



UNIVERSITY  
OF WOLLONGONG  
AUSTRALIA

University of Wollongong  
Research Online

---

Faculty of Engineering and Information Sciences -  
Papers: Part B

Faculty of Engineering and Information Sciences

---

2017

# Split Questionnaire Designs: collecting only the data that you need through MCAR and MAR designs

James O. Chipperfield  
*Australian Bureau of Statistics*

Margo L. Barr  
*New South Wales Ministry of Health*

David G. Steel  
*University of Wollongong, [dsteel@uow.edu.au](mailto:dsteel@uow.edu.au)*

---

## Publication Details

Chipperfield, J. O., Barr, M. L. & Steel, D. G. (2018). Split Questionnaire Designs: collecting only the data that you need through MCAR and MAR designs. *Journal of Applied Statistics*, 45 (8), 1465-1475.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: [research-pubs@uow.edu.au](mailto:research-pubs@uow.edu.au)

---

# Split Questionnaire Designs: collecting only the data that you need through MCAR and MAR designs

## **Abstract**

We call a sample design that allows for different patterns, or sets, of data items to be collected from different sample units a Split Questionnaire Design (SQD). SQDs can be thought of as incorporating missing data into survey design. This paper examines the situation where data that are not collected by an SQD can be treated as Missing Completely At Random or Missing At Random, targets are regression coefficients in a generalised linear model fitted to binary variables, and targets are estimated using Maximum Likelihood. A key finding is that it can be easy to measure the relative contribution of a respondent to the accuracy of estimated model parameters before collecting all the respondent's model covariates. We show empirically and theoretically that we could achieve a significant reduction in respondent burden with a negligible impact on the accuracy of estimates by not collecting model covariates from respondents who we identify as contributing little to the accuracy of estimates. We discuss the general implications for SQDs.

## **Disciplines**

Engineering | Science and Technology Studies

## **Publication Details**

Chipperfield, J. O., Barr, M. L. & Steel, D. G. (2018). Split Questionnaire Designs: collecting only the data that you need through MCAR and MAR designs. *Journal of Applied Statistics*, 45 (8), 1465-1475.

# Split Questionnaire Designs: Collecting only the data that you need through MCAR and MAR designs

James O. Chipperfield<sup>1,2,4</sup>, Margo L. Barr<sup>3</sup>, David G. Steel<sup>4</sup>

<sup>1</sup> Australian Bureau of Statistics

<sup>2</sup>Corresponding author: James O. Chipperfield, email  
james.chipperfield@abs.gov.au

<sup>3</sup> New South Wales Ministry of Health, Australia

<sup>4</sup> University of Wollongong, Australia

## Abstract

We call a sample design that allows for different patterns, or sets, of data items to be collected from different sample units a Split Questionnaire Design (SQD). SQDs can be used to accommodate constraints on respondent burden and to maximise survey design efficiency, commonly measured by the trade-off between the survey cost and the accuracy of target estimates. This paper examines these issues where the data that are not collected by an SQD can be treated as Missing Completely At Random or Missing At Random, targets are regression coefficients in a generalised linear model fitted to binary variables, and targets are estimated using Maximum Likelihood. A key finding is that some respondents may contribute relatively little to the information about regression coefficients; consequently, collecting all data items from these respondents can not only be inefficient but may also impose unnecessary burden. It is better to concentrate the data collection on respondents who make a large contribution to the information matrix. This paper illustrates how to exploit this key finding through an SQD, using Australia's NSW Population Health Survey.

**Key words:** sample design, missing data, multi-matrix sampling

## 1 What is a Split Questionnaire Design?

Consider a survey which collects information from respondents on  $M$  questionnaire modules, where the  $m$ th module collects the  $K_m$  data items denoted by

$\mathbf{y}_m = (y_{m1}, \dots, y_{mk}, \dots, y_{mK_m})'$ ,  $k = 1, \dots, K_m$  and  $m = 1, \dots, M$ . We will call a

sample design that allows for different patterns, or sets, of modules to be collected from different sample units a Split Questionnaire Design (SQD). In a survey that collects data from  $M$  modules, an SQD in theory allows the use of all  $J = \sum_{p=1}^M {}^M C_p$  different combinations in which data on the  $M$  different modules can be collected.

However, in many situations only a relatively small number of different patterns may be used for practical reasons (e.g. form design) and for tractability of estimation.

The sample allocation for an SQD is defined by  $\mathbf{n} = (n^{(1)}, n^{(2)}, \dots, n^{(2)}, \dots, n^{(J)})'$ , where  $n^{(j)}$  is the number of sample units from which the  $j$  th pattern (or combination) of modules are collected. For example, when  $M=3$  the entries in Table 1 show the 7 different patterns available to an SQD, where  $j = 1$  indicates the pattern where only  $\mathbf{y}_1$  is collected from  $n^{(1)}$  sample units.

Historically SQDs were used to reduce response burden and increase response rates (see Munger & Lloyd, 1988). In recent times there has been considerable research into SQDs, much of which has been driven by contemporary realities facing many statistical organisations. These include: increasing non-response rates; increasing demand for more information to be collected as analysts become more sophisticated; tight budget or cost constraints; and variables may be very expensive to collect and intrusive (e.g. require medical procedures).

Some authors fix the allocation,  $\mathbf{n}$ , and consider estimating the mean (see Renssen & Nieuwenbroek, 1997, Merkouris, 2004, Wretman, 1994, and Gelman, King, & Liu, 1998). Thomas, Raghunathan, Schenker, Katzoff, et Johnson (2006) consider forming patterns, where those data items belonging to a pattern are predictive of those data items that do not belong to the pattern. Also, Gonzales et Eltinge (2008) consider the relative efficiency of allocations that follow a monotonic pattern.

Very little work in the literature on SQDs allows  $\mathbf{n}$  to vary. Chipperfield et Steel (2009) considered the approach of finding the optimal allocation for an SQD by trading-off survey costs against accuracy of population estimates, using Best Linear Unbiased Estimation. The different patterns were allocated randomly to survey respondents such that the data not collected by the SQD were considered to be Miss-

ing Completely At Random (MCAR). Chipperfield et Steel (2012) considered the same issue when survey targets were analytic parameters such as linear regression coefficients. Through simulation, they found that by allowing  $\mathbf{n}$  to vary, substantial gains were possible. The size of the gains depended, amongst other things, on the interaction between the marginal cost of collecting each data item and the accuracy requirements imposed on the target estimates.

This paper expands and improves upon the work of Chipperfield and Steel (2009, 2012) by considering survey targets that are non-linear regression parameters, by allowing data patterns to be allocated to respondents based on their characteristics (e.g. age, sex and diabetes status) such that the data not collected by the SQD are Missing At Random (MAR). This approach is evaluated using data from the NSW Population Health Survey (PHS) rather than simulated data. The key findings in this paper are that while MCAR-SQD designs can be worthwhile, MAR designs can be *extremely* efficient, and finding an efficient MAR-SQD allocation is simple and intuitive. These findings have wide implications for survey designs where analytic targets are important.

It is also worth mentioning that Wacholder, Carroll, Pee, et Gail (1994) considers the impact of a handful of different SQD allocations on the accuracy of estimated regression coefficients. In reporting these impacts for case-control studies, Wacholder et al. (1994) uses empirical findings to make general conclusions about which patterns may be worth considering under an SQD. Here we provide a general insight into this issue by deriving a simple expression that shows how the contribution each respondent makes to the observed information matrix of regression coefficients depends upon the outcome variable,  $y$ . While the focus of this paper is on SQD for

Table 1: SQD Data Patterns for Three Modules ( $K = 3$ )

Data pattern ( $j$ )	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	Sample size
1	X			$n^{(1)}$
2		X		$n^{(2)}$
3	X	X		$n^{(3)}$
4			X	$n^{(4)}$
5		X	X	$n^{(5)}$
6	X		X	$n^{(6)}$
7	X	X	X	$n^{(7)}$

random samples of the population, we do briefly consider cases-control sampling.

Section 2 describes two key aspects of optimal SQD allocation, namely the cost of data collection and the accuracy of estimation. Section 3 introduces the PHS and Section 4 explores the potential efficiency of an SQD in the PHS. Section 5 makes some concluding remarks.

## 2 Framework

### 2.1 The Variance of ML Estimates- random samples from the population

Denote the complete set of data that could be collected from the  $i$ th respondent by  $(y, \mathbf{x}_i)$ , where  $y_i$  is the binary outcome variable,  $\mathbf{x}_i$  is a  $K$  vector of covariates collected from the  $i$ th respondent. With the exception of Section 2.3, we assume that respondents are selected using an unstratified simple random sample. Let  $(y, \mathbf{x}_i)$  for  $i = 1, \dots, n$  define a  $(K + 1)$ -way contingency table with up to  $Q = 2^{K+1}$  cells indexed by  $q = 1, \dots, Q$ . The distribution of the cell counts in the contingency table is assumed to be multinomial with parameter  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_q, \dots, \pi_Q)'$ .

#### 2.1.1 Complete Data

The complete data,  $\mathbf{d}_c$ , arises if all variables are collected from all  $n$  respondents in the sample. Specifically,  $\mathbf{d}_c = (\mathbf{y}, \mathbf{X})$ , where  $\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)'$ , and  $\mathbf{X} =$

$(\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n)'$ . Here the regression coefficients,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_r, \dots, \beta_{K-1})'$ , are the design targets. The ML estimate of  $\boldsymbol{\beta}$ , denoted by  $\hat{\boldsymbol{\beta}}$ , based on  $\mathbf{d}_c$  is obtained by solving  $Sc(\boldsymbol{\beta}; \mathbf{d}_c) = \mathbf{0}$  for  $\boldsymbol{\beta}$ , where

$$Sc(\boldsymbol{\beta}; \mathbf{d}_c) = \sum_i \mathbf{x}_i (y_i - \mu_i),$$

and  $\mu_i = f(\mathbf{x}'_i \boldsymbol{\beta}) = E(y_i)$  and  $f^{-1}$  is a link function suitable for predicting binary outcomes.

A common expression for the observed information matrix for  $\boldsymbol{\beta}$  from the sample is  $\mathbf{X}'\mathbf{W}\mathbf{X}$ , where  $\mathbf{W}$  is diagonal with  $i$  th element  $w_i = n\mu_i(1 - \mu_i)$  and  $\mu = f(\mathbf{x}'_i \boldsymbol{\beta})$ . An alternative expression for the observed information matrix on  $\boldsymbol{\beta}$  from the sample is  $\mathbf{I} = \sum_i \mathbf{I}_i$ , where  $\mathbf{I}_i = \sum_i Sc'_i Sc_i$  and  $Sc_i = \mathbf{x}'_i (y_i - \mu_i)$  (see Louis, 1982). Here we use the observed information because it explicitly shows (see section 2.2 for an example) how different values of  $y$  may drive extreme differences in the respondent-level contribution to the observed information.

### 2.1.2 Observed Data

Under an SQD, only  $\mathbf{d}_o$  is collected. The observed data,  $\mathbf{d}_o$ , arises from collecting some subset of  $\mathbf{d}_c$ . The ML estimate of  $\boldsymbol{\beta}$ ,  $\hat{\boldsymbol{\beta}}$ , under  $\mathbf{d}_o$  is obtained by solving

$$E_{\mathbf{d}_c | \mathbf{d}_o} [Sc(\boldsymbol{\beta}; \mathbf{d}_o)] = \mathbf{0} \tag{1}$$

for  $\boldsymbol{\beta}$  (see Rubin et Little (1987)). Breckling, Chambers, Dorfman, Tam, et Welsh (1994) and L. Chambers R., Steel, Wang, et Welsh (2011) show that  $Var(\boldsymbol{\beta}; \mathbf{d}_o) = Info^{-1}(\boldsymbol{\beta}; \mathbf{d}_o)$ , where observed information on  $\boldsymbol{\beta}$  from  $\mathbf{d}_o$  is

$$Info(\boldsymbol{\beta}; \mathbf{d}_o) = \mathbf{I} - \mathbf{L}, \tag{2}$$

where  $\mathbf{L} = Var_{\mathbf{d}_c | \mathbf{d}_o} [Sc(\boldsymbol{\beta}; \mathbf{d}_o)]$  is the observed information loss due to collecting incomplete data,  $\mathbf{d}_o$ , rather than the complete data,  $\mathbf{d}_c$ . Let  $\mathbf{L} = \sum_i \mathbf{L}_i$ , where  $\mathbf{L}_i$  is

the observed loss of information for the  $i$ th record. Next we develop an expression for  $\mathbf{L}_i$ .

When not collecting all data items from the  $i$  th respondent there is uncertainty about which one of the  $Q$  cells in the contingency table the  $i$  th respondent belongs. Let  $Q_i$  be the set of cells to which respondent  $i$  could belong given the data collected from the  $i$  th respondent,  $\hat{\mu}_q$  be the expected value of  $y$  under the logistic model if respondent  $i$  belongs to the  $q$  th cell and  $p_{iq} = \pi_q / (\sum_{s \in Q_i} \pi_s)$  be the probability that the  $i$  th respondent belongs to the  $q$  th cell. If we assume observations are for different respondents are independent, it follows that  $\mathbf{L}_i = \sum_{q \in Q_i} p_{iq} (Sc_{iq} - E_i)(Sc_{iq} - E_i)'$ , where  $Sc_{iq}$  is the value for  $Sc_i$  if the  $i$  th respondent belongs to the  $q$  cell and  $E_i = \sum_{q \in Q_i} p_{iq} Sc_{iq}$ . If all variables are collected from the  $i$ th respondent then  $\mathbf{L}_i = \mathbf{0}$ , where  $\mathbf{0}$  is a vector of zeros, and if  $y_i$  is not collected then  $\mathbf{L}_i = \mathbf{I}_i$ .

## 2.2 Example of Unstratified Sampling

Consider collected all variables from two respondents  $i$  and  $j$  where  $\mathbf{x}_i = \mathbf{x}_j$ ,  $\mu_i = \mu_j = 0.05$ ,  $y_i = 1$ , and  $y_j = 0$ . The predicted value of 0.05 was chosen as it is the prevalence of diabetes in NSW and would be a reasonable value for  $\mu$  if the model had weak predicted power. It is easy to see that, because  $y_i$  is a greater distance from its predicted value of 0.05 when compared with  $y_j$ , all elements in  $\mathbf{L}_i$  are a factor of 361 ( $= (0.95/0.05)^2$ ) times the corresponding elements in  $\mathbf{L}_j$ . This means that, due to the different values of  $y$  alone, the information collected from respondent  $i$  is 361 times greater than the information collected from respondent  $j$ . Consequently, significantly greater information could be lost if only a subset of  $\mathbf{x}$  were collected from respondent  $i$  compared with respondent  $j$ . This is confirmed in the empirical study, where significant cost savings are possible by collecting only

some elements of  $\mathbf{x}_i$  when  $y_i = 0$ . If we changed the value of 0.05 to 0.01 the ratio of the information collected from respondent  $i$  relative to that collected from respondent  $j$  would become greater still.

### 2.3 Example of Case-Control or Stratified Sampling

In case control studies an objective is to estimate the coefficients  $\beta_r$  for  $r = 1, \dots, K - 1$  in  $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1, \beta_k, \dots, \beta_K)'$ , which is the same as  $\boldsymbol{\beta}$  except that, as indicated, the coefficient corresponding to the intercept term may be different. Instead of selecting an unstratified random sample from the population, as in Section 2.1, a stratified random sample of cases ( $y = 1$ ) and a random sample of controls ( $y = 0$ ) are selected. Sampling cases and controls in this way can be efficient when cases (e.g. diabetes is present) are rare in the population. Here we consider whether, like unstratified random sampling, cases are likely to contribute considerably more to the information about  $\beta_r$  than controls.

Since case-control sampling is informative sampling (see R. L. Chambers & Skinner, 2003), a different approach to estimation is called for. Using Bayes' theorem, Wacholder et al. (1994) consider the complete data likelihood based on the factorisation

$$p(\mathbf{x}|y; \boldsymbol{\gamma}) = p(y|\mathbf{x}; \boldsymbol{\beta}^*)p(\mathbf{x}; \boldsymbol{\zeta}^*)/p(y).$$

where  $p(y|\mathbf{x}; \boldsymbol{\beta}^*)$ ,  $p(y)$  and  $p(\mathbf{x}; \boldsymbol{\zeta}^*)$  are the distributions in the case-control sample. The observed information matrix for the parameters  $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{*'}, \boldsymbol{\zeta}^{*'})'$  can be partitioned as

$$\mathbf{I}^* = \begin{pmatrix} \mathbf{I}_{\boldsymbol{\beta}^* \boldsymbol{\beta}^*} & \mathbf{I}_{\boldsymbol{\beta}^* \boldsymbol{\zeta}^*} \\ \mathbf{I}_{\boldsymbol{\zeta}^* \boldsymbol{\beta}^*} & \mathbf{I}_{\boldsymbol{\zeta}^* \boldsymbol{\zeta}^*} \end{pmatrix},$$

It is easy to show that  $\mathbf{I}_{\beta^* \beta^*} = \sum_i S c_i^*{}' S c_i^*$  where  $S c_i^* = p(\mathbf{x}_i; \zeta^*) \mathbf{x}_i' (y_i - \mu_i^*)$ , where  $\mu_i^* = f(\mathbf{x}_i' \beta^*)$  is the expectation of  $y$  in the case-control sample (not the population). Estimation of the parameters proceeds by maximising the likelihood  $l$ .

Using heuristics we now consider the extent to which diabetes status affects a respondent's contribution to  $\mathbf{I}_{\beta^* \beta^*}$  if there are equal number of cases and controls in the sample and the model has weak predictive power. Again consider two respondents  $i$  and  $j$  where  $\mathbf{x}_i = \mathbf{x}_j$ . While  $\mu_i = \mu_j = 0.05$  in the unstratified random sampling case,  $\mu_j^* \approx 0.5$ , since the prevalence of diabetes in the case-control sample is 50%. The residual term  $(y_i - \mu_i^*)$  will now be about 0.5 whether  $y_i = 0$  or 1. It follows that the two respondents will contribute about the same amount of information to  $\mathbf{I}_{\beta^* \beta^*}$ . In other words, in this simple situation, cases and controls contribute about the same amount to the information matrix.

This has implications for SQD design. It suggest that the information loss from only collecting a subset of data on  $\mathbf{x}$  would be much less for cases (e.g. respondents with diabetes) in stratified case-control sampling than for unstratified sampling. More empirical work is required to understand how to design an SQD for studies stratified by cases.

## 2.4 Cost Function

First consider cost for an MCAR-SQD. Denote  $c_m$  as the average time an interviewer spends collecting the data from module  $m$  and  $\Delta n_m$  as the reduction in the number of respondents from whom module  $m$  is collected. The change in cost is given by

$$\Delta C_{MCAR} = \sum_m c_m \Delta n_m \quad (3)$$

Next we can consider an MAR-SQD, where the probability of collecting the

$m$ th module is allowed to depend upon respondent characteristic  $z$ , where  $z = 1, \dots, p, \dots, P$ . Denote  $c_{mp}$  as the average time an interviewer spends collecting the data in module  $m$  from a respondent with  $z = p$  and denote  $\Delta n_{mp}$  as the reduction in the number of respondents with  $z = p$  from whom module  $m$  is collected. The change in cost is given by

$$\Delta C_{MAR} = \sum_{m,z} c_{mz} \Delta n_{mz} \quad (4)$$

There would be no difference between (3) and (4) if the time to collect each module was constant across all values of  $z$ .

### 3 The New South Wales Population Health Survey

The PHS aims to provide detailed information on the health of people living in the Australian state of NSW to support planning, implementation and evaluation of health services and programs (see Barr, Gorringer, & Fritsche, 2005). In 2009, the PHS sample size was 12,000. Though the PHS is designed to meet accuracy targets for annual population estimates of key health variables and risk factors, it is used extensively for multi-variate analysis. New questions are added or removed from the PHS each year according to changing stakeholder priorities and the sample size is increased or decreased according to funding levels. The PHS has a two stage design: the first stage is a random sample of household telephone numbers and the second stage is a random sample of one person per household.

In 2009, the PHS was made up of 43 modules collected using computer-assisted telephone interviewing. Each module is constructed so that it is stand-alone, making it feasible to allocate individual modules to respondents. Here we use the PHS to

simulate the potential costs savings by using an SQD.

### 3.1 Design Targets

To illustrate the ideas in this paper, we consider the situation which requires reducing the number of respondents that can be asked one or more of the Alcohol, Nutrition, Weight and Smoke modules. This will mean one or more of the ALCOHOL, VEG, OVERWEIGHT and SMOKE variables (see Table 2) will be collected from fewer respondents, but the overall number of respondents will remain unchanged. Consequently, there is concern about how this will affect analysis of the association between these variables and DIABETES.

For a survey with lots of data items and many analysts, an important problem here is to specify survey targets. This problem is usually addressed by choosing a small number of key targets. With this in mind, consider a logistic model with DIABETES as the binary outcome variable and the following binary covariates: AGE 54-65, AGE 66+, SEX, moderate alcohol consumption (ALCOHOL), adequate vegetable consumption (VEG), moderate smoker (SMOKE), and overweight (OVERWEIGHT). The design targets of interest here include one or more of the regression coefficients for ALCOHOL, VEG, SMOKE and OVERWEIGHT. The variables in the DIABETES model defines a contingency table with  $Q = 163$ .

### 3.2 The Cost Model

Recall that we are considering a scenario in which the PHS reduces cost by reducing the number of respondents that can be asked up to four modules. For this scenario, we define cost in terms of the total time spent by interviewers undertaking data collection activities, as it would largely explain the change in the marginal monetary

cost of the survey. However, for the PHS, we see from Tables 2 and 3 that older respondents have longer interview times than younger respondents.

Table 2 shows that across the modules, the average interview times for respondents with age= 66+ are between 10 - 50% longer than for respondents with age=0-54. For example, Table 2 shows that the average interview time of 1.72 ( $c_{41}$ ) for respondents with age=0-54 is 20% lower than the average interview time for respondents with age = 66+ ( $c_{43} = 2.16$ ). The time taken and cost of collecting the Nutrition, Smoke and Alcohol modules from respondents with age= 66+ is 27% longer than for respondents with age=0-54.

Table 3 shows the distribution of interview times for all respondents, and by age group. It shows that the average, median and maximum interview times were 26, 25 and 70 minutes, respectively. Interestingly, the distribution of interview times for respondents with age= 66+ has a much longer upwards tail than the other age groups.

From the perspective of simply reducing cost or respondent burden, it clearly follows that the amount of data collected from older respondents should be minimised.

## 4 SQD and the PHS

This section uses the PHS to examine the potential of an MCAR-SQD and an MAR-SQD to minimise the reduction in accuracy of the estimated target regression coefficients, given one or more of the ALCOHOL, VEG, OVERWEIGHT and SMOKE variables will be collected from fewer respondents.

Table 2: Average Interview Times (minutes) for Select Modules of the NSW Population Health Survey

Module ( $m$ )	Module (Variables Collected)	Average Interview Time by Characteristic			
		All $c_m$	0-54 $c_{m1}$	54-65 $c_{m2}$	66+ $c_{m3}$
1	Demographic (AGE 54-65, AGE 66+, SEX)	0.61	0.59	0.56	0.66
2	Diabetes and Blood Pressure (DIABETES)	0.23	0.18	0.25	0.28
3	Weight (OVERWEIGHT)	0.54	0.48	0.55	0.64
4	Nutrition (VEG )	1.91	1.72	1.97	2.16
5	Alcohol (ALCOHOL)	0.45	0.40	0.51	0.49
6	Smoke (SMOKE)	0.38	0.32	0.40	0.46

Table 3: Distribution of Interview times (minutes) for NSW Population Health Survey

Age Group (yrs)	Average	Min	25%	50%	75%	Max
-	26	11	22	25	29	70
0-19	25	11	21	24	28	47
20-53	25	13	21	24	28	47
54-65	27	14	21	24	28	58
66+	29	13	23	27	32	70

## 4.1 MCAR Design

Consider the reduction in the accuracy of ML-estimated regression coefficients from (1) when not collecting ALCO, VEG and SMOKE from 100 respondents, selected completely at random from all PHS respondents. The accuracy of ML-estimated coefficients is obtained from (2). The reduction in the accuracy of an estimate can be expressed as the reduction in sample size that would be required to achieve the same reduction in accuracy. The reduction in the effective sample size for each regression coefficient is shown in Table 4 for each possible combination of variables that are not collected. Table 4 shows that not collecting ALCO, VEG and SMOKE from the 100 respondents and reducing the PHS sample size by 98 respondents would both reduce the accuracy of the ML-estimated coefficient (obtained from (2)) for SMOKE by the same amount. It is clear from Table 4 that when a variable is not collected almost all information about its corresponding model coefficient is lost. Conversely, when a variable is collected almost no information about its corresponding coefficient is lost- this is almost independent of which other covariates are collected (e.g. across all patterns where ALCOHOL is not collected, the reduction in the effective sample size for the estimate of the ALCOHOL coefficient is within the narrow range of 96 to 100).

If estimates used only respondents from whom all variables were collected, the so called *complete cases*, all figures in Table 4 would be 100. The benefit of instead using ML estimation, which uses all respondents, is apparent when the reduction in effective sample size is less than 100.

It is also worthwhile noting that under an MCAR-SQD, it is not possible to disproportionately reduce the amount of data collected from respondents with age=66+.

Table 4: Reduction in Effective Sample Size for Select Regression Coefficients, when the data that are not collected from 100 Respondents are MCAR

Variables not collected	ALCOHOL	VEG	SMOKE	OVER-WEIGHT
OVERWEIGHT	2	2	2	96
SMOKE	1	0	93	0
VEG	0	96	0	0
ALCOHOL	96	0	2	0
ALCOHOL, VEG	98	97	2	0
VEG, SMOKE	1	97	95	0
OVERWEIGHT , SMOKE	3	2	96	98
ALCOHOL, SMOKE	100	0	99	1
OVERWEIGHT , ALCOHOL	98	2	3	98
OVERWEIGHT , VEG	1	97	2	97
OVERWEIGHT , VEG, SMOKE	3	98	97	99
OVERWEIGHT , ALCOHOL, SMOKE	100	2	100	100
OVERWEIGHT , ALCOHOL, VEG	100	98	4	99
ALCOHOL, VEG, SMOKE	100	98	98	1

## 4.2 MAR Design

Here we consider the expected reduction in the accuracy of ML-estimated regression coefficients when not collecting ALCO, VEG and SMOKE from 100 respondents which are randomly selected from a sub-group defined in terms of age group, sex and/or DIABETES. In practice, this would require age, sex and DIABETES to be collected *before* a decision is made about which of the Alcohol, Nutrition, Weight and Smoke modules to collect. Modern computer-assisted collection methods make this approach feasible.

Table 5 shows that not collecting ALCO, VEG and SMOKE from 100 randomly sub-sampled respondents who are male, with age= 0-53 and DIABETES=No *or* decreasing the PHS sample size by 8 respondents, reduce the accuracy of the ML estimate of the coefficient for SMOKE by the same amount. This figure of 8 increases significantly to 1463 if the 100 respondents instead had DIABETES=Yes.

In other words, not collecting ALCO, VEG and SMOKE from one respondent with DIABETES=Yes or 183 (=1463/8) respondents with DIABETES=No, reduce the accuracy of the estimated coefficient for SMOKE by the same amount. While both have the same impact on accuracy, clearly the latter would result in significantly greater cost savings.

Age has a significant impact on the effective sample size. If we consider the above example with age = 66+ instead of age = 0-53, not collecting ALCO, VEG and SMOKE from one respondent with DIABETES=Yes or 5 (=176/32) respondents with DIABETES=No, reduce the accuracy of the estimated coefficient for SMOKE by the same amount. While both have the same impact on accuracy, clearly the latter would result in greater cost savings, though not as significant as for the above case where age=0-53.

Table 5 shows that not collecting ALCO, VEG and SMOKE from respondents who are female with age= 0-53 and DIABETES=No is the optimal MAR-SQD approach - this does not change despite the fact that the cost of collecting all these variables is 27% lower compared with respondents with age=66+.

For the case where ALCO, VEG and SMOKE are not collected, the MAR-SQDs are always more efficient than the MCAR-SQD (see Table 4), as long as the MAR-SQD collects all variables from respondents with DIABETES=Yes. Namely, the optimal MAR-SQD reduces the effective sample size for the ALCO, VEG and SMOKE coefficients are by 2, 2 and 3, respectively- considerably smaller than 100, 98 and 98, respectively, for the MCAR-SQD.

The conclusion that an MAR-SQD is more efficient than an MCAR-SQD is also true across all possible missing patterns in OVERWEIGHT, ALCO, VEG and

Table 5: Reduction in Effective Sample Size for Select Regression Coefficients when not Collecting ALCOHOL, VEG and SMOKE on 100 Respondents by Sex, Age, and Diabetes

SEX	AGE	DIABETES	ALCOHOL	VEG	SMOKE	OVERWEIGHT
		No	15	17	15	0.2
		Yes	503	650	744	2
Male	0-53	No	7	2	8	0.1
Male	54-65	No	54	29	48	0.7
Male	66+	No	62	39	32	1
Male	0-53	Yes	1452	423	1463	5
Male	54-65	Yes	1195	342	975	4
Male	66+	Yes	733	472	176	1
Female	0-53	No	2	2	3	0.1
Female	54-65	No	8	24	17	0.2
Female	66+	No	4	33	13	0.1
Female	0-53	Yes	1117	737	1774	4
Female	54-65	Yes	633	1140	851	2.6
Female	66+	Yes	176	929	396	1.2

SMOKE, not just the pattern considered in Table 5 for illustration. This was true as long as the MAR-SQD collects all variables from respondents with DIABETES=Yes (results not provided).

Alternatively, we could have considered an MAR-SQD where only DIABETES is used to decide which of the four modules are collected from a respondent. Table 5 shows that if we did not collect ALCO, VEG and SMOKE from 100 randomly selected respondents with DIABETES=No, the reduction in the effective sample size for the estimated coefficients for ALCO, VEG and SMOKE would be 15, 17 and 15, respectively- slightly higher than for the optimal MAR-SQD.

## 5 Summary

This paper shows theoretically and empirically that respondents with an outcome variable that is further from its expected values (e.g. diabetes is present), condi-

tional on  $\mathbf{x}$  and the model, contribute significantly more information than other respondents. As a consequence, collecting only some covariates from respondents in the latter case, would only have a relatively small impact on the accuracy of estimates regression coefficients. This important finding should be kept in mind when designing a survey for analytic parameters.

### Acknowledgements

Thanks to the NSW Ministry of Health for access to the 2009 NSW Population Health Survey data and the Australian Bureau of Statistics for part-funding this research.

### Références

- Barr, M. L., Gorringer, D., & Fritsche, L. (2005). Nsw population survey: Description of methods. *Centre for Epidemiology and Research, NSW Department of Health*.
- Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M., & Welsh, A. H. (1994). Maximum likelihood inference from sample survey data. *Internatoinal Statistical Review*, 62, 349-63.
- Chambers, L., R., Steel, D. G., Wang, S., & Welsh, A. (2011). *Maximum likelihood estimation for sample surveys*. CRC Press Taylor and Francis Group, United States of America.
- Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. John Wiley and Sons.
- Chipperfield, J. O., & Steel, D. G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 227-244.
- Chipperfield, J. O., & Steel, D. G. (2012). Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference*, 141, 1925-1933.
- Gelman, A., King, G., & Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93, 846-857.
- Gonzales, J. M., & Eltinge, J. L. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. *Proceedings of the Joint Statistical Meeting*.
- Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society*, 2, 226-233.

- Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, *99*, 1131-1139.
- Munger, G., & Lloyd, B. H. (1988). The use of multiple matrix sampling for survey research. *Journal of Experimental Education*(56), 187-191.
- Renssen, R. H., & Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more surveys. *Journal of the American Statistical Association*(92), 368-374.
- Rubin, D. B., & Little, R. J. A. (1987). *Statistical analysis of missing data*. John Wiley and Sons.
- Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the national health and nutrition examination survey. *Survey Methodology*, *32*, 217-231.
- Wacholder, S., Carroll, R. J., Pee, D., & Gail, H., M. (1994). The partial questionnaire design for case control studies. *Statistics in Medicine*, *13*, 623-634.
- Wretman, J. (1994). Estimation in sample surveys with split questionnaires. *Research Report, University of Stockholm*, *3*, 1-11.