Split Questionnaire Designs: are they an efficient design choice?

James O. Chipperfield [1,2], Margo Barr[3], David G. Steel [4]
[1] Australian Bureau of Statistics
[2]Corresponding author: James O. Chipperfield, email james.chipperfield@abs.gov.au
[3] Centre for Epidemiology and Evidence, Australia
[4] University of Wollongong, Australia

Abstract

Abstract

Split Questionnaire Designs (SQD) have been historically used to accommodate constraints on respondent burden. This paper discusses how an SQD can be an efficient design option in practice to give more flexibility in meeting the range of objectives of a survey and focuses on the NSW Population Health Survey, which has previously used SQDs. The efficiency of a design can be measured by the cost required to meet constraints on the accuracy of estimates. The targets of interest to the design are analytic parameters, such as regression coefficients. It is assumed that these targets of interest are estimated using Maximum Likelihood methods.

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 $\mathbf{y}_m = (y_{m1}, \ldots, y_{mk}, \ldots, y_{mK_m})'$, $k = 1, \ldots, K_m$ and $m = 1, \ldots, M$. 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 information from M modules, an SQD allows the use of all $J = \sum_{p=1}^{M} {}^{M}C_p$ different combinations in which information on the M different modules can be collected. The sample allocation for an SQD is defined by $\mathbf{n} = (n^{(1)}, n^{(2)}, \ldots, n^{(2)}, \ldots, n^{(J)})'$, 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.

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; and tight budget or cost constraints.

Some authors fix the allocation, $\mathbf{n}$, and consider estiimation issues (see for example Renssen and Nieuwenbroek (1997) and Merkouris (2004)). Thomas et. al (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. Gonzales and Eltinge (2008) consider the relative efficiency of alternative allocations for which multi-phase estimation is suitable. Chipperfield and Steel (2009, 2012) considered the approach of finding the optimal allocation for an SQD by trading-off survey against the surveys target estimates.

Section 2 considers how the theory of SQDs can be applied to the NSW Population Health Survey (PHS), a major survey, within the framework of Chipperfield and Steel (2009, 2012). Section 3 outlines future work.

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

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

2 The New South Wales Health Survey

The PHS aims to provide detailed information on the health of people living in the Australian state of NSW and its health regions to support planning, implementation and evaluation of health services and programs (see Barr et. al, 2005). In 2009, the annual PHS sample size is 12, 000 at the NSW level and 1,500 at each of its health regions. Though the PHS is used extensively for multi-variate analysis, it is designed to meet accuracy targets for annual population estimates of key health variables and risk factors for each health region and for NSW. New questions may be added to the PHS each year in response to stakeholders priorities. The PHS has a two stage design: the first stage is a random sample of telephone numbers within a health region and the second stage is a random sample of one person per household. The PHS estimates are model-assisted under a post-stratified model with age and sex covariates.

In 2009, the PHS was made up of 43 modules collected using computer-assisted telephone interviewing (CATI). There is a core set of 31 modules which are asked of all respondents. There is an additional set of $M = 12$ *SQD modules*, nine of which are selected at random for each respondent. Therefore there are $J = {}^{12}C_9 = 220$ patterns or combinations of modules, where the expected allocation for pattern j is $n_{PHS}^{(j)} = n/J$. The PHS's SQD design allowed an increase in the total number of modules collected while controlling respondent burden.

The rest of this section discusses features of the PHS that are important when considering an optimal allocation for the SQD modules.

2.1 The Distribution of the Data

All data collected by the PHS are categorical. Some questions within a module may or not be answered, due to sequencing. Define the $T = \Sigma_m k_m$ vector of variables $\mathbf{y} = (y_{11}, \ldots, y_{mk_m}, \ldots, y_{Mk_M})'$ for the i th person by $\mathbf{y}_i$, where y_{mk_m} has l_{km} levels such that $\mathbf{y}_i$ for $i = 1, \ldots, n$ defines a K-way contingency table with $Q = \Pi_k \Pi_m l_{km}$ cells and $\mathbf{y}_i$ and $\mathbf{y}_{i'}$ are independent for $i \neq i'$. We define $\mathbf{W}_i = (W_{i1}, \ldots, W_{iq}, \ldots, W_{iQ})'$ to be a Q x 1 vector where $W_{iq} = 1$ if unit i belongs to the q th cell of a contingency table and $W_{iq} = 0$ otherwise, where $q = 1, \ldots Q$. The distribution of the cell counts in the contingency table is assumed to be multinomial with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_q, \ldots, \pi_Q)'$.

2.2 The Cost Model for Data Collection

Here we are only interested in how the cost of the PHS changes in response to a change in the allocation, $\mathbf{n}$. In particular, we define cost in terms of time (e.g. hours) spent by interviewers undertaking data collection activities, as it would largely explain the change in the monetary cost of the survey due to a change in allocation.

Some costs do not vary with the allocation. Examples includes the cost of developing the questionnaire, field testing supervision, and purchasing the technological infrastructure to process the survey data. Such costs are ignored here as they would not affect the optimal allocation here.

We now describe the two parameters in our cost model. The parameter c_0 is the average time an interviewer spends per respondent before any of the information about the SQD modules are collected. For the PHS this would include the time an interviewer spends waiting for calls to be answered and the time taken to introduce the survey, and would be a function of the proportion of answered calls that end in a refusal. Up to 7 calls are made to establish initial contact with a household and up to 5 calls are made to make contact with a selected person within the household. In the Health Survey, the average time taken to introduce the survey is 30 seconds, the average time an interviewer spends waiting for calls to be answered is 20 seconds and the refusal rate is 40%. Hence 100 (=[30 + 30] / {1-0.40}) seconds is a crude estimate of the average time spent by interviewers prior to collecting any information from a respondent. In addition, interviewers spend on average 17 minutes collecting information on the 31 core modules. Therefore we could calculate c_0 to be almost 19 minutes.

The parameter $c^{(j)}$ is the average time an interviewer spends collecting data from the SQD modules assigned to pattern j. If we denote c_m as the average time an interviewer spends collecting $\mathbf{y}_m$ and denote s_j to be the set of modules allocated to pattern j, then we may write $c^{(j)} = \Sigma_{m \in s_j} c_m$.

The cost, C, measured in interviewers' time, of an SQD with allocation $\mathbf{n}$ is

$$C \quad = c_0 n + \textstyle\sum_j c^{(j)} n^{(j)} \tag{1}$$

Table 2 shows the distribution of interview times for all respondents, for respondents aged 66 +, and for respondents in different age and sex categories. The interview times are based on 1800 interviews conducted for the 2009 PHS. It shows that the average, median and maximum interview times were 26, 25 and 70 minutes, respectively. The interview times for females under 66 years of age tend to have the same distribution regardless of age. This is also the case for males under 66 years of age, who tend to have slightly shorter interview times than females under 66+. Interestingly, males and females in the 66+ age group have significantly longer average interview times than the other age-sex categories.

The average interview time across the 12 SQD modules varied widely. The Air Polution module, which was only applicable for respondents living in the Sydney metropolitan, Illawarra and Hunter region, took only 2 second on average: this was because when the module was not applicable it required zero seconds of interviewer time. SQD modules with the longest interviewer times include: Nutrition (N) with 13 questions and average interview time of 2.8 minutes (3.1 minutes for 66+), Social Capital (SC) with average interviewer time of 2.7 minutes (3.6 minutes for 66+) and Oral Health (OH) with average interviewer time of 2.0 minutes (2.2 for 66+).

The average interview time to complete a random sample of 9 SQD modules was 8.1 minutes (9.6 minutes for 66+). If instead all 12 SQD questions were asked the average

Table 2: Distribution of Interview times (minutes) for NSW Population Health Survey

Sex	Age	Average	Min	25%	50%	75%	Max
-	-	26	11	22	25	29	70
-	66+	29	13	23	27	32	70
Females	0-19	25	11	21	25	29	43
Females	20-53	25	15	21	24	28	45
Females	54-65	26	16	22	25	29	44
Females	66+	29	17	24	28	33	70
Males	0-19	24	13	20	23	27	47
Males	20-53	24	13	20	23	27	47
Males	54-65	27	14	20	23	27	58
Males	66+	28	13	23	27	32	69

interview time would increase to 10.9 minutes (12.9 for 66+). Therefore asking 9 instead of 12 of the modules saves on average 1.8 (2.3 minutes for 66+) minutes per respondent.

2.3 Design Targets

For a survey with hundreds of data items and many potential analysts it is difficult to specify a small set of design targets. For an organisation funding one PHS SQD module with a small number of data items, this would be less difficult. This is not a new problem and is usually addressed by choosing a small number of key design targets of interest.

To this end, consider if we define a single categorical variable for each of the 12 SQD modules, called a *module variable*, which aims to contain key information collected by the module. For example, the Nutrition module, with 13 questions, could be summarised into a *Nutrition variable* with categories: "Eat fast food 2 times a week or less and eats red meat at least three times a week", "Eat fast food 2 times a week or less and does not eat red meat at least three times a week" and "other". Similarly, the Social Capital variable could have categories: "participates in sporting activities and feels safe at night", "participates in sporting activities and does not feel safe at night" and "other" otherwise. And the Oral health module variable could have categories: "All natural teeth and visited a denitist in the last 12 months", "All natural teeth and has not visited a denitist in the last 12 months" and "other". The design targets, whether marginal proportions or regression coefficients, could be defined in terms of the multinomial distribution with 27 parameters (3x3x3).

2.4 The Variance of the Estimates of Design Targets

The complete data, $\mathbf{d}_c$, is collected if all T data items are collected from all n respondents. The observed data, $\mathbf{d}_o$, arises from not collecting all T data items from all n respondents. Under an SQD, $\mathbf{d}_o$ is collected.

Below the parameter for the multinomial distribution (i.e. proportions) and regression coefficients as considered as design targets. The variance of estimates of the design targets depend upon $\mathbf{d}_o$ or, equivalently, $\mathbf{n}$. Historical survey data is very important to evaluate these expressions for the variance at the design stage.

2.4.1 Proportions

Let $\hat{\boldsymbol{\pi}}$ be the ML estimate of $\boldsymbol{\pi}$ from $\mathbf{d}_o$ (see Rubin & Little, 2002 for details). Chipperfield et Steel (2012) gives an expression for $Var(\hat{\boldsymbol{\pi}}; \mathbf{d}_o)) = Info^{-1}(\boldsymbol{\pi}; \mathbf{d}_o)$, where $Info(\boldsymbol{\pi}; \mathbf{d}_o)$ is the information on $\boldsymbol{\pi}$ from the observed data $\mathbf{d}_o$. The ML estimator of the number of people in the sample belonging to each of the Q cells is $\hat{\mathbf{r}} = n\hat{\boldsymbol{\pi}}$.

2.4.2 Regression Coefficients

Consider fitting a regression model to the counts $\hat{\mathbf{r}}$ to obtain the ML estimate of a regression parameter, $\boldsymbol{\beta}$, based on $\mathbf{d}_o$. This involves solving $Sc(\boldsymbol{\beta}; \mathbf{d}_o) = \mathbf{0}$ where

$$Sc(\boldsymbol{\beta}\mathbf{d}_o) = \mathbf{X}[\hat{\mathbf{r}}_s - diag(\hat{\mathbf{r}}_s + \hat{\mathbf{r}}_f)\boldsymbol{\mu}],$$

$\hat{\mathbf{r}} = (\hat{\mathbf{r}}_s, \hat{\mathbf{r}}_f)$, $\hat{\mathbf{r}}_s = (r_{sl})$ and $\hat{\mathbf{r}}_f = (\hat{r}_{fl})$ are column vectors of length L, $\hat{r}_{sl}$ and $\hat{r}_{fl}$ are respectively the number of successes and failures conditional on the the l th covariate pattern $\mathbf{x}'_l$, $\mathbf{X} = (\mathbf{x}'_1, , \ldots, \mathbf{x}'_l, \ldots, \mathbf{x}'_L)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_l, \ldots, \mu_L)'$, $\mu_l = f(\mathbf{x}'_l\boldsymbol{\beta})$ and μ is the link function. From Breckling, Chambers, Dorfman, Tam, et Welsh (1994), the $Var(\boldsymbol{\beta}; \mathbf{d}_o) = Info^{-1}(\boldsymbol{\beta}; \mathbf{d}_o)$, where information on $\boldsymbol{\beta}$ from $\mathbf{d}_o$ is

$$Info(\boldsymbol{\beta}; \mathbf{d}_o) = Info(\boldsymbol{\beta}; \mathbf{d}_c) - Var\Big[Sc(\boldsymbol{\beta}; \mathbf{d}_o)\Big]$$

where $Info(\boldsymbol{\beta}; \mathbf{d}_c) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})'$ is the information that would be available from the complete data $\mathbf{d}_c$, $\hat{\mathbf{W}}$ is diagonal with l th element $\hat{w}_l = n\hat{\mu}_l(1 - \hat{\mu})$, $\hat{\mu} = f(\mathbf{x}'_l\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$,

$$Var\Big[Sc(\boldsymbol{\beta}; \mathbf{d}_c)\Big] \quad = \mathbf{X}'Var\Big[diag(\mathbf{1}_L - \boldsymbol{\mu})\hat{\mathbf{r}}_s - diag(\hat{\mathbf{r}}_\mathbf{f})\boldsymbol{\mu}\Big]\mathbf{X} \tag{2}$$

and

$$\begin{aligned} &Var\Big[diag(\mathbf{1}_L - \boldsymbol{\mu})\hat{\mathbf{r}}_s - diag(\hat{\mathbf{r}}_f)\boldsymbol{\mu}\Big] \\ &\quad = diag(\mathbf{1}_L - \boldsymbol{\mu})Var[\mathbf{y}]diag(\mathbf{1}_L - \boldsymbol{\mu}] + diag(\boldsymbol{\mu})Var\Big[\hat{\mathbf{r}}_\mathbf{f}\Big]diag(\boldsymbol{\mu}) \\ &\quad\quad -2diag(\mathbf{1}_L)Cov[(\hat{\mathbf{r}}_s, \hat{\mathbf{r}}_f)diag(\boldsymbol{\mu}] \end{aligned} \tag{3}$$

These terms in the above equation can be evaluated using $Var(\hat{\boldsymbol{\pi}}; \mathbf{d}_o))$.

2.5 Assigning Patterns to Respondents

For a sample size of 12,000, the expected allocation for pattern j, under the current approach of randomly allocating 9 of the 12 SQD modules to a respondent, is $n^{(j)}_{PHS} \approx 55$. While the PHS SQD allocation is simple it has some note worthy features. First, by collecting all possible patterns with 9 modules it collects all interactions between the SQD modules up to order 9 and uses approximately the same sample size for each combination. Second, the data not collected by the SQD modules are missing Completely At Random (MCAR) (see Rubin & Little, 2002) and means that inferences using the *available cases*, a simple and popular way of dealing with non-response, are valid. It is possible by collecting particular combinations of modules more than others and collecting some modules more than others, that the cost, the accuracy of estimates and the respondent burden could be more effectively managed.

From the outset it is useful to consider removing some patterns from the design. First, some patterns could be excluded from the design on the basis that they lead to unacceptablely long average interview times. For the 66+ we could consider discarding the patterns which include all of the OH, N and SC modules. Since a 66+ respondent would only be allocated at most two of these modules, the average intervew time would reduce from about 27 to under 26 minutes. Excluding patterns based on the age of the respondent would mean the data not collected are Missing At Random (MAR) (see Rubin & Little, 2002).

Second, if joint analysis of variables collected in two different modules is an important to the design, then it could be desirable to always collect these modules from the same respondent.

Third, other patterns can be ruled out on the basis that they are inefficient for the purposes of the design and so unlikely to feature in the optimal allocation. This is particularly important in order to ensure the search for the optimal allocation is computationally feasible. However, for the PHS situation where J is relatively small, this is not likely to be a concern.

3 Summary and Future Work

This paper identifies some potential areas of development to the NSWs Population Health Survey's SQD design. Changing the SQD allocation could improve the accuracy of design targets for fixed cost and reduce respondent burden for older respondents. In future work we will consider the issues in section 2 in more detail and aim to develop a practical and efficient alternative to the current SQD design for the NSW Health Survey.

Références

Barr, M., Gorringe, 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.

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.

Gonzales, J. M., & Eltinge, J. L. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. *Proceedings of the Joint Statistical Meeting*.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, *99*, 1131-1139.

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. (2002). *Statistical analysis of missing data, 2nd edition*. 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.