**Adaptive Survey Design in the Dutch Health Survey 2021**

Kees van Berkel, Suzanne van der Doef and Barry Schouten

*Statistics Netherlands, July 2020*

**Abstract**

Challenges that surveys are facing are increasing data collection costs and declining budgets. During the past years, many surveys at Statistics Netherlands were redesigned to reduce cost and to increase or maintain response rates. Currently, alternative approaches are investigated to produce more accurate estimates within the same budget. Adaptive survey design is proposed for achieving this goal.

In 2018 adaptive survey design has been implemented successfully for the Dutch Health survey. In this survey the sequential mixed mode strategy internet observation followed by face-to-face observation is applied. The feature to adapt is the face-to-face sample: through stratified selection of internet non-respondents for face-to-face follow-up, nonresponse bias can potentially be reduced.

In 2021 changes in the adaptive survey design for the Dutch Health survey will be implemented. This paper contains the highlights of the design. This includes the estimation of response probabilities, the motivation for the choice of the strata in the face-to-face sampling design, how to reduce nonresponse bias with constraints on precision and costs, and the effect of the adaptive survey design on the main survey estimates.

*Key words:* balanced response, nonresponse bias, accuracy, data collection costs.

1. **Introduction**

Adaptive survey design assumes that differentiation of effort over relevant population subgroups is either effective in improving survey quality or efficient in reducing survey costs. The designs have received a lot of interest over the last decade in response to budget pressure due to gradual but persistent declines of response rates, see Chun et al. (2018). Despite the interest in the designs, the number of implemented case studies described in the literature is still relatively small. Schouten et al. present a general overview of adaptive survey designs.

Statistics Netherlands introduced adaptive survey design in the Health Survey 2018, see Van Berkel et al. (2020). The designs for 2019 and 2020 are almost identical and do not differ much from the 2018 design, although the definition of the target groups has been slightly modified. In 2021 a larger change is made, with substantially less face-to-face observation.

This paper reads as follows. Section 2 describes some methodology with emphasis on assumptions, bias estimation and feature to adapt. Section 3 presents the development of the Health survey design 2021, with expected impact on survey results. Section 4 ends with a discussion of limitations and possibilities for improvement.

1. **Methodology**

The aim of adaptive survey design is to get a better balanced response by putting different effort in different groups of the population. Adaptive survey design is effective in improving survey results, or reducing survey costs.

*Assumptions*

When designing the observation strategy, the following assumptions are made.

1. The sample is a probability sample of fixed size *n*.
2. Each person in the population has a positive inclusion probability $π\_{k}$.
3. Response follows the ‘Random response model’ in which person *k* responds with response probability $ρ\_{k}$ known only to person *k*.
4. Respondents' answers to questions are independent of the method of observation.

Assumption four states that there are no mode-specific measurement errors. However, there are indications that some indicators of the Health Survey are affected by these errors. We will return to this in the discussion.

*Estimating bias*

The aim of the survey is to estimate population means for several target variables. An estimator for the population mean is the modified Horvitz-Thompson estimator ${\left(\sum\_{kϵΡ}^{}{y\_{k}}/{π\_{k}}\right)}/{\left(\sum\_{kϵΡ}^{}{1}/{π\_{k}}\right)}$, where $Ρ$ denotes the set of all respondents, and $y\_{k}$ the value of the target variable $Y$ for person *k*. Observe that in the case of simple random sampling, the estimator equals the response mean.

In general the estimator is biased, unless all response probabilities $ρ\_{k}$ are equal. Bethlehem (1988) showed that the bias can be approximated by$ {R\left(ρ,Y\right) S\_{ρ }S\_{Y}}/{\overbar{ρ}}$ . Here $R$ denotes Pearson’s correlation coefficient, $S$ the population standard deviation, and $\overbar{ρ}$ the population response mean.

Since Pearson's correlation coefficient is between -1 and 1, an upper limit for the bias can be given by ${ S\_{ρ }S\_{Y}}/{\overbar{ρ}}$ . Since the standard deviation the target variables cannot be influenced by the design of the observation strategy, the only way to minimise the upper limit for the bias is to minimise the coefficient of variation of the response probabilities $CV\left(ρ\right)={ S\_{ρ }}/{\overbar{ρ}}$.

*Design feature to adapt*

To minimise $CV\left(ρ\right)$, the observation strategy is adapted. The ideal situation is to obtain a high overall response rate, with little variation in the individual response probabilities. Consider a simple random sample with the sequential mixed-mode strategy CAWI → CAPI. All sampled people are first asked by letter to participate in the survey by completing a questionnaire on the internet. People who have not responded to this request after no more than two reminders are visited at home by interviewers to conduct the survey.

The design feature to adapt in this strategy is the CAPI follow-up. To reduce the variation in response rates, more CAPI follow-up is deployed for groups of people who react badly via the internet than for those who do well. Observe that the adaptation may be at the expense of the overall response rate. The identification of the so-called target groups is carried out using cluster analysis.

1. **Developing the Health Survey Design 2021**

In order to predict people’s response behaviour, results of the Health Survey 2019 were used. People were clustered into target groups based on personal characteristics, in such a way that within each group there is little variation in response behaviour per mode, and between two groups response behaviour differs in at least one mode. After the clustering, the coefficient of variation of response probabilities had to be minimised subject to constraints on e.g. budget, response numbers or rates, and sample sizes per mode of observation. The solution of this problem contains the CAWI sample sizes and CAPI-sampling fractions per target group, and an estimate of $CV(ρ)$.

*Understanding response behaviour*

To explain response behaviour of people, a classification tree algorithm was applied to Health Survey 2019 data. Demographic and regional characteristics were linked to the sample. Examples are ethnicity, income, age, gender, marital status, household composition, urbanity and province. The most explanatory variables for response behaviour turned out to be ethnicity, income and age. The algorithm distinguished ethnicity categories ‘migrants’ and ‘Dutch nationals’, income categories quintiles 1, 2, and 3-5, and age categories 0-11, 12-24, 25-44, 45-64 and 65+.

*Clustering the target population*

K-means clustering was applied with the selected characteristics and categories by the classification tree. This method divides people into groups which are homogeneous according to response behaviour, where outliers can be detected. The advantage of this method is that small groups with extremely high or low response rates can be identified as target groups. A disadvantage may be that the target groups are less homogeneous according to the characteristics used. Figure 1 shows, as a result of the clustering, the eight target groups for the Health Survey 2021.

Figure 1 Target groups Health Survey 2021

|  |  |  |  |
| --- | --- | --- | --- |
|   | Dutch nationals | Migrants |  |
| Age         Income | 1 | 2 | 3 – 5 | 1 | 2 | 3 – 5 |
| 0-11 | 2 | 2 | 2 | 1 | 1 | 7 |
| 12-24 | 3 | 3 | 7 | 8 | 1 | 3 |
| 25-44 | 3 | 3 | 3 | 5 | 5 | 5 |
| 45-64 | 3 | 7 | 4 | 5 | 5 | 5 |
| 65+ | 3 | 7 | 6 | 5 | 5 | 4 |

*Minimising* $CV\left(ρ\right)$ *subject to constraints*

The problem to be solved is to minimise $CV\left(ρ\right)=S(ρ)/\overbar{ρ}$ subject to the constraint that the lower bound of the 95% confidence interval of the expected number of respondents equals 9,500. This has been done to be fairly sure that at least 9,500 responses will be achieved. It was further assumed that the CAWI-sample is a simple random sample of people living in the Netherlands, and that 3% of this sample is not eligible for CAPI follow-up. Contrary to previous years, no logistical and cost restrictions have been set for 2021. The response probabilities produced by the classification tree algorithm were used to estimate $CV\left(ρ\right)$.

The problem was solved with the R package Alabama using a Lagrangian minimization algorithm for optimizing smooth nonlinear functions with constraints. Different random starting values for the CAWI sample size and the CAPI sampling fractions per target group were used, since the algorithm may end in a local minimum. The smallest $CV\left(ρ\right)$ was found at 0.103. The results of the corresponding solution are summarized in Table 1. Column n CAWI contains the CAWI sample size, r CAWI the expected number of CAWI respondents, p CAWI the expected CAWI response rate, n elig the number of CAWI nonrespondents eligible for face-to-face follow-up, n CAPI the CAPI sample size, f CAPI the CAPI sampling fraction n CAPI / n elig, r CAPI the expected number of CAPI respondents, p CAPI the expected CAPI response rate, r tot the final number of expected responses, and p tot the final response rates per target group.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| group | n CAWI  |  r CAWI | p CAWI | n elig | n CAPI | f CAPI | r CAPI | p CAPI | r tot | p tot |
|  |  |  | *%* |  |  | *%* |  | *%* |  | *%* |
| 1 | 604 | 146 | 24.2 | 440 | 318 | 72.4 | 151 | 47.3 | 297 | 49.1 |
| 2 | 1,723 | 819 | 47.5 | 853 | 74 | 8.7 | 46 | 61.6 | 864 | 50.2 |
| 3 | 5,201 | 1,606 | 30.9 | 3,439 | 2,276 | 66.2 | 967 | 42.5 | 2,573 | 49.5 |
| 4 | 3,808 | 1,596 | 41.9 | 2,098 | 681 | 32.4 | 292 | 42.9 | 1,888 | 49.6 |
| 5 | 2,891 | 622 | 21.5 | 2,182 | 2,182 | 100.0 | 662 | 30.4 | 1,285 | 44.4 |
| 6 | 1,972 | 1,062 | 53.8 | 851 | 0 | 0.0 | 0 | 47.2 | 1,062 | 53.8 |
| 7 | 3,066 | 1,125 | 36.7 | 1,849 | 808 | 43.7 | 383 | 47.4 | 1,508 | 49.2 |
| 8 | 363 | 76 | 21.0 | 276 | 276 | 100.0 | 79 | 28.6 | 155 | 42.7 |
| total | 19,629 | 7,052 | 35.9 | 11,988 | 6,615 | 55.2 | 2,580 | 39.0 | 9,632 | 49.1 |

Table 1 Adaptive survey design, Health Survey 2021

As expected, the lower the CAWI response rate in a target group, the more CAPI follow-up is applied. For the groups with the lowest CAWI response rates, groups 5 and 8, the CAPI sampling fraction equals 100%, resulting in final response rates of 44.4 and 42.7% respectively. These are considerably lower than the overall response rate of 49.1% and cannot be raised, unless other response improvement measures are taken. For the group with the highest CAWI response rate, group 6, the CAPI sampling fraction equals 0%. So for this group, the final response rate equals the CAWI response rate of 53.8%. This is higher than the overall survey response rate of 49.1%.

*Effect of adaptive survey design on survey results*

To get an idea of the effect of the adaptive survey design on the results of the Health Survey, simulations were carried out using bootstrapping. To that end, 1000 samples with replacement were drawn from the 2016 sample. The year 2016 was chosen, because then all CAWI non-respondents were approached via CAPI. Each bootstrap sample has the right CAWI and CAPI sizes per target group according to the adaptive survey design in Table 1.

After weighting the response per sample, ten core variables were estimated. These estimates were compared with the results of the Health Survey 2016. Most of the survey results with adaptive survey design do not differ much from those without adaptation. Significant shifts were found in using not-prescribed medicines, smoking, and physical activity. In the adaptive design, it is expected that the proportion of non-prescribed medicine users decreases by 1.6 percent points, the proportion of smokers decreases by 1.3 percent points, and the proportion of people who meet the standard of physical activity increases by 2.3 percent points.

1. **Discussion**

*Mode-specific measurement errors*

In the presented adaptive survey design, the existence of mode-specific measurement errors is not taken into account. This is a shortcoming because there are indications that these errors do exist.

The differences between estimates with CAWI response and those with CAPI response are relatively large for some variables. These differences are due to a combination of selection and mode effects. A selection effect may be that the composition of CAWI respondents by health characteristics is different from that of CAPI respondents. In the extreme case, only healthy people participate via the internet, and therefore only unhealthy people are eligible for a face-to-face interview. A mode effect could be that social desirability affects answers in face-to-face interviews.

Mode effects can be complex and subject to interactions between the demographics of the respondent and the subject. It is difficult to quantify their size and qualitative judgments by experts familiar with the subject and the observation modes are required.

A method for separating selection and mode effects is a survey design with re-interviews. Both CAWI non-respondents and a sample of CAWI respondents are approached via CAPI. Klausch et al. (2017) show that a design with re-interviews and a smaller initial sample can increase the accuracy of survey results with the same budget.

*Clustering the target population*

An important decision is the choice of the target groups. In this paper, the focus was on explanation of nonresponse. Target groups were based on administrative variables used in post-survey adjustments. When clustering the target population, it is probably better to also make use of administrative variables related to both response behaviour and survey results. Are there suggestions for such variables and how to use them in clustering?

*Uncertainty in estimated response rates*

Estimates of response rates are subject to uncertainties. Furthermore, response behaviour is not constant over time and depends on unexpected events such as the failure of the statistical office's server or the outbreak of an epidemic. The question is, how should the selection fractions be adjusted when estimates of response probabilities differ significantly from realisation figures?

1. **References**

Berkel van, K., S. van der Doef, and B. Schouten. 2020. “Implementing Adaptive Survey Design With an Application to the Dutch Health Survey.” *Journal of Official Statistics*, 36 (3): 609 – 629.

 Available at <http://dx.doi.org/10.2478/JOS-2020-0031> from September 2020.

Preview at <https://content.sciendo.com/downloadpdf/journals/jos/36/3/article-p609.xml>

Bethlehem, J.G. 1988. “Reduction of Nonresponse Bias Through Regression Estimation.” *Journal of Official Statistics* 4 (3): 251 – 260. Available at https://www.scb.se/contentassets/-

ca21efb41fee47d293bbee5bf7be7fb3/reduction-of-nonresponse-bias-throughregression-

estimation.pdf

Chun, A.Y., S.G. Heeringa, and B. Schouten. 2018. “Responsive and Adaptive Design for Survey Optimization.” *Journal of Official Statistics*, 34 (3): 581 – 597.

Available at <http://dx.doi.org/10.2478/JOS-2018-0028>

Klausch, T., B. Schouten, B. Buelens, and J. van den Brakel. 2017. “Adjusting measurement bias in sequential mixed-mode surveys using re-interview data.” *Journal of Survey Statistics and Methodology*, 5(4): 409-432. Available at <https://doi.org/10.1093/jssam/smx022>

Schouten, B., Peytchev, A., Wagner, J. (2017): Adaptive Survey Design, Series on Statistics Handbooks, Chapman and Hall/CRC Press.