power prior distributions for regression models, *Statistical Science* 15 (1), 46 – 60.

Ibrahim, J.G., Chen, M.H., Gwon, Y., Chen, F. (2015), The power prior: theory and applications, *Statistics in Medicine*, 34 (28), 3724 – 3749.

Kreuter, F. (2013), *Improving Surveys with Paradata*. Analytic Uses of Process Information, Edited monograph, John Wiley and Sons, Hoboken, New Jersey, USA.

Laber, E.B., Lizotte, D.J., Qian, M., Pelham, W.E., Murphy, S.A. (2014), Dynamic Treatment Regimes: technical challenges and applications, *Electronic Journal of Statistics*, 8 (1), 1125 – 1272.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., Murphy, S.A. (2012), A SMART design for building individualized treatment sequences, *Annual Review of Clinical Psychology*, 8, 21 – 48.

Luiten, A., Schouten, B. (2013). Adaptive fieldwork design to increase representative household survey response: A pilot study in the Survey of Consumer Satisfaction. *Journal of Royal Statistical Society, Series A*, 176 (1), 169 – 190.

Lundquist, P., Särndal, C.E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey, *Journal of Official Statistics*, 29 (4), 557 – 582.

Murphy, S.A. (2003), Optimal dynamic treatment regimes, *Journal of the Royal Statistical Society: Series B*, 65, 331 – 355.

Nishimura, R., Wagner, J., Elliott, M.R. (2016), Alternative indicators for the risk of nonresponse bias. A simulation study, *International Statistical Review*, 84 (1), 43 – 62.

O’Quigley, J., Shen, L.Z. (1996), Continual reassessment method: A likelihood approach, *Biometrics*, 52 (2), 673 – 684.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21 – 29.

Raftery, A.E., Lewis, S. (1992), How many iterations in the Gibbs sampler, In *Bayesian Statistics 4* (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford University Press, Oxford, 763 – 773.

Rosen, J. A., Murphy, J. J., Peytchev, A., Holder, T. E., Dever, J. A., Herget, D. R., Pratt, D. J. (2014), Prioritizing low-propensity sample members in a survey: Implications for nonresponse bias, *Survey Practice*, 7(1), 1–8.

Särndal, C.E., Lundquist, P. (2014). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2, 361 – 387.

Schafer, J.L. (2013). Bayesian penalized spline models for statistical process monitoring of survey paradata quality indicators, Chapter 13, 311 – 340, In *Improving surveys with paradata*. Analytic uses of process information, ed. F. Kreuter, Wiley, New York, USA.

Scharfstein, D.O., Daniels, M.J., Robins, J.M. (2003), Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes, *Biostatistics*, 4 (4), 495 – 512.

Schouten, B., Calinescu, M., Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39 (1), 29 – 58.

Schouten, J.G., Cobben, F., Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35 (1), 101 – 113.

Schouten, B., Shlomo, N. (2015), Selecting adaptive survey design strata with partial R-indicators, *International Statistical Review*, available on early view.

Schulte, P.L., Tsiatis, A.A., Laber, E.B., Davidian, M. (2014), Q- and A-learning methods for estimating optimal dynamic treatment regimes, *Statistical Science*, 29 (4), 640 – 661.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, PhD thesis, University of Michigan, USA.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias, *Public Opinion Quarterly*, 76 (3), 555 – 575.

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G., Kruger Ndiaye, S. (2013). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28 (4), 477 – 499.

## Appendix A: The full conditionals of the regression parameters in the Gibbs sampler

In this appendix, we provide expressions for the full conditionals of the various regression model parameters for response and costs.

### A.1 Full conditionals for regression parameters in response propensity models

Contact propensity and participation propensity models have the following form

$$Z_{t,i}(s_{1,t}) = \alpha_t(s_t)x_{t,i} + \varepsilon_{t,i} + \delta_t(s_{1,t})$$

where  $Z_{t,i}(s_{1,t})$  is a latent variable and  $x_{t,i}$  is a column vector of baseline and paradata covariates of length  $m$ . The regression parameters are the slope parameters in the vector  $\alpha_t$ , and the random effect  $\delta_t$ . Apart from these, also the latent variable  $Z_{t,i}$  is updated in the Gibbs sampler, although it is not of direct interest. In a survey, we only observe whether  $Z_{t,i}(s_{1,t}) > 0$  occurs or not. In the main text, the

superscripts “C” and “P” are added to distinguish contact and participation models, but for the derivation of full conditionals this distinction is not important; it is fully analogous.

In the following, the vector of random effects in phase  $t$  for all possible histories of actions,  $s_{1,t-1}$ , is denoted by  $\delta_t$ . Obviously, each survey sample unit received just one treatment series of actions. We add the subscript  $i$  to indicate the strategy that was assigned to unit  $i$ , i.e.  $s_{i,1,t}$  is the series of actions assigned to unit  $i$  in phases 1 to  $t$ . We let  $\delta_{t,i}$  be the random effect that applies to unit  $i$ .

### A.1.1 Slope parameters in contact model

For  $\alpha_t(s)$ , the prior distribution is normal  $\alpha_t(s) \sim N(\mu(s), \Sigma(s))$ . The full conditional distribution is also normal, and we denote it as

$$(\alpha_t(s) | u_t, z_t, \delta_t, x, s_{1,t}) \sim N(\mu_{FULL}(s), \Sigma_{FULL}(s)) \quad (A1)$$

To derive the expectation and covariance of the full conditional distribution for action  $s$ , we need to restrict to sample units that reached phase  $t$  and for which  $s_{i,t} = s$ . Let this number be  $n_t(s)$ . For convenience, we label the units  $i = 1, 2, \dots, n_t(s)$ . Let  $z_{t,i}^* = z_{t,i} - \delta_{t,i}$  and let  $z_t^* = (z_{t,1}^*, \dots, z_{t,n_t(s)}^*)^T$ . Furthermore, let  $X$  be the  $n_t(s) \times m$  covariate matrix with sample units as rows.

It follows that the parameters in (A1) can be written as

$$\Sigma_{FULL}(s) = \left( (\Sigma(s))^{-1} + X^T X \right)^{-1}, \quad (A2)$$

$$\mu_{FULL}(s) = \Sigma_{FULL}(s) \left( (\Sigma(s))^{-1} \mu(s) + X^T z_t^* \right). \quad (A3)$$

### A.1.2 Random effects

We will now derive the posterior distribution of  $\delta_t(s_{1,t})$  for all possible strategies up to phase  $t$ ,  $s_{1,t}$ . We assume that the prior distribution is the same for all strategies,  $\delta_t(s_{1,t}) \sim N(0, \tau_t^2)$ . Furthermore, we assume that the random effects are independent. Hence, the full conditional distribution depends only on the outcomes of all sample units that reached phase  $t$  and that have exactly the strategy  $s_{1,t}$ , i.e.  $s_{1,t} = s_{i,1,t}$ . Let there be  $n_t(s_{1,t})$  such units, labelled  $i = 1, 2, \dots, n_t(s_{1,t})$ . The full conditional distribution

$$(\delta_t(s_{1,t}) | \alpha_t(s_t), u_t, z_t, x, s_{1,t}) \sim N(\mu_{FULL}(s), \Sigma_{FULL}(s)). \quad (A4)$$

Let  $\tilde{z}_{t,i}(s_{1,t}) = z_{t,i}(s_{1,t}) - \alpha_t(s_t)x_{t,i}$ , then it follows that the parameters in (A4) can be written as

$$\Sigma_{FULL}(s) = \left( (\tau_t^2)^{-1} + n_t(s_{1,t}) \right)^{-1}, \quad (A5)$$

$$\mu_{FULL}(s) = \Sigma_{FULL}(s) \sum_{i=1}^{n_t(s_{1,t})} \tilde{z}_{t,i}(s_{1,t}). \quad (A6)$$

### A.1.3 Latent response propensity

The last variables to update for the propensity models are the latent variables  $Z_{t,i}$  for the sample units that reached phase  $t$ . It holds that

$(Z_{t,i} | \alpha_t, \delta_t, x, s_{1,t}) \sim N(\alpha_t(s_{i,t})x_i + \delta_{t,i}, 1)$  distributed. When  $u_{t,i} = 1$ , then  $Z_{t,i} > 0$

and  $(Z_{t,i}|u_t, \alpha_t, \delta_t, x, s_{1,t})$  is the normal distribution restricted to the positive real axis. For  $u_{t,i} = 0$ ,  $(Z_{t,i}|u_t, \alpha_t, \delta_t, x, s_{1,t})$  is the normal distribution restricted to the non-positive real axis.

There are no explicit expressions for this distribution. When the outcome is  $u_{t,i} = 1$ , draws from  $N(\alpha_t(s_{i,t})x_i + \delta_{t,i}, 1)$  are repeated until a draw is positive. For  $u_{t,i} = 0$ , draws are repeated until a non-negative value is found.

## A.2. Full conditionals for regression parameters in costs models

We will derive the parameters of the posterior distributions of the parameters in the costs models. There are three such models, one for contact, one for participation and one for refusal. The derivation of full conditionals is fully analogous, so that we omit reference to the specific type of costs. The model has the form

$$C_i(s) = \gamma(s)x_i + \varepsilon_i(s), \quad (\text{A7})$$

$$\varepsilon_i(s) \sim N(0, \sigma^2(s)), \quad (\text{A8})$$

where  $x_i$  is the column vector of baseline covariates of length  $m$  and  $\varepsilon_i(s)$  is action-dependent error term. In this paper, the costs depend on the action that is applied but not on the phase in which it is applied. For actions that are applied in multiple phases, we, therefore, consider the average costs over all phases. The parameters that need to be updated and that require full conditional distributions are the slope parameters  $\gamma(s)$  and the dispersion parameters  $\sigma^2(s)$ .

In updating regression parameters for a specific action  $s$ , we need to restrict ourselves to sample units that were treated by action  $s$  at least once during the survey. Let there be  $n(s)$  such units. The observed  $c_i$  then is the average cost for the sample unit over all phases in which the action has been applied. We let  $c$  be the column vector of length  $n(s)$  containing the values of the sample units.

### A.2.1 Slope parameters

The prior distribution for the slope parameters is multivariate normal,  $\gamma(s) \sim N(\mu(s), \Sigma^2(s))$ . The full conditional distribution is also normal, and we denote it as

$$(\gamma(s)|c, x, \sigma^2) \sim N(\mu_{FULL}(s), \Sigma_{FULL}(s)). \quad (\text{A9})$$

Let  $X$  be the  $n(s) \times m$  covariate matrix with sample units as rows.

It follows for the parameters in (A9) that

$$\Sigma_{FULL}(s) = \left( \frac{1}{\sigma^2(s)} X^T X + (\Sigma^2(s))^{-1} \right)^{-1}, \quad (\text{A10})$$

$$\mu_{FULL}(s) = \Sigma_{FULL}(s) \left( \frac{1}{\sigma^2(s)} X^T c + (\Sigma^2(s))^{-1} \mu(s) \right). \quad (\text{A11})$$

### A.2.2 Variance of error term

The prior distribution is inverse gamma  $\sigma^2(s) \sim \Gamma^{-1}(a(s), b(s))$ . The full conditional is also inverse gamma

$$(\sigma^2(s)|\gamma, c, x) \sim \Gamma^{-1}(a_{FULL}(s), b_{FULL}(s)). \quad (A12)$$

Given the notation introduced earlier, we have for the parameters in (A12)

$$a_{FULL}(s) = a(s) + \frac{n(s)}{2} \quad (A13)$$

$$b_{FULL}(s) = b(s) + \frac{1}{2} (c(s) - X\gamma(s))^T (c(s) - X\gamma(s)). \quad (A14)$$

## Appendix B: Elicitation of hyperparameters in the Health Survey simulation study

In the simulation study, we have probit regression models for contact and participation and linear regression models for costs. In phase 1 there is only a contact model whereas in phases 2 and 3 there are also models for participation. Hence, in total there are ten models. For an informative prior, hyperparameters are needed for all regression coefficients in all models. We elicited informative priors by assuming that a historic Health survey data set of sample size  $n = 10000$  was available. Per type of regression coefficient, we explain how we proceeded in constructing priors. Regression slope parameters in cost models: The slope parameters  $\gamma_C(s)$  and  $\gamma_R(s)$  are normally distributed and a saturated model with variable AgeGender is applied. In a saturated linear model, each parameter is estimated using only the sample units in the corresponding population stratum. Consider the parameter  $\gamma_C$  for a particular stratum and a particular strategy. Based on historic data, the parameter is estimated as the average of the observed individual costs  $c_{C,i}$  for sample units in the stratum that received the specified strategy. The average stratum costs is approximately normally distributed with expectation equal to the true  $\gamma_C$  and variance equal to  $\frac{\sigma_C^2}{n}$ . Given that we simulate data ourselves in this paper, we can derive hyperparameters directly from the simulation model values.

Regression dispersion parameters in cost models: The dispersion parameters  $\sigma_C^2(s)$  and  $\sigma_R^2(s)$  have an inverse Gamma distribution and are constant over population strata. Consider  $\sigma_C^2$  for a particular strategy. Given historic data, it is estimated as the sample variance over the observed individual costs  $c_{C,i}$  for sample units that received the specified strategy. The sample variance divided over the true variance  $\sigma_C^2$  and multiplied with  $n - 1$  is approximately  $\chi_{n-1}^2$  distributed. The expectation and variance of a  $\chi_{n-1}^2$  distribution are, respectively,  $n - 1$  and  $2(n - 1)$ . This means that the sample variance has an expectation and variance equal to, respectively,  $\sigma_C^2$  and  $2\sigma_C^4/(n - 1)$ . An inverse Gamma distribution,  $\Gamma^{-1}(\alpha, \beta)$ , has expectation and variance equal to, respectively,  $\beta/(\alpha - 1)$  and  $\beta^2/((\alpha - 1)^2(\alpha - 2))$ . Hence,  $\alpha$  and  $\beta$  can be derived as  $\alpha = 2 + \frac{1}{2}(n - 1)/\sigma_C^2$  and  $\beta = \sigma_C^2 + \frac{1}{2}(n - 1)$ . Again under the simulation model, these hyperparameters can be derived directly from the simulation values.

Regression slope parameters in contact/participation models: The elicitation of hyperparameters is analogous for contact and participation models. Because of the probit link function, there is no explicit expression for estimators for the regression



slope parameters. Given that we include only main effects for baseline covariates,  $x_0$ , and paradata,  $x_1$ , it is, therefore, not straightforward how to choose hyperparameters based on historic data in an analytic way. For this reason, we simulated 2000 datasets of size 10000 and fitted probit regression models to each dataset. Over the 2000 fitted vectors of parameters, means and variances were computed for single slope parameters and covariances for pairs of slope parameters. These means, variances and covariances were used as hyperparameters. Somewhat surprisingly, absolute covariances were sometimes quite large, especially between the slope parameters of the two vectors  $x_0$  and  $x_1$ . Obviously this approach can only be applied in a simulation study; for a real historic dataset another approach is needed.

## Appendix C: Simulation study propensities and costs

Tables C.1 and C.2 present the simulation contact propensities, participation propensities and contact and participation costs, respectively, per phase and AgeGender  $\times$  Break-off subgroup.

**Table C.1: Contact and participation propensities per phase and subgroup.**

	Phase 1	Phase 2		Phase 3	
	Response	Contact	Participation	Contact	Participation
15-24, F, no break-off	0.30	0.70	0.58	0.50	0.30
15-24, F, break-off	0	0.70	0.87	0.50	0.59
15-24, M, no break-off	0.28	0.70	0.64	0.50	0.36
15-24, M, break-off	0	0.70	0.93	0.50	0.65
25-44, F, no break-off	0.33	0.74	0.49	0.62	0.22
25-44, F, break-off	0	0.74	0.94	0.62	0.66
25-44, M, no break-off	0.31	0.74	0.44	0.62	0.16
25-44, M, break-off	0	0.74	0.94	0.62	0.66
45-65, F, no break-off	0.35	0.80	0.38	0.90	0.10
45-65, F, break-off	0	0.80	0.95	0.90	0.72
45-65, M, no break-off	0.40	0.80	0.36	0.90	0.08
45-65, M, break-off	0	0.80	0.95	0.90	0.69

**Table C.2: Contact and participation costs per unit per phase and subgroup. Standard deviations are given within brackets.**

	Phase 1	Phase 2		Phase 3	
	Response	Contact	Participation	Contact	Participation
15-24, F, no break-off	3 (1)	11 (1)	16 (1)	14 (1)	16 (1)
15-24, F, break-off	3 (1)	11 (1)	12 (1)	14 (1)	12 (1)
15-24, M, no break-off	3 (1)	12 (1)	14 (1)	15 (1)	14 (1)
15-24, M, break-off	3 (1)	12(1)	10 (1)	15 (1)	10 (1)
25-44, F, no break-off	3 (1)	10 (1)	21 (1)	13 (1)	21 (1)
25-44, F, break-off	3 (1)	10 (1)	19 (1)	13 (1)	19 (1)

25-44, M, no break-off	3 (1)	10 (1)	20 (1)	13 (1)	20 (1)
25-44, M, break-off	3 (1)	10 (1)	18 (1)	13 (1)	26 (1)
45-65, F, no break-off	3 (1)	8 (1)	26 (1)	11 (1)	18 (1)
45-65, F, break-off	3 (1)	8 (1)	21 (1)	11 (1)	21 (1)
45-65, M, no break-off	3 (1)	9 (1)	24 (1)	12 (1)	24 (1)
45-65, M, break-off	3 (1)	9 (1)	20 (1)	12(1)	20 (1)

---

## Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/'16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/'14–2015/'16	Crop year, financial year, etc., 2013/'14 to 2015/'16 inclusive

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

## Colofon

### *Publisher*

Statistics Netherlands  
Henri Faasdreef 312, 2492 JP The Hague  
[www.cbs.nl](http://www.cbs.nl)

### *Prepress*

Statistics Netherlands, Studio BCO

### *Design*

Edenspiekermann

### *Information*

Telephone +31 88 570 70 70, fax +31 70 337 59 94  
Via contactform: [www.cbsl.nl/information](http://www.cbsl.nl/information)

© Statistics Netherlands, The Hague/Heerlen/Bonaire 2017.

Reproduction is permitted, provided Statistics Netherlands is quoted as the source.