DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 3, 2017, pp. 687–708, http://dx.doi.org/10.1515/JOS-2017-0032

Robustness of Adaptive Survey Designs to Inaccuracy of Design Parameters

Joep Burger[1], *Koen Perryck*[2], *and Barry Schouten*[2,3]

Adaptive survey designs (ASDs) optimize design features, given 1) the interactions between the design features and characteristics of sampling units and 2) a set of constraints, such as a budget and a minimum number of respondents. Estimation of the interactions is subject to both random and systematic error. In this article, we propose and evaluate four viewpoints to assess robustness of ASDs to inaccuracy of design parameter estimates: the effect of both imprecision and bias on both ASD structure and ASD performance. We additionally propose three distance measures to compare the structure of ASDs. The methodology is illustrated using a simple simulation study and a more complex but realistic case study on the Dutch Travel Survey. The proposed methodology can be applied to other ASD optimization problems. In our simulation study and case study, the ASD was fairly robust to imprecision, but not to realistic dynamics in the design parameters. To deal with the sensitivity of ASDs to changing design parameters, we recommend to learn and update the design parameters.

Key words: Mixed-mode survey; mode effect; optimization problem; sensitivity analysis; Travel Survey.

1. Introduction

Adaptive survey designs (ASDs) are based on the presumption that population units vary in their reaction to survey questionnaires. This variation in response and answering behavior leads to a suboptimal allocation of limited resources in a uniform, one-size-fits-all survey design. Adaptive survey designs attempt to optimize resource allocation by adapting design features to characteristics of the units that are known at the outset of the survey or that are observed during data collection (Wagner 2008; Schouten et al. 2013a; Chesnut 2013; Särndal and Lundquist 2013; Tourangeau et al. 2017). An optimal allocation can also further reduce nonresponse bias on survey target variables than calibration alone (Schouten et al. 2016).

In order to be able to adapt, a range of relevant characteristics needs to be known, a set of design features needs to be at hand, and the interactions between these two need to be observed in earlier waves of the same, or a similar, survey, or in earlier phases of data

[1] Statistics Netherlands, Department of Process Development and Methodology, CBS-weg 11, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. Email: j.burger@cbs.nl
[2] Statistics Netherlands, Department of Process Development and Methodology, The Hague, The Netherlands. Emails: koenperryck@hotmail.com and jg.schouten@cbs.nl
[3] Utrecht University, Faculty of Social and Behavioral Sciences, Utrecht, The Netherlands.
Acknowledgments: We thank Coen van Heukelingen, Carin Zwaneveld and Miranda de Vree for data and information about the Dutch Travel Survey, and the associate editor and two anonymous reviewers for useful comments and suggestions.

collection. Obviously, these may be strong requirements, so that it must be assumed that such design parameters are subject to imprecision and bias; estimated interactions may be noisy due to relatively small sample sizes and may be too optimistic or pessimistic. In this article, we investigate the robustness of optimal adaptive survey designs to imprecision and bias in design parameters.

An ASD requires estimates of key design parameters such as contact and participation propensities, and survey costs per sampling unit (Wagner 2013; Peytchev et al. 2010). These estimates are subject to estimation error and temporal changes, which may lead to imprecision and bias. Inaccurate design parameter estimates will affect the optimal choices, that is, the structure of an ASD, or the performance of an ASD when inaccuracy of design parameters is not accounted for. A change in the optimal strategy allocation to sampling units, that is, the structure of the ASD, may lead to practical and logistical problems when running the survey. A lower performance of the ASD may lead to lower response rates and representativeness, a budget overrun, and/or insufficient precision in survey estimates. One option to account for such change in structure and performance is to define data collection phases and postpone decisions on the exact implementation of strategies to the beginning of each phase (Groves and Heeringa 2006; Laflamme and Karaganis 2010). Another option is to explicitly account for uncertainty in the optimization itself (Bruin et al. 2016). In this article, we first explore and assess robustness of an ASD against imprecise and biased design parameters.

We address three questions: 1) How can we sensibly measure change in structure and performance of an ASD? 2) How robust are the structure and performance of an ASD to random error, that is, imprecision, in parameters? 3) How robust are the structure and performance of an ASD to bias in parameters? Inaccuracy of design parameters affects all survey designs, but since ASDs require estimates at a more detailed level, they may be more sensitive than non-adaptive, uniform designs. Nonetheless, the research questions are important also for non-adaptive (uniform) designs.

We set the investigation in the context of multi-mode surveys, because we see the survey mode as one of the most powerful and influential design features. We follow Calinescu et al. (2013), Calinescu and Schouten (2015), and Perryck (2015) in that we formulate mathematical optimization problems and consider the combined effect of mode-specific selection biases and mode-specific measurement biases (Schouten et al. 2013b). In this setting, survey design parameters are, thus, not restricted to response propensities and cost parameters, but they also include measurement differences. However, the proposed methodology and the strategy adopted to evaluate robustness is general and may be applied to any range of survey design features.

We motivate the answers to the research questions using a simple simulation study and a more complex but realistic application to the Dutch Travel Survey in the years 2010–2013. The Travel Survey showed a strong decline over these years in two of the three survey modes, web and telephone, and, hence, design parameter estimates in any given year are biased in subsequent years. The Travel Survey case study is an example of a static ASD (Bethlehem et al. 2011, 396), that is, the stratification of the population is based solely on data that are available at the start of data collection. The case study could be extended in a relatively straightforward way to dynamic designs, which include paradata

to stratify the population during data collection. An example of a dynamic ASD is given in the simulation study.

The outline of the article is as follows. In Section 2, we set the framework for ASD optimization and discuss strategies and measures for the evaluation of robustness. In Section 3, we illustrate the methodology with a simple simulation study. In Section 4, we present the case study linked to the Dutch Travel Survey. We end with a discussion in Section 5.

2. Framework for Exploring Robustness of Adaptive Survey Designs

In this section, we formulate ASD and set basic notation (2.1), discuss approaches to assessing robustness of ASDs to inaccuracy of design parameter estimates (2.2), and define three measures for comparing the structure of ASDs (2.3).

2.1. Setting and Notation

We formulate ASD as a mathematical optimization problem and we largely follow the notation of Calinescu and Schouten (2015). ASDs can also be formulated more loosely, without a strict notation, but to explore robustness a framework will be convenient.

In constrained optimization, an optimization parameter needs to be found so that an objective function is either minimized or maximized, given one or more constraints. In an ASD, the optimization parameter is the probability $p_{s,g}$ of allocating strategy $s \in \{1, 2, \ldots, S\}$, say the choice of survey mode or the amount of an incentive, to group $g \in \{1, 2, \ldots, G\}$, say an age class or a neighborhood characteristic as observed by an interviewer. To avoid confusion with strategy s, we use the term group rather than stratum or subpopulation. The objective function $f(p_{s,g})$ can be a cost or error function that is minimized, or a quality function that is maximized. In the latter case, an obvious constraint function $h(p_{s,g})$ is a cost function that may not exceed a budget.

How to stratify the population into G groups is not at all straightforward and depends on the focus on survey outcome variables, the quality and cost criteria in the ASD, and, obviously, on the availability of auxiliary variables from administrative data and paradata. In surveys with a wide range of outcome variables or in panels, the traditional focus is on the explanation of nonresponse: variables are selected in the ASD that explain nonresponse (see Schouten and Shlomo 2015 for a discussion and example). In surveys with a few outcome variables, the resulting set of auxiliary variables may be subsampled on the basis of their association to the outcome variables; groups are only formed when they are homogeneous both in nonresponse and in outcomes (e.g., Särndal and Lundquist 2013). ASDs do not need to be limited to a focus on nonresponse. First attempts to extend ASD to moderate measurement error have been made (e.g., Calinescu and Schouten 2016). Any ASD optimization problem also has one or more dual problems, that is, by switching the roles of objective function and constraints; the most obvious being cost minimization given constraints on quality. Hence, ASD groups may also be formed based on their homogeneity in terms of costs and survey outcome variables.

The stratification into population groups need not be limited to auxiliary variables available at the outset of data collection. Paradata variables, for example interviewer observations on the dwelling or neighborhood, break-off in a mode, for contact history,

may be used to form subgroups. Once they become available, the strategy allocation is further refined.

A data collection strategy s is the whole of design choices that are considered. Typically, the strategies in an ASD differ in certain design features, for example the application of a nonresponse follow-up or an incentive boost, but not in others, for example the contact strategy or the content of the questionnaire. A group may receive a specific strategy or a randomized mix of the selected strategies. Ideally, the strategies are chosen such that they counteract nonresponse and/or measurement error of the targeted population groups.

In order to determine the optimal design, we need input information about strategy- and group-specific design parameters. What effect does strategy s have on the response propensity and answering behavior of units in group g, and at what cost? Such strategy- and group-specific design parameters need to be estimated in order to evaluate the objective function $f(p_{s,g})$ and constraint functions $h(p_{s,g})$. These estimates may be inaccurate. In the next subsection, we discuss causes for inaccuracy and approaches to assess robustness of ASDs to inaccuracy.

2.2. *Inaccuracy of Design Parameters and Assessment of Robustness*

We see three causes for inaccuracy of estimates of design parameters. First, the estimates may be imprecise, because they are estimated on a sample. Second, the estimates may be biased because they were based on a different survey, or on the same survey in an experimental setting. Third, the estimates may be unbiased at the time they were made, but in time the true values have changed. The inaccuracy may affect both the structure of the optimal strategy allocation and the performance of the ASD. The impact on structure is undesirable from an operational perspective, and the impact on performance is undesirable from a quality and cost perspective. We sketch four approaches to assess robustness. Before we proceed, we make two remarks.

We consider the robustness of adaptive designs. However, an assessment of robustness is equally sensible and valuable for a non-adaptive design. Also the parameters of non-adaptive designs may be inaccurate and deviations of expected quality and/or costs occur frequently. The main difference is that adaptive designs require parameter estimates at a more detailed level, that is, the population groups. Hence, parameter estimates are more susceptible to inaccuracy of adaptive designs than in non-adaptive designs.

Second, estimation of design parameters need not be done using a single experiment or parallel run, but may be part of a more flexible design in which some of the data collection resources are allocated to learn and update. Such continuous learning would be most natural in a Bayesian context, in which historic survey data and expert knowledge are employed to set prior distributions for design parameters. Resources would then be allocated to promising strategies with weakly informative priors. In reinforcement and machine learning, this is called the exploration-exploitation tradeoff (Sutton and Barto 2012). ASD optimization would, however, have to be reconsidered (Bruin et al. 2016), which is beyond the scope of this article.

We see four different, but complementary, approaches to evaluate robustness in a non-Bayesian setting; two of them focus on imprecision and the other two on bias in design

parameters. The impact of imprecision can be assessed in two ways. First, by resampling the survey data and optimizing the ASD for each resample, we may evaluate the variability in the structure of the strategy allocation. Second, we may optimize the ASD for the original sample and evaluate the variability in its performance when applied to resamples of the survey data in terms of quality and costs, for example how often is the budget overrun or are response rates below the specified threshold.

The impact of bias can also be assessed in two ways by considering different waves of the same survey. First, we can optimize the ASD in both waves and compare the structure of the strategy allocation between waves. Second, we can simulate and assess the performance of an ASD optimized in one wave of the survey in a new wave. The optimal strategy allocation probabilities are applied to the new wave of the survey and quality and costs are compared.

Obviously, the simultaneous impact of imprecision and bias can also be assessed, but is harder to interpret and to convert to counter measures. In the next subsection, we make clear what we mean by assessing the structure of the strategy allocation, which is very important when implementing ASD in practice.

2.3. Measures for Comparing the Structure of Adaptive Survey Designs

Since an ASD can consist of many $(G \times S)$ allocation probabilities, we introduce three distance measures, labeled d_1 to d_3, to compare the structure of two designs, say design A and B. All three measures are bounded between 0 and 1, where 0 means that the designs are the same and 1 that they are maximally different.

The first measure is the relative difference in sample size between design A and design B:

$$d_1(\mathrm{A}, \mathrm{B}) = \frac{|n(\mathrm{A}) - n(\mathrm{B})|}{n(\mathrm{A}) + n(\mathrm{B})}, \tag{1}$$

where $n(\delta)$ is the sample size for design δ.

For the second and third measure, we partition the probability $p_{s,g}$ of allocating strategy s to group g into the probability of sampling a population unit in group g, $\sum_s p_{s,g} = 1 - p_{\emptyset,g}$ and the conditional allocation probability given being sampled, $\tilde{p}_{s,g} = \frac{p_{s,g}}{1-p_{\emptyset,g}}$. The second measure is thus the Euclidean distance in group sampling probabilities between design A and design B:

$$d_2(\mathrm{A}, \mathrm{B}) = \frac{N}{n(\mathrm{A}) + n(\mathrm{B})} \sqrt{\Sigma_g w_g (p_{\emptyset,g}(\mathrm{B}) - p_{\emptyset,g}(\mathrm{A}))^2} \tag{2}$$

where $p_{\emptyset,g}(\delta)$ is the probability that a population unit in group g is not sampled in design δ, w_g the weight or relative frequency of group g in the population $\left(w_g = \frac{N_g}{N}\right)$ and N the population size. For simplicity we assume that N and w_g are the same in both designs. We use sampling probability and inclusion probability interchangeably $\left(\sum_s p_{s,g}\right)$.

The third measure is the Euclidean distance in conditional allocation probabilities between design 1 and design 2:

$$d_3(\mathrm{A}, \mathrm{B}) = \sqrt{\frac{1}{2G} \sum_{g=1}^{G} \sum_{s=1}^{S} (\tilde{p}_{s,g}(\mathrm{A}) - \tilde{p}_{s,g}(\mathrm{B}))^2} \tag{3}$$

where $\tilde{p}_{s,g}(\delta) = \frac{p_{s,g}(\delta)}{1-p_{\emptyset,g}(\delta)}$ is the allocation probability conditional on being sampled in design δ. The scaling factor $\frac{1}{2}$ is added to bound d_3 by 1.

In Sections 3 and 4, we demonstrate the utility of the three measures.

3. Simulation Study

We present a simplified example to demonstrate sensitivity analyses and the utility of the design structure measures d_1 to d_3. Assume a survey with two data collection phases, where the second phase is expensive and optional, for example a first phase using a web questionnaire followed by a second phase using face-to-face interviewing. At the onset of the survey, one relevant auxiliary variable, say age, is available for all sampling units. During the first phase one paradata observation is made, say a binary indicator for a web break-off. We divide the sample into four groups based on the two variables, say {$\leq$25 years, $>$25 years} $\times$ {no break-off, break-off}, and seek to optimize the follow-up probability per group. For simplicity, we assume that the groups have an equal population size. Tables 1 and 2 give fictive estimates of the design parameters response propensity and costs per sampling unit for each of the four groups in the two phases and in two consecutive years, say t and $t+1$.

For year t, standard deviations are given in Tables 1 and 2 to mimic the uncertainty in the estimated values. The response propensities are drawn from a Beta(α,β) distribution. The two shape parameters are chosen such that the expectation and the square root of the variance match the values in Table 1. The costs per sampling unit are drawn from a normal distribution $N(\mu,\sigma^2)$, where μ and σ are given in Table 2. The response propensities and costs per sampling unit in year $t+1$ are used to mimic bias through time change.

Now we have set the simulation framework, we specify the optimization problem. We maximize the response rate, subject to constraints on the total costs and the number of respondents in the two age groups. The budget is set at 38 per responding unit and 750 respondents are required per age group. The budget is too low to allocate all Phase 1 nonrespondents to Phase 2. A budget of 40 per responding unit would be needed to do a full follow-up.

In order to further simplify the analyses, we consider only follow-up probabilities that are multiples of 0.1, that is, in the set $\{0, 0.1, 0.2, \ldots, 1\}$. Given that we consider four groups and eleven follow-up probabilities, there are $11^4 = 14{,}641$ possible designs. This relatively small number of candidate designs allows us to do a brute force optimization, that is, evaluate all candidate designs and select the best design.

Table 1. Fictive group response propensities for two phases in two consecutive years t and t + 1. For the first year, standard deviations are given in parentheses.

			Year t		Year $t+1$	
Group	Baseline	Paradata	Phase 1	Phase 2	Phase 1	Phase 2
1	$\leq$25	No break-off	0.20 (0.02)	0.40 (0.04)	0.18	0.40
2		Break-off	0	0.70 (0.04)	0	0.70
3	$>$25	No break-off	0.40 (0.02)	0.50 (0.04)	0.35	0.50
4		Break-off	0	0.80 (0.04)	0	0.80

Table 2. Fictive group costs per sampling unit for two phases in two consecutive years t and t + 1. For the first year, standard deviations are given in parentheses.

			Year t		Year $t+1$	
Group	Baseline	Paradata	Phase 1	Phase 2	Phase 1	Phase 2
1	≤ 25	No break-off	4 (0.1)	25 (2)	4	30
2		Break-off	5 (0.1)	30 (2)	5	32
3	>25	No break-off	4 (0.1)	20 (1)	4	22
4		Break-off	5 (0.1)	25 (2)	5	26

We perform four analyses:

1. The effect of *imprecise* estimates of design parameters on
   a. ASD structure
   b. ASD performance
2. The effect of *biased* estimates of design parameters on
   a. ASD structure
   b. ASD performance

The effect of imprecise estimates of the design parameters (1) is analyzed by sampling 1,000 replications from the Beta and normal distributions of the response propensity and costs. The effect on ASD structure (1a) is analyzed by optimizing each replication and comparing the 1,000 ASDs using the distance measures. The effect on ASD performance (1b) is analyzed by solving the optimization problem once using the point estimates of the design parameters, and evaluating the response rate, costs and number of respondents per group of the resulting ASD when applied to each of the 1,000 sets of design parameters.

The effect of bias in design parameter estimates (2) is analyzed by using propensities and costs from two years. The effect on ASD structure (2a) is analyzed by optimizing each year and comparing the two ASDs using the distance measures. The effect on ASD performance (2b) is analyzed by solving the optimization problem in one year, and evaluating the response rate, costs and number of respondents per group of the resulting ASD when applied to the next year.

1a) Effect of Imprecise Input on ASD Structure

The optimal sample size (d_1) and sample allocation probabilities (d_2) vary relatively little from one sample to the other (Figure 1); both measures exhibit large peaks close to zero. In contrast, the strategy allocation structure (d_3) shows a lot of variation and the median distance is 0.35. In the example, the optimal sampling design does not vary a lot, but optimal Phase 2 follow-up probabilities vary considerably between replications.

1b) Effect of Imprecise Input on ASD Performance

The variation in response rate is modest: the interquartile range is $63 - 61 = 2$%-point (Figure 2). As a result, the number of respondents per age group (not shown) is fairly stable. In 52% of the draws the number of respondents met the threshold and in only four percent of the draws the number of respondents was below 95% of the threshold. However, the costs show more variation: in 44% of the draws the costs exceeded the threshold, in

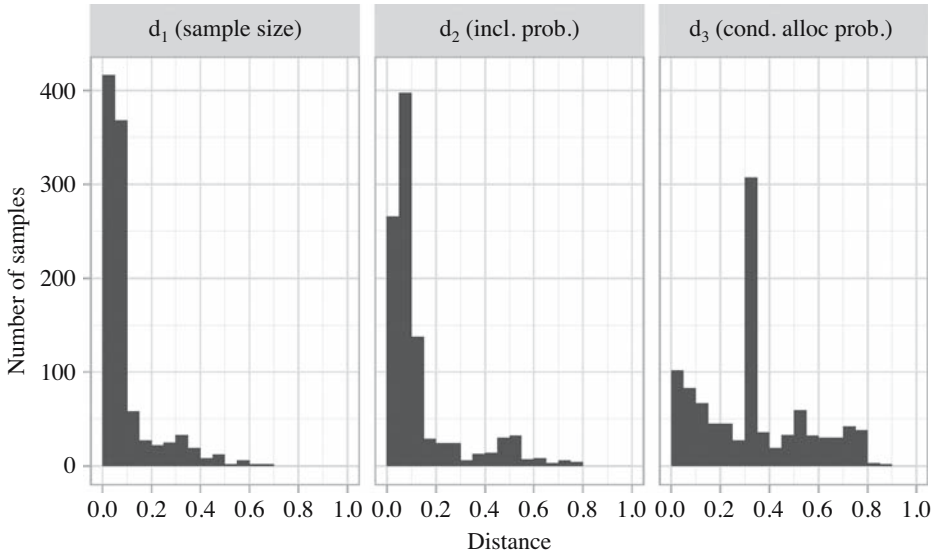

Fig. 1. Histograms for the three metrics d_1, d_2 and d_3 for 1,000 simulated sets of design parameters in year t.

18% of the draws the costs exceeded five percent of the budget, and in five percent of the draws the costs exceeded ten percent of the budget. As is usually done in survey contracts, some level of variation is explicitly acknowledged. A ten percent increase in budget is common, but even then in five percent of cases this would be insufficient.

2a) Effect of Biased Input on ASD Structure

The distance measures in optimal designs between years are $d_1(t, t+1) = 0.19$, $d_2(t, t+1) = 0.23$ and $d_3(t, t+1) = 0.58$. Starting with the sample sizes, d_1 points at a fairly large change. The optimal sample sizes are 2,505 and 3,705, respectively, for year t and $t+1$, and are indeed quite different. Since the $t+1$ design has limited follow-up in Phase 2, most of the response must come from Phase 1, and, hence, requires a much larger sample size. Moving to the group inclusion probabilities, d_2 also points at a fairly large change. The relative shares of the groups in the sample are $n_g(t)/n(t) = (0.30, 0.30, 0.20, 0.20)$

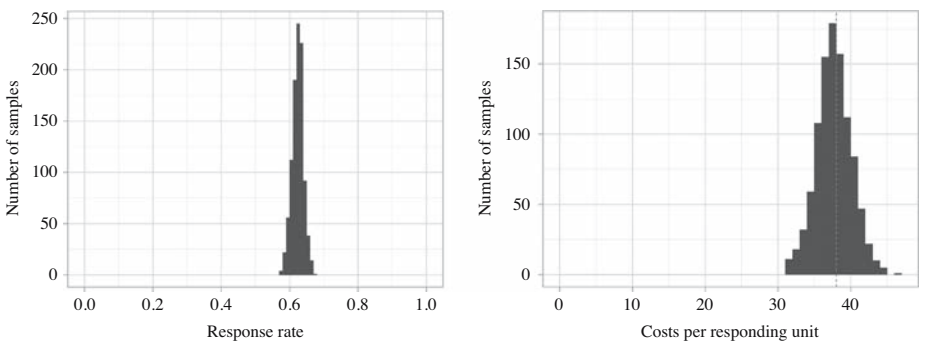

Fig. 2. Histograms for the response rate (left) and costs per responding unit (right) of the optimal design in year t applied to 1,000 simulated sets of design parameters. Dashed line shows budget.

and $n_g(t+1)/n(t+1) = \big(0.23,\ 0.23,\ 0.27,\ 0.27\big)$. Note that population units can be allocated based only on age, so that the relative shares for Phase 1 break-offs and non-break-offs are the same. In year t, persons of 25 and under are oversampled, whereas in year $t+1$ persons over 25 years are oversampled. Indeed, the sampling design is quite different. Finally, the d_3 value points at a large difference in strategy allocation structure. The optimal follow-up probabilities conditional upon being sampled are $\tilde{p}_g(t) = (0.3,\ 1,\ 1,\ 1)$ and $\tilde{p}_g(t+1) = (0,\ 1,\ 0,\ 0.5)$; the allocation structure to Phase 2 has clearly changed in all groups, except the break-offs by persons of 25 years and under. In this example, the decrease in Phase 1 response propensities implies a big change in the form of the optimal design.

2b) Effect of Biased Input on ASD Performance

When the optimal design for year t is applied in year $t+1$, then the response rate drops by 1.1%-point from 62.4% to 61.3%, the number of respondents in the age groups drops on average also by 1.1%, and the costs per responding unit increase to 41.4 and lead to a budget overrun of almost 10%. We can conclude that the drop in Phase 1 response propensities has a strong impact on the performance.

In Section 4, we analyze the effects on imprecise and biased input on ASD structure and performance for a real example linked to the Dutch Travel Survey.

4. Case Study

We apply the proposed methodology to the Dutch Travel Survey (DTS, abbreviated to OViN in Dutch) for the years 2010 through 2013. In this case study, we minimize mode-effects subject to constraints on costs, precision and response rate by optimizing follow-up probabilities. After briefly introducing the survey (4.1), we will explain how subpopulations and strategies are defined (4.2), specify the optimization problem (4.3), estimate the design parameters across time, including the sampling error from bootstrap samples (4.4), solve the optimization problem (4.5), introduce a number of scenarios with different constraints (4.6), and assess the robustness against imprecise input (4.7) and biased input (4.8).

4.1. Dutch Travel Survey

The DTS monitors the travel behavior of the Dutch non-institutional population of all ages. The sampling design is a stratified random sample, where strata are defined by region. The current survey design is a uniform sequential mixed-mode design. Each sampling unit is requested by mail to fill out a web questionnaire. After two mail reminders, nonrespondents are approached for a telephone interview (CATI) if a telephone number is available and for a face-to-face interview (CAPI) if not. For CATI, three contact attempts are made, each consisting of a calling attempt in the morning, afternoon and evening. For CAPI, up to seven visits are made.

4.2. Population Groups and Strategies

In an ASD, the design features are adapted to each stratum. As explained in Section 2, we refer to stratum or subpopulation as group g to avoid confusion with strategy s. Population

groups were defined using a CHAID/regression tree analysis, maximizing the between-group variance and minimizing the within-group variance in a key target variable: total distance traveled on a particular day. As auxiliary variables, we used gender (2), age (18), degree of urbanization (5), region (12), origin (3), car ownership (3), household income (6), day of the week (7), household size (6), socioeconomic status (5), and availability of a telephone number (2), with the number of categories in parentheses. The resulting stratification into $G = 8$ groups is shown in supplementary Table S1, including the labels used in subsequent figures (supplementary file available online at: http://dx.doi.org/10.1515/JOS-2017-0032).

From the sequential mixed-mode design Web–CATI/CAPI (s_3), we could deduce two more strategies without the need for additional experiments: Web only (s_1) and Web–CATI (s_2). The current Web–CATI/CAPI design was set as benchmark strategy and consequently has no mode effect by definition. It has the highest response propensities, but is also the most expensive design.

4.3. The DTS Optimization Problem

The DTS considers removing part of the CATI and CAPI follow-up in order to reduce the costs of the survey. The impact on DTS survey statistics, in this study the total distance travelled on a given day, must, however, be minimized. DTS users demand a minimum precision in relevant population groups, which is translated to minimum numbers of respondents. Furthermore, DTS users and Statistics Netherlands have also agreed on a minimum response rate. Hence, the DTS constrained optimization problem is defined by one cost function (survey costs) and three quality functions (method effect on distance travelled, numbers of respondents, and response rate).

We choose to minimize the absolute method effect under constraints on the total costs, the numbers of respondents in the eight groups defined in Subsection 4.2, and the overall response rate. The roles of costs and method effect could be switched to define a dual problem, but, since the budget cut is most real, we focus on optimizing quality. The constraint parameters are denoted as B (budget), R_g (minimum number of respondents in group g) and Γ (minimum response rate).

In the optimization, we search for the best allocation probabilities, $p_{s,g}$, of the three possible strategies s (Web only, Web–CATI, Web–CATI/CAPI) to each of the eight groups g. The optimization problems in terms of the strategy allocation probabilities are formulated in the Supplement (available online at: http://dx.doi.org/10.1515/JOS-2017-0032).

The quality and cost function depend on three design parameters: the method effects $D_{s,g}$, the response propensities $\rho_{s,g}$ and the costs per sampling unit $c_{s,g}$. These parameters need to be estimated.

4.4. Estimating the Design Parameters

We discuss estimation of each of the three types of design parameters.

The first design parameter, the relative mode effect $D_{s,g}$ is defined as the difference between two population parameter estimates, $\hat{\theta}$ estimated from the response by group g

using either strategy s or a benchmark strategy BM:

$$D_{s,g} = \hat{\theta}_{s,g} - \hat{\theta}_{\mathrm{BM},g}. \tag{4}$$

The population parameter $\hat{\theta}_{s,g}$ is a key target variable, here the total distance traveled on a particular day. It is estimated using the inclusion weights of the sampling design, corrected for selective nonresponse by multiplicatively weighting the appropriate subsamples to known population distributions of auxiliary variables (Deming and Stephan 1940). The auxiliary variables are the same as those used to define the population groups (Subsection 4.2). Since the population parameters for strategies s_1 and s_2 are estimated from subsamples of the original sample, the original weighting model of the DTS is too detailed and is reduced to the main effects.

Thus, the mode effect is operationalized as the difference between weighted estimates, and is the net effect of a difference in selection and a difference in measurement. The latter, we would call a pure mode effect or measurement effect. The more effective the weighting model, the more it will correct for the selection effect and the more similar the mode effect will be to the measurement effect (of strategy s relative to the benchmark strategy).

The second design parameter, the response propensity $\rho_{s,g}$, is estimated by multiplying the overall response rate $\frac{r}{n}$ by the cumulative proportion of respondents that participated in strategy s:

$$\rho_{s,g} = \frac{r}{n} \frac{\sum_{i \in \mathcal{R}} w_i I_{gi} J_{si}}{\sum_{i \in \mathcal{R}} w_i I_{gi}}, \tag{5}$$

where w_i is the calibration weight of responding unit i, I_{gi} the binary indicator for responding unit i being an element of group g and J_{si} is the binary indicator for responding unit i participating in strategy s, and $\mathcal{R}$ the response set.

Population sizes were estimated by summing the calibration weights of the respondents. The calibration weights are produced in the regular statistical production process based on generalized regression calibration (Särndal et al. 1992). In the calibration, the selected auxiliary variables from administrative data are included as well as a range of additional variables (see CBS 2015 for the exact weighting model).

The third design parameter, the cost per sampling unit $c_{s,g}$ is derived from cost calculation sheets provided by data collection staff. In addition to fixed costs that are similar for each group, costs differ between groups, because they differ in response propensities and telephone registration propensities.

The design parameters $D_{s,g}$, $\rho_{s,g}$ and $c_{s,g}$ were estimated for $S = 3$ strategies times $G = 8$ groups times four years (Figures 3–5). The uncertainty in the estimates was quantified by drawing 1,000 bootstrap samples from the data and recalculating the design parameters for each bootstrap sample. Multicollinearity between the design parameters is thus taken into account.

The estimates of the relative mode effects, $D_{s,g}$ (Figure 3) show wide confidence intervals relative to the point estimates. They generally did not show obvious trends over time, except for units in group g_7 (0_18 + _<30k: unemployed adults with relatively low household income), for whom the total distance traveled was estimated about one hundred

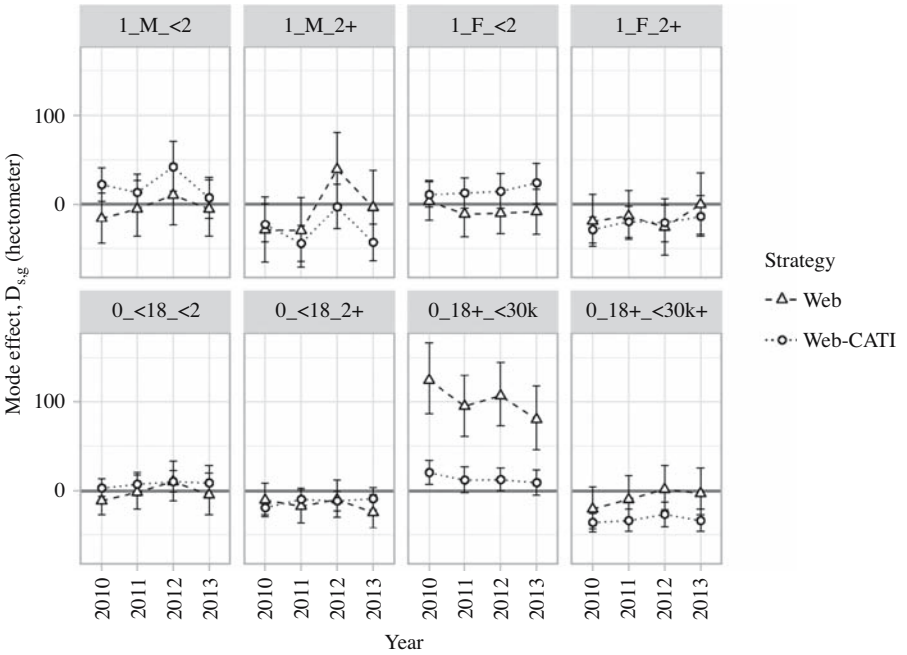

Fig. 3. Uncertainty and dynamics in estimated mode effect, $\hat{D}_{s,g}$ per strategy s (legend) and group g (panels) in the Dutch Travel Survey. Mode effect is defined relative to benchmark strategy Web–CATI/CAPI. Shown are 2.5th, 50th and 97.5th percentiles.

hectometers (ten kilometers) higher in Web than in benchmark strategy Web–CATI/CAPI.

The estimates of the response propensities, $\rho_{s,g}$ (Figure 4) show narrower confidence intervals relative to the point estimates than the estimates of the mode effects (see Figure 3). The response propensity estimates declined for Web and Web–CATI, but not for Web–CATI/CAPI. The decline in response to Web–CATI is caused by a decline in the availability of telephone numbers to call web nonrespondents. As a result, more web nonrespondents receive a personal interview (CAPI).

The costs, $c_{s,g}$ were estimated with narrow confidence intervals relative to the point estimates (Figure 5). Web is obviously the cheapest strategy with on average 2.3 euro per sampling unit, whereas Web–CATI is about 3 to 4 times more expensive, and the benchmark Web–CATI/CAPI even 10 to 13 times in 2010. For Web–CATI/CAPI costs strongly rose over time, because both the web response and the availability of telephone numbers declined and more sampling units were assigned to the most expensive CAPI mode.

4.5. *Solving the Optimization Problem*

To find the optimal allocation probabilities, $p_{s,g}$, the optimization problem (supplementary Eq. S1, available online at: http://dx.doi.org/10.1515/JOS-2017-0032) was solved using the R package `nloptr` (R Core Team 2014; Ypma 2015), which is an R interface to NLopt, an open-source library for nonlinear optimization (Johnson 2016). As starting

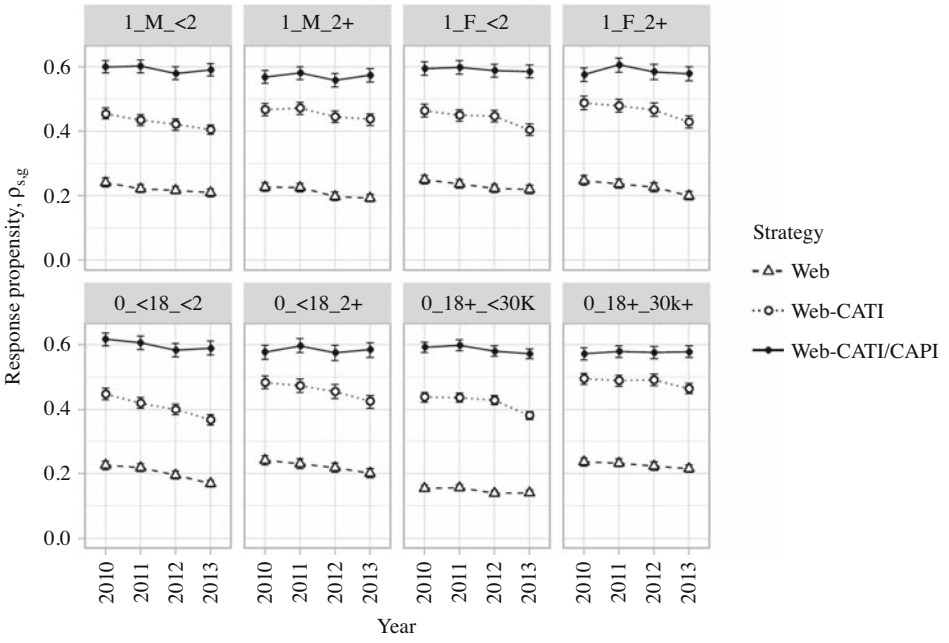

Fig. 4. Uncertainty and dynamics in estimated response propensity, $\hat{\rho}_{s,g}$ per strategy s (legend) and group g (panels) in the Dutch Travel Survey. Shown are 2.5th, 50th, and 97.5th percentiles.

values we chose the allocation probabilities under the current uniform Web−CATI/CAPI design. Based on previous experiences with global and local search algorithms in the NLopt library (Calinescu and Schouten 2015), we selected the local search algorithm Constrained Optimization BY Linear Approximations (COBYLA, Powell 1994). An optimization was run using the design parameters $D_{s,g}$, $\rho_{s,g}$, and $c_{s,g}$ estimated from the original sample (resulting in a design referred to as ORt or $p_{s,g}^{\mathrm{OR}t}$ for a given year t) or from a bootstrap sample (resulting in a design referred to as BSt or $p_{s,g}^{\mathrm{BS}t}$). To reduce runtime, optimizations were executed in parallel using the R packages `foreach` and `doParallel` (Calaway et al. 2015a; Calaway et al. 2015b) on an eight-core machine.

4.6. Scenarios

In all scenarios (Table 3), we set the minimum total number of respondents, R and allocate this over groups according to what has been realized in the field: $R_g = \frac{r_g^*}{\sum_g r_g^*} R$. The first scenario was defined by the current DTS, which has a budget of $B = 1.8$ mln euro and is aimed at obtaining at least $R = 35{,}000$ respondents and a response rate of at least $\Gamma = 0.585$. To reduce costs, we studied to what extent R and Γ need to be compromised when the budget would be cut to $B = 1.5$ mln euro – a reduction of almost 17%. Survey stakeholders have indicated that they are willing to compromise to $\Gamma = 0.565$. Moreover, from 2010 through 2013 the average number of respondents was closer to about $R = 32{,}000$. Reducing the budget without compromising R or Γ resulted in no solutions (supplementary Fig. S1, available online at: http://dx.doi.org/10.1515/JOS-2017-0032): none of the bootstraps converged and met the constraints. Either R or Γ has to be relaxed,

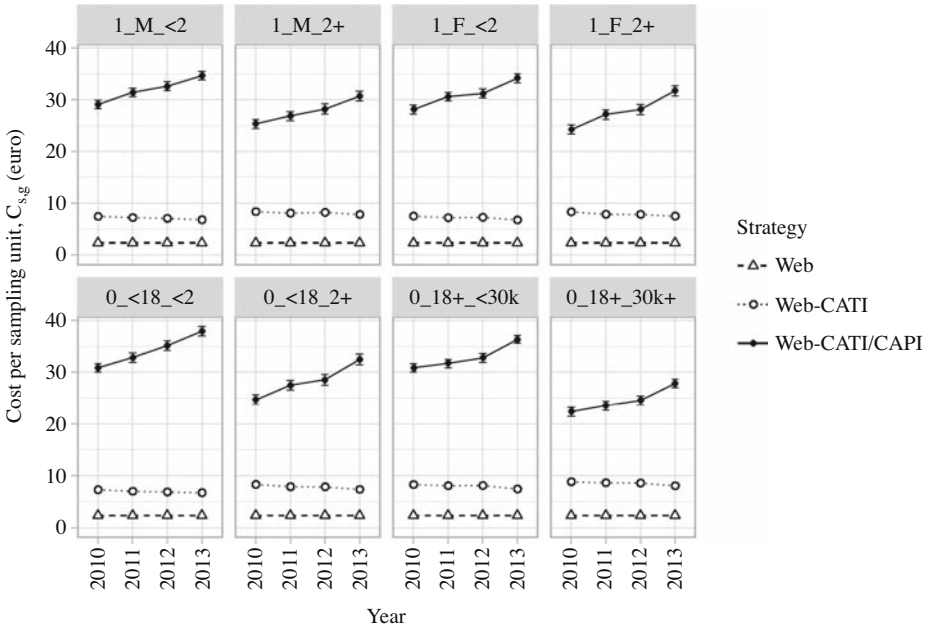

Fig. 5. Uncertainty and dynamics in estimated cost per sampling unit, $\hat{c}_{s,g}$ per strategy s (legend) and group g (panels) in the Dutch Travel Survey. Shown are 2.5th, 50th and 97.5th percentiles.

or both to some lesser extent. Scenarios 2 and 3 were chosen along the diagonal of supplementary Fig. S1 for more detailed analyses.

4.7. *Robustness Against Imprecise Input*

In this subsection, we assess the effect of imprecise estimates of design parameters on ASD structure and performance. The effect of biased estimates will be discussed in the next subsection.

ASD Structure

Here we compare the structure of the bootstrapped ASDs with the structure of the original ASD. In all three scenarios and eight groups the inclusion probability ($\sum_s p_{s,g}$) was around 0.004 and slightly decreased with decreasing R (see supplementary Fig. S2, available online at: http://dx.doi.org/10.1515/JOS-2017-0032). Within-group variance in inclusion

Table 3. Scenarios defined by the constraint parameters.

Scenario	Label†	Budget, B (mln euro)	Minimum number of respondents, R	Minimum response rate, Γ
1	Current	1.8	35,000	0.585
2	B ↓ RR ↓	1.5	35,000	0.570
3	B ↓ SE ↑	1.5	32,000	0.585

†B: budget, RR: response rate, SE: standard error, ↓ (↑): decrease (increase) relative to Scenario 1.

probabilities generally inflates the variance of the population parameter estimates (Kish 1987). This variance in inclusion probabilities was highest in Scenario 1 ('Current'), where the inclusion probabilities were most sensitive to inaccuracy of the design parameters.

In Scenario 1 ('Current'), the budget is sufficient to almost uniformly administer the benchmark strategy Web−CATI/CAPI (see supplementary Fig. S3 available online at: http://dx.doi.org/10.1515/JOS-2017-0032). When the budget is cut, some portion of each group should be administered the Web−CATI strategy. This means that some group members will not be visited for a personal interview (CAPI) if they do not respond to the web questionnaire and their telephone number is unavailable for a telephone interview (CATI). This effect is strongest when the constraint on the precision remains fixed and only the constraint on the response rate is relaxed (Scenario 2: 'B ↓ RR ↓ '). In that case the conditional allocation probability is fairly sensitive to inaccuracy of the design parameter estimates. When the constraint on the precision is relaxed and the constraint on the response rate remains fixed (Scenario 3: 'B ↓ SE ↑ ') the optimal design remains almost uniformly the benchmark strategy.

The distance measures summarize these patterns (Figure 6). The distance in overall sample size (d_1, Eq. 1) was small but highest in Scenario 1 ('Current'), which also had the highest variation in inclusion probabilities (see Fig. S2). This is also expressed in more detail by the distance in inclusion probabilities (d_2, Eq. 2). The distance in conditional allocation probabilities (d_3, Eq. 3) corroborate the details in supplementary Fig. S3 (available online at: http://dx.doi.org/10.1515/JOS-2017-0032).

ASD Performance

When applying the original ASD to the bootstrap samples, the objective function, that is, the population average mode effect $\bar{D}_{\mathrm{BM}}$ (Eq. 4) was successfully eliminated in all three scenarios (not shown here but integrated in Figure 8 below). Given the constraints, the objective function was very robust against inaccuracy of the design parameter

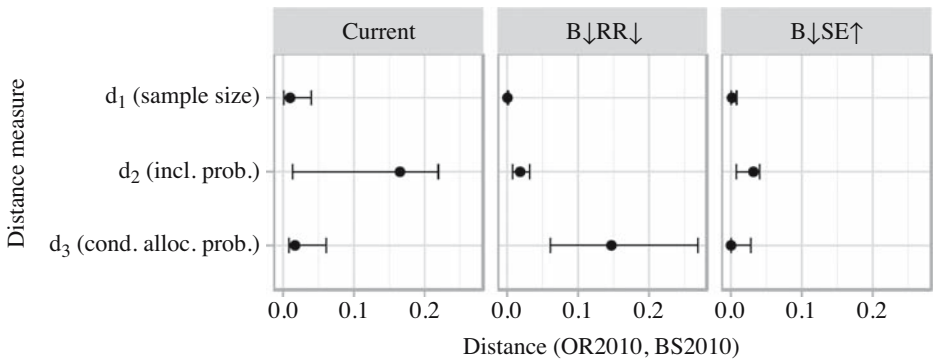

Fig. 6. Robustness of ASDs against imprecision of design parameter estimates of the 2010 Dutch Travel Survey, for three scenarios (panels). Each distance is measured between two optimal designs, one using the design parameters from the original sample (OR2010), the other using the design parameters estimated from a bootstrap sample (BS2010). Shown are 2.5th, 50th, and 97.5th percentiles of distances between ASDs that converged and met the constraints.

estimates – if a solution was found (see supplementary Fig. S1, available online at: http://dx.doi.org/10.1515/JOS-2017-0032). Only when a lower budget was compensated by relaxing the constraint on the response rate (Scenario 2: 'B ↓ RR ↓') some bootstraps resulted in a population average mode effect of up to about one hectometer.

In 95% of the samples, the budget is not overrun by more than eleven thousand euro (Scenario 1: 'Current'), that is, less than 0.7% of the budget in all scenarios (supplementary Fig. S4, available online at: http://dx.doi.org/10.1515/JOS-2017-0032). In all scenarios, the precision is still met when design parameters are estimated from a bootstrap sample. In 95% of the samples, the design is no more than 145 respondents short to meet the constraint on the response rate (Scenario 1: 'Current'), that is, less than 0.5% in all scenarios.

We conclude that the performance of the ASD is robust against imprecise input. The uniform benchmark design, on the other hand, either structurally overruns the budget by 20 thousand euro, that is, 1.3% of the budget (low-budget Scenarios 2 and 3), or is structurally short of more than two thousand respondents, that is, 6% of the minimum (high-precision Scenarios 1 and 2), or both (Scenario 2).

4.8. Robustness Against Biased Input

In this subsection, we assess the effect of biased estimates of design parameters on ASD structure and performance.

ASD Structure

For the DTS, the design parameter estimates changed over time: response rates decreased in the strategies Web and Web–CATI (see Figure 4), and costs increased in the benchmark strategy Web–CATI/CAPI (see Figure 5). These changes were so severe that in 2011 the majority of bootstraps no longer yielded a solution in the acceptable range of R and Γ at $B = 1.5$ mln euro (see supplementary Fig. S1, available online at: http://dx.doi.org/10.1515/JOS-2017-0032). When both R and Γ were relaxed – a scenario not included in Table 3 – the valid ASDs for 2010 and 2011 differed mostly in the conditional allocation probabilities (Figure 7). Further loosening of the constraints B, R or Γ would be necessary to find more solutions that could meet those constraints. The design was, thus, not robust against strong but real temporal changes in design parameter estimates. Note that this holds for both the uniform and the adaptive design.

ASD Performance

When the ASDs of 2010 from the previous section were applied to subsequent years, the population average mode effect, $\bar{D}_{s3}$ could no longer be eliminated (Figure 8). Again, this effect is strongest when a lower budget was compensated by relaxing the constraint on the response rate (Scenario 2: 'B ↓ RR ↓'). By definition, the uniform benchmark design causes no mode effect and its population average mode effect is therefore insensitive to dynamics in design parameter estimates.

The ASDs of 2010 administered high proportions of the benchmark strategy Web–CATI/CAPI (see supplementary Fig. S3 available online at: http://dx.doi.org/10.1515/JOS-2017-0032). The costs of this strategy, $c_{s3,g}$ rose strongly over time (see

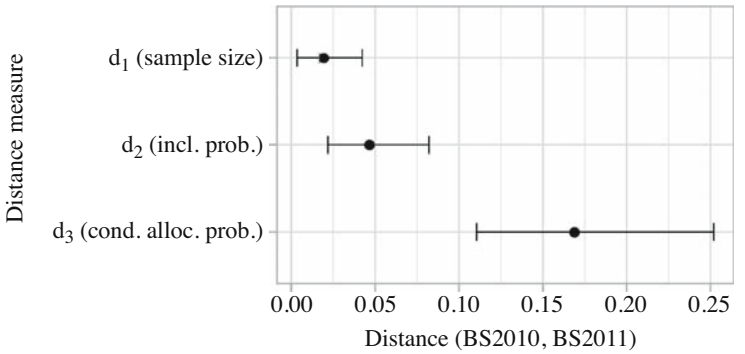

Fig. 7. Robustness of ASDs against dynamics in design parameter estimates of the Dutch Travel Survey. Each distance is measured between two optimal designs, one using the design parameters estimated from a bootstrap sample in 2010 (BS2010), the other from a bootstrap sample in 2011 (BS2011). Shown are 2.5th, 50th, and 97.5th percentiles of distances between ASDs that converged and met the constraints in both years (only 282 of 1,000 bootstrap samples). $B = 1.5$ mln euro, $R = 32{,}000$, $\Gamma = 0.565$ ('$B\downarrow SE\uparrow RR\downarrow$').

Figure 5). As a result, the budget would be overrun considerably when the ASDs of 2010 were applied in subsequent years (Figure 9A: $C - B > 0$). Note that in the low-budget Scenarios 2 and 3, the uniform benchmark design would overrun the budget even more (and as we have seen already overran the budget in 2010). In the more expensive Scenario 1 ('Current') the uniform benchmark design seemed cheaper than the adaptive designs, but there it did not meet the constraint on the minimum number of respondents (Figure 9B: $R - r > 0$). The variation is caused by the variation in the design parameter estimates, $D_{s,g}$, $\rho_{s,g}$ and $c_{s,g}$ (both designs) and the variation in the allocation probabilities, $p_{s,g}$ (adaptive design).

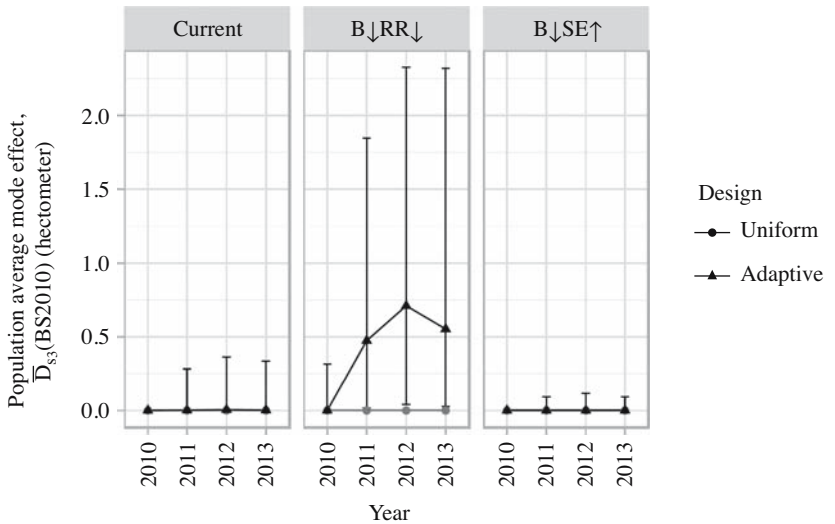

Fig. 8. Performance of the uniform benchmark design 2010 and ASDs 2010 (legend) in subsequent years of the Dutch Travel Survey, for three scenarios (panels). Performance is the ability to minimize the objective function, that is, the population average mode effect, $\bar{D}_{s3}$. Only bootstraps that converged and met the constraints for an ASD are shown for both designs. Shown are 2.5th, 50th, and 97.5th percentiles.

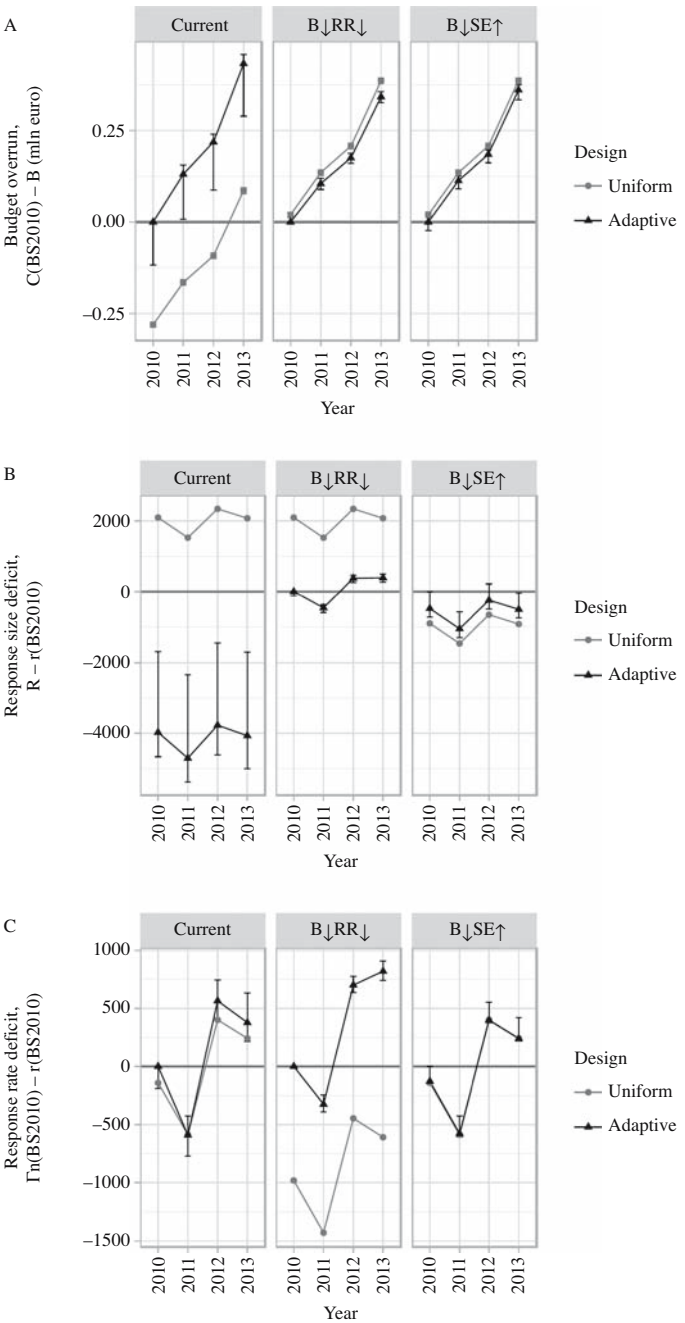

Fig. 9. Performance of the uniform benchmark design 2010 and ASDs 2010 (legend) in subsequent years of the Dutch Travel Survey, for three scenarios (panels). (A) budget overrun, $C - B$, (B) response size deficit, $R - r$, (C) response rate deficit, $\Gamma n - r$. Results at or below zero meet the constraint parameters, B, R, or Γ. Only bootstraps that converged and met the constraints for an ASD are shown for both designs. Shown are 2.5th, 50th, and 97.5th percentiles.

The response propensities of the benchmark strategy Web–CATI/CAPI, $\rho_{s3,g}$ were fairly stable over time (see Figure 4). As a result, in most groups the ASDs of 2010 still obtained (almost) enough respondents in subsequent years (Figure 9B: $R - r \leq 0$; supplementary Fig. S5: $R_g - r_g \leq 0$). As we have discussed in the previous section, in the scenarios with a tight constraint on precision (1 and 2), a uniform benchmark design was already in 2010 deficient in the number of respondents ($R - r > 0$).

The ASDs of 2010 would still meet the constraint on the response rate in 2011 (Figure 9C: $\Gamma n - r \leq 0$), but no longer in subsequent years ($\Gamma n - r > 0$). The uniform benchmark design could meet the constraint on the response rate, but only in Scenario 2 ('B ↓ RR ↓') where it could not meet the constraint on the response size (Figure 9B).

We conclude that the performance of the ASD is not robust against (strongly) biased input, although it outperforms the uniform design.

5. Discussion

To optimally adapt a survey design to the response and answering behavior of sampling units, design parameters need to be estimated. In this article, we propose and evaluate four viewpoints to assess robustness of adaptive survey designs to inaccuracy of design parameter estimates: the effect of both imprecision and bias on both ASD structure and ASD performance. We additionally propose three distance measures to compare the structure of ASDs. The methodology is illustrated using a simple simulation study and a more complex but realistic case study on the Dutch Travel Survey.

By bootstrapping the survey data and recalculating the design parameters, realistic combinations are obtained to assess an ASD's sensitivity to estimation error. The simulation study showed that optimal sample sizes and inclusion probabilities were robust to estimation error, but conditional allocation probabilities were not. The objective function was fairly robust to estimation error, but the optimal design could not meet all restrictions when applied to resampled data. The case study on the Dutch Travel Survey – with high variance in mode effect estimates – also showed that optimal sample sizes were robust to imprecision, but the robustness of the allocation probabilities depended on the constraint parameters. When applied to resampled data, the optimal DTS design remained successful in minimizing the objective function, and fairly successful in meeting the constraints. Variation in optimal allocation probabilities is undesirable for logistical and process-control reasons; it is practically not feasible to implement and monitor a constantly changing allocation structure.

When it comes to bias in design parameter estimates, the simulation study showed that the structure of the optimal design can change substantially in the next year. The optimal design also performed considerably poorer when applied to the next year. The DTS case study confirmed that both ASD structure and performance are very sensitive to bias. Scenarios that were feasible in 2010 became unsolvable in subsequent years because of sharply declining response rates in web and availability of phone numbers, unless one or more of the constraints were relaxed. Applying the optimal design to subsequent years revealed that the objective function could no longer be eliminated, the budget would be overrun and the minimum response rate would not be met, especially when the budget is cut without compromising on precision. It is important to note, however, that the uniform

benchmark design is a special, suboptimal case of an ASD and already did not meet the constraints in 2010.

Robustness will obviously improve when design parameters are estimated with greater precision, for example by larger sample sizes. Robustness might also improve when more constraints are added, such as a lower bound on the costs in addition to a budget. This might, however, reduce the probability of convergence and the number of solutions. Third, robustness could be improved by creating homogenous groups that differ in more than one survey outcome variable, and/or that differ in the design parameters, like the mode effects, response propensities, or costs. The more groups differ in the design parameters, the more an adaptive design can gain over a uniform design. Large between-group variances also urge the need for up to date information used to assign sampling units to groups. Finally, time series of design parameters could be modeled to predict potential problems such as declining response rates and availability of valid telephone numbers. When choosing the constraints, one could anticipate a narrower solution space than at present.

In conclusion, sensitivity of adaptive survey designs to inaccuracy of key input parameters urges the need for a more sophisticated choice of population groups, and for a frequent updating of the design parameters and the optimal ASD. Even when input parameters evolve greatly, mode effects might still be minimized to obtain unbiased statistics, but only at the expense of lower response sizes and consequently less precise statistics. Hence, robustness of the resulting (adaptive) survey design needs to be balanced against the precision and bias of the key statistics.

Nonlinear optimization problems can result in local minima, and, thus, suboptimal solutions. Optimization may also be stopped before a solution is found if the objective function or optimization parameters change too little between iterations. The probability of finding the global minimum could be increased by choosing a larger set of starting values, possibly after a brute force optimization as employed in the simulation study, or choosing another search algorithm. Future research should be directed at more sophisticated, more informed methods to optimize ASDs.

Future research should also set ASD optimization in a Bayesian context where historic survey data and expert knowledge are entered through prior distributions on the design parameters. Some of the costs that can be saved by switching from a uniform to an adaptive design should be used to learn the design parameters. The design parameter distributions would then be updated during and after data collection, leading to an integral analysis of sensitivity. The faster the design parameters evolve, the more frequently they need to be updated. A Bayesian framework is currently under development to allow the learning and updating of the design parameters (Bruin et al. 2016). Performing optimization of ASDs in a Bayesian setting is not straightforward, however, and would have to be combined with a more flexible design of surveys.

This article is inspired from realistic and practical quality-costs tradeoffs. The framework and metrics are especially designed for data collection staff involved in monitoring and analysis and in (re)designing surveys. The analyses in this article help assessing uncertainty in quality and costs and may support or guide design decisions. However, a broader application is warranted in order to evaluate the utility and use in practice.

6. References

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken: Wiley.

Bruin, L., N. Mushkudiani, and B. Schouten. 2016. "A Bayesian Analysis of Mixed-Mode Data Collection." Paper presented at the 71st Annual Conference of the American Association for Public Opinion Research, Austin TX, May 12–15. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/BADEN/workshop-2016/AAPOR_Bruin-Mushkudiani-Schouten.pdf (accessed June 2017).

Calaway, R., Revolution Analytics, and S. Weston. 2015a. "Package 'foreach." Available at: https://cran.r-project.org/package=foreach (accessed June 2017).

Calaway, R., Revolution Analytics, S. Weston, and D. Tenenbaum. 2015b. "Package 'doParallel'." Available at: https://cran.r-project.org/package=doParallel (accessed June 2017).

Calinescu, M. and B. Schouten. 2015. "Adaptive Survey Designs to Minimize Survey Mode Effects—a Case Study on the Dutch Labor Force Survey." *Survey Methodology* 41: 403–425.

Calinescu, M. and B. Schouten. 2016. "Adaptive Survey Designs for Nonresponse and Measurement Error in Multi-Purpose Surveys." *Survey Research Methods* 10: 35–47. Doi: http://dx.doi.org/10.18148/srm/2016.v10i1.6157.

Calinescu, M., S. Bhulai, and B. Schouten. 2013. "Optimal Resource Allocation in Survey Designs." *European Journal of Operational Research* 226: 115–121. Doi: http://dx.doi.org/10.1016/j.ejor.2012.10.046.

CBS. 2015. *Onderzoek Verplaatsingen in Nederland 2015. Onderzoeksbeschrijving*. The Hague/Heerlen: Statistics Netherlands. Available at: https://www.cbs.nl/-/media/_pdf/2016/38/2016ep27.pdf (accessed June 2017).

Chesnut, J. 2013. *Model-Based Mode of Data Collection Switching from Internet to Mail in the American Community Survey*. Washington: US Census Bureau. Available at: https://census.gov/content/dam/Census/library/working-papers/2013/acs/2013_Chesnut_01.pdf (accessed June 2017).

Deming, W.E. and F.F. Stephan. 1940. "On a Least Squares of Adjustment of a Sampled Frequency Table When the Expected Totals Are Known." *Annals of Mathematical Statistics* 11: 427–444. Doi: http://dx.doi.org/10.1214/aoms/1177731829.

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society A* 169: 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00423.x.

Johnson, S.G. 2016. "The NLopt Nonlinear-Optimization Package." Available at: http://ab-initio.mit.edu/wiki/index.php/NLopt (accessed June 2017).

Kish, K. 1987. "Weighting in Deft2." *The Survey Statistician* 17: 26–30.

Laflamme, F. and M. Karaganis. 2010. "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." Paper presented at the European Conference on Quality in Official Statistics (Q2010), Helsinki, May 4–6. Available at: https://q2010.stat.fi/sessions/session-29 (accessed June 2017).

Perryck, K. 2015. *Assessing the Impact of Inaccuracy in Design Parameters on the Performance of Adaptive Survey Designs*. Utrecht: Utrecht University. (MSc thesis.)

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias through Case Prioritization." *Survey Research Methods* 4: 21–29. Doi: http://dx.doi.org/10.18148/srm/2010.v4i1.3037.

Powell, M.J.D. 1994. "A Direct Search Optimization Method that Models the Objective and Constraint Functions by Linear Interpolation." In *Advances in Optimization and Numerical Analysis*, edited by S. Gomez and J.-P. Hennart, 51–67. Dordrecht: Kluwer Academic.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.r-project.org/ (accessed June 2017).

Särndal, C.-E. and P. Lundquist. 2013. "Aspects of Responsive Survey Design with Applications to the Swedish Living Conditions Survey." *Journal of Official Statistics* 29: 557–582. Doi: http://dx.doi.org/10.2478/jos-2013-0040.

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

Schouten, B. and N. Shlomo. 2015. *Selecting Adaptive Survey Design Strata with Partial R-indicators*. The Hague/Heerlen: Statistics Netherlands. (Discussion paper 201521.)

Schouten, B., M. Calinescu, and A. Luiten. 2013a. "Optimizing Quality of Response Through Adaptive Survey Designs." *Survey Methodology* 39: 29–58.

Schouten, B., J. van den Brakel, B. Buelens, J. van der Laan, and Th. Klausch. 2013b. "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys." *Social Science Research* 42: 1555–1570. Doi: http://dx.doi.org/10.1016/j.ssresearch.2013.07.005.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner. 2016. "Does More Balanced Survey Response Imply Less Non-Response Bias?" *Journal of the Royal Statistical Society A* 179: 727–748. Doi: http://dx.doi.org/10.1111/rssa.12152.

Sutton, R.S. and A.G. Barto. 2012. *Reinforcement Learning: An Introduction*. Second edition. Cambridge MA: MIT Press.

Tourangeau, R., J.M. Brick, S. Lohr, and J. Li. 2017. "Adaptive and Responsive Survey Designs: a Review and Assessment." *Journal of the Royal Statistical Society Series A* 180: 203–223. Doi: http://dx.doi.org/10.1111/rssa.12186.

Wagner, J. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*. Ann Arbor: University of Michigan. (Doctoral thesis.)

Wagner, J. 2013. "Adaptive Contact Strategies in Telephone and Face-to-face Surveys." *Survey Research Methods* 7: 45–55. Doi: http://dx.doi.org/10.18148/srm/2013.v7i1.5037.

Ypma, J. 2015. "Package 'nloptr'." Available at: https://cran.r-project.org/package=nloptr (accessed June 2017).

Received March 2016
Revised June 2017
Accepted June 2017