2009

Design and estimation for split questionnaire surveys

James O. Chipperfield
Australian Bureau of Statistics

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Design and estimation for split questionnaire surveys

Abstract

When sampling from a finite population to estimate the means or totals of *K* population characteristics of interest, survey designs typically impose the constraint that information on *all K* characteristics (or data items) is collected from *all* units in the sample. Relaxing this constraint means that information on a subset of the *K* data items may be collected from any given unit in the sample. Such a design, called a split questionnaire design (SQD), has three advantages over the typical design: increased efficiency with which design objectives can be met, by allowing the number of sample units from which information on a particular data item is collected to vary; improved efficiency in estimation through exploiting the correlation between the *K* data items; and flexibility to restrict the maximum number of data items collected from a unit to be less than *K*. An SQD can be viewed as designing for the missing pattern of data. This article considers sevaral estimators, including the Best Linear Unbiased Estimator (BLUE), for an SQD. The results show that significant gains can be achieved. The size of the SQD gains depends upon the function describing the survey costs, the design constraints, and the covariance matrix of the data items of interest. These methods are evaluated in a simulation study with four data items.

Keywords

era2015

Disciplines

Physical Sciences and Mathematics

Publication Details

Chipperfield, J. O. & Steel, D. G. (2009). Design and estimation for split questionnaire surveys. Journal of Official Statistics: an international quarterly, 25 (2), 227-244.

Journal of Official Statistics, Vol. 25, No. 2, 2009, pp. 227–244

Design and Estimation for Split Questionnaire Surveys

James O. Chipperfield[1] *and David G. Steel*[2]

When sampling from a finite population to estimate the means or totals of K population characteristics of interest, survey designs typically impose the constraint that information on *all* K characteristics (or data items) is collected from *all* units in the sample. Relaxing this constraint means that information on a subset of the K data items may be collected from any given unit in the sample. Such a design, called a split questionnaire design (SQD), has three advantages over the typical design: increased efficiency with which design objectives can be met, by allowing the number of sample units from which information on a particular data item is collected to vary; improved efficiency in estimation through exploiting the correlation between the K data items; and flexibility to restrict the maximum number of data items collected from a unit to be less than K. An SQD can be viewed as designing for the missing pattern of data. This article considers sevaral estimators, including the Best Linear Unbiased Estimator (BLUE), for an SQD. The results show that significant gains can be achieved. The size of the SQD gains depends upon the function describing the survey costs, the design constraints, and the covariance matrix of the data items of interest. These methods are evaluated in a simulation study with four data items.

Key words: Missing data; multi-phase; optimal sample design; cost function.

1. Introduction

Consider a large population of N units with K data items or characteristics of interest with population totals $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)'$, where $Y_k = \sum_{i=1}^{N} Y_{ki}$ and Y_{ki} is the kth data item corresponding to the ith population unit, and with a known covariance structure, $S_{kk'} = 1/(N-1)\sum_{i=1}^{N}(Y_{ki} - \bar{Y}_k)(Y_{k'i} - \bar{Y}_{k'})$.

With few exceptions, survey designs involve collecting information on *all* K data items from a sample of n units selected from the population; the information collected is denoted by $\boldsymbol{y}_i = (y_{1i}, y_{2i}, \ldots, y_{Ki})'$, where $i = 1, 2, \ldots, n$. This single-phase constraint leads to simplicity in the survey design and estimation and the requirement that only one questionnaire is developed, pilot tested and, perhaps, printed.

The sample size necessary to obtain the required precision for the estimation of the population total may differ for each data item but, because of the single-phase constraint, the same sample size is often used for all data items. Imposing the single-phase constraint results in suboptimal outcomes. Relaxing the single-phase constraint allows information

[1] Methodology Division, Australian Bureau of Statistics, P O Box 10, Belconnen, ACT 2616, Australia. Email: james.chipperfield@abs.gov.au
[2] University of Wollongong, Centre for Statistical and Survey Methodology, Northfields Avenue, Wollongong NSW 2522, Australia. Email: dsteel@uow.edu.au
Acknowledgments: The authors would like to thank Robert Clark for his constructive suggestions and comments and the Australian Bureau of Statistics for supporting this research.

on data items with relatively low variability and high cost to be collected from fewer units than data items with relatively high variability and low cost. We will call a design that allows for different patterns, or sets, of data items to be collected from different units a Split Questionnaire Design (SQD). An SQD allows the use of all $J = \sum_{p=1}^{K} {}^{K}C_p = 2^K - 1$ different combinations in which the K different data items can be collected.

Some established statistical methods have a relationship with SQDs, but none of these methods enable the full benefits of an SQD to be realised. In a multi-phase design (see Cochran 1977, p. 327 and Särndal, Swensson, and Wretman 1992, p. 343) the possible combinations for which the K data items can be collected, referred to as patterns, are restricted to follow the monotonic pattern shown in Table 1. The Xs in the table indicate the data items collected from each pattern: when y_k is collected, information on $y_{k-1}, y_{k-2}, \ldots, y_1$ is always collected. Any given multi-phase design allows at most K patterns. An SQD is more flexible. This article shows that the benefit of this flexibility can be substantial.

Multiple Matrix Sampling (MMS) (Shoemaker 1973) in theory allows the use of all J patterns and focuses on estimating differences between groups in situations where a single-phase design is impractical or would result in concerns about the quality of responses, due to, say, respondent fatigue. The examples in Shoemaker (1973) do not implement all J patterns and focus on situations where the different data items are measures of the same underlying characteristic. For example, the difference between levels of literacy in schools could be measured by giving a sample of students random subsets of a large number ($K = 500$) of words to spell (see Munger and Lloyd 1988, for an application). Shoemaker (1973) suggests MMS could be useful for sample design when there are no such concerns.

While the single-phase constraint is typically imposed at the design stage, nonresponse means the observed survey data do not follow a single-phase pattern. For example, when $K = 2$ some units may respond to only y_1, or to only y_2, and others may respond to both. Estimators of population totals and their sampling errors in the presence of missing data have been widely discussed in the literature (see Shao and Sitter 1996; Rao 1996; Fay 1996). These methods minimise the variance and bias of estimators of population totals *given* the observed data. They do not exploit benefits from specifying the pattern of observed data.

Renssen and Nieuwenbroek (1997) and Merkouris (2004) suggested methods of improving the consistency and accuracy of estimates from independent surveys that have data items in common. Renssen and Nieuwenbroek (1997) noted the application of their method to SQDs, where the independent surveys can collectively be interpreted as one

Table 1. Multi-phase design patterns

Pattern	y_1	y_2	. . .	y_k
1	X		. . .	
2	X	X	. . .	
⋮			. . .	
K	X	X	. . .	X

survey with independent samples collecting a different subset of the K data items. Also, Wretman (1994) considered estimation using patterns with only two data items, where one of the data items was common to all the patterns. Again, these papers are focused on estimation, not on design issues.

There is a link between SQDs and sampling on two occasions. Sampling on two occasions (Särndal et al. 1992, p. 368) involves collecting information on the *same* data item on two occasions, denoted by y_1 and y_2. One objective of such a sampling scheme is to estimate the current level, say Y_2. In contrast, an SQD involves collecting two data items, y_1 and y_2, that correspond to *different* characteristics on the *same* occasion. Design issues focus on selecting a rotation pattern for panel designs, a characteristic of Labour Force Surveys of many national statistical agencies (see for example Bell 2001; McLaren and Steel 2001).

Raghunathan and Grizzle (1995) and Gelman, King, and Liu (1998) relaxed the single-phase constraint and imputed the unobserved data within a Bayesian framework and estimated variances using multiple imputation (Rubin and Little 1987). However, they do not solve the problem of optimal allocation with multivariate constraints and their results rely on model assumptions. Adiguzel and Wedel (2008) consider optimal splitting of a questionnaire when the number of data items, K, is very large. In this article we consider the problem of optimal allocation for SQDs in a design-based (Särndal et al. 1992, p. 219), rather than model-based, framework. Consequently, the methods used do not rely on model assumptions.

Section 2 introduces several design-based estimators, including the Best Linear Unbiased Estimator (BLUE), of $\mathbf{Y}$ when $K = 2$ for an SQD. Section 3 compares the efficiency of the two-phase, single-phase and split questionnaire designs in an empirical study when $K = 2$. Section 4 extends the BLUE to arbitrary K and compares the efficiency of the SQD with the multi-phase design in a simulation study with $K = 4$. Section 5 summarises the article's findings.

2. Split Questionnaire Design and Estimation for $K = 2$

In the case of $K = 2$ an SQD selects three nonoverlapping Simple Random Samples Without Replacement (SRSWOR) denoted by $s^{(1)}$, $s^{(2)}$, and $s^{(3)}$, of size $n^{(1)}$, $n^{(2)}$, and $n^{(3)}$, that collect information on only y_1, only y_2, and both y_1 and y_2, respectively. The three ways information on the data items can be collected are denoted by Patterns 1, 2, and 3, respectively, and are illustrated in Table 2. We define the fixed cost per sample unit as c_0, which is independent of the cost of collecting the information about y_1 or y_2. The marginal cost of collecting the data items from pattern j is denoted by $c^{(j)}$. We make the reasonable

Table 2. SQD Patterns

Pattern	y_1	y_2	Sample size	Marginal cost
1	X		$n^{(1)}$	$c^{(1)}$
2		X	$n^{(2)}$	$c^{(2)}$
3	X	X	$n^{(3)}$	$c^{(3)}$

assumption that $c^{(3)} = c^{(1)} + c^{(2)}$. The total cost per unit, made up of the fixed and marginal cost, is denoted by $t^{(j)} = c_o + c^{(j)}$.

The design is specified by $\boldsymbol{n} = (n^{(1)}, n^{(2)}, n^{(3)})'$ with total sample size $n = n^{(1)} + n^{(2)} + n^{(3)}$. We define $s^{(13)} = s^{(1)} \cup s^{(3)}$, $s^{(23)} = s^{(2)} \cup s^{(3)}$, $n^{(13)} = n^{(1)} + n^{(3)}$, and $n^{(23)} = n^{(2)} + n^{(3)}$. For simplicity, we assume that the sampling fraction n/N is small so that the finite population correction factor can be ignored.

The total cost of the survey is $C = c_f + c_0 n + c^{(1)} n^{(1)} + c^{(2)} n^{(2)} + c^{(3)} n^{(3)}$, where c_f is the fixed cost for the survey that is independent of the sample size. Cost can be defined in terms of payments incurred by the statistical organisation. The coefficient $c^{(1)}$ would then be the marginal cost of collecting information on *only* y_1 from each unit given it is selected in the sample. Cost can also be defined in terms of the reporting load on the responding unit, measured in terms of interview time. The coefficient $c^{(1)}$ would then be the interview time required to collect *only* y_1 from each unit, given that the purpose of the survey has been explained and basic information, such as age and sex, has been collected. For convenience we assume that $c_f = 0$ or equivalently that c_f has been subtracted from C; the presence of a fixed cost does not affect the optimisation algorithms developed in this article.

When $K=2$, we consider three designs: (a) Single-phase design: $n^{(1)} = 0, n^{(2)} = 0$ and $n^{(3)} > 0$; (b) Two-phase design: $n^{(1)} > 0, n^{(2)} = 0$ and $n^{(3)} > 0$ (or by symmetry $n^{(1)} = 0, n^{(2)} > 0$ and $n^{(3)} > 0$); and (c) SQD: $n^{(1)} \geq 0$, $n^{(2)} \geq 0$ and $n^{(3)} \geq 0$. Designs (a) and (b) are special cases of (c). In the next three subsections we consider these designs.

2.1. Single-Phase Design

The single-phase design involves selecting $n = n^{(3)}$ units by SRSWOR. From each selected unit, information on both y_1 and y_2 is collected. The cost function simplifies to $C = t^{(3)} n^{(3)} = t^{(3)} n$. The estimator of Y_k is $\hat{Y}_k^{sp} = \hat{Y}_k^{(3)}$, where $\hat{Y}_k^{(j)} = N/n^{(j)} \sum_{i \in s^{(j)}} y_{ki}$, the Horvitz-Thompson estimator of Y_k based on sample $s^{(j)}$. The variance is given by $Var(\hat{Y}_k^{sp}) = V_k^{(3)}$, where $V_k^{(j)} = N^2 S_k^2 / n^{(j)}$ and $S_k^2 = S_{kk}$.

2.2. Two-Phase Design

The two-phase design involves selecting $n^{(3)}$ units by SRSWOR and collecting information on y_1 and y_2 and selecting $n^{(1)}$ units by SRSWOR and collecting information only on y_1. The cost function is $C = c_0 n + c^{(1)} n^{(1)} + c^{(3)} n^{(3)}$. The estimator of Y_1 is $\hat{Y}_1^{tp} = \hat{Y}_1^{(13)}$, where $\hat{Y}_1^{(13)} = N/n^{(13)} \sum_{i \in s^{(13)}} y_{ki}$, and has variance $Var\left(\hat{Y}_1^{tp}\right) = V_1^{(13)} = N^2 S_1^2 / n^{(13)}$. The two-phase regression estimator (see Sitter 1997) of Y_2 is $\hat{Y}_2^{tp} = \hat{Y}_2^{(3)} + B\left(\hat{Y}_1^{(13)} - \hat{Y}_1^{(3)}\right)$ with linearised variance $Var\left(\hat{Y}_2^{tp}\right) = N^2 S_2^2 / n^{(13)} + N^2 S_{2 \cdot 1}^2 (1 - n^{(3)}/n^{(13)})/n^{(3)}$, where $B = S_{12}/S_1^2$, $S_{2 \cdot 1}^2 = S_2^2(1 - \rho^2)$, and $\rho = S_{12}/(S_1 S_2)$. An alternative expression is $Var\left(\hat{Y}_2^{tp}\right) = N^2 S_2^2 / n_2^*$, where $n_2^* = n^{(3)} + n^{(1)} \rho^2 / \left(1 + n^{(1)}(1 - \rho^2)/n^{(3)}\right)$ can be regarded as an effective sample size. This shows how $\hat{Y}_2^{tp}$ exploits the information collected from the $n^{(1)}$ units in $s^{(1)}$ through the correlation. The two-phase estimator is typically used when $c^{(1)}$ is significantly smaller than $c^{(2)}$ and when ρ is large (Cochran 1977).

2.3. Split Questionnaire Designs

Next we consider two estimators when Patterns 1, 2, and 3 are used.

2.3.1. Simple Estimator

The simple estimator for a population characteristic is the Horvitz-Thompson estimator based on the sample of units from which information on that characteristic was collected. The simple estimators of Y_1 and Y_2 are $\hat{Y}_1^{se} = \hat{Y}_1^{(13)}$ and $\hat{Y}_2^{se} = \hat{Y}_2^{(23)}$, respectively, where $Var\left(\hat{Y}_1^{se}\right) = V_1^{(13)} = N^2 S_1^2 / n^{(13)}$ and $Var\left(\hat{Y}_2^{se}\right) = V_2^{(23)} = N^2 S_2^2 / n^{(23)}$.

2.3.2. Best Linear Unbiased Estimator (BLUE)

Best linear unbiased estimation (BLUE) is a general approach for combining different estimates in an optimal way (see Srivastava and Carter 1986; Fuller 1990). Here, the BLUE optimally combines the four estimates in $\boldsymbol{\alpha} = \left(\hat{Y}_1^{(1)}, \hat{Y}_2^{(2)}, \hat{Y}_1^{(3)}, \hat{Y}_2^{(3)}\right)'$ by taking into account their covariance structure.

The BLUE of $\boldsymbol{Y} = (Y_1, Y_2)'$ is $\left(\hat{\boldsymbol{Y}}^{\boldsymbol{sq}}\right) = \left(\hat{Y}_1^{sq}, \hat{Y}_2^{sq}\right)'$, where

$$\hat{\mathbf{Y}}^{\boldsymbol{sq}} = \mathbf{A}'\boldsymbol{\alpha} \tag{1}$$

and

$$\mathbf{A}' = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1} \tag{2}$$

is a 2×4 vector of composite factors. The matrix

$$\mathbf{V} = \begin{pmatrix} V_1^{(1)} & 0 & 0 & 0 \\ 0 & V_2^{(2)} & 0 & 0 \\ 0 & 0 & V_1^{(3)} & V_{1,2}^{(3)} \\ 0 & 0 & V_{1,2}^{(3)} & V_2^{(3)} \end{pmatrix} \tag{3}$$

is the variance – covariance matrix of $\boldsymbol{\alpha}$,

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}'$$

and $V_{1,2}^{(3)} = N^2 S_{12} / n^{(3)}$

The properties of the BLUE only depend on $E(\mathbf{Y}) = \boldsymbol{W}'\alpha$ and $Var(\mathbf{Y}) = \mathbf{V}$ and do not assume normality of any of the variables.

Note: when n/N is not approximately zero, (1) still applies but the 0 s in $\mathbf{V}$ would instead be negative.

The variance-covariance matrix of $\hat{\boldsymbol{Y}}^{\boldsymbol{sq}}$ is

$$Var\left(\hat{\mathbf{Y}}^{\boldsymbol{sq}}\right) = \mathbf{A}'\mathbf{V}\mathbf{A} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1} \tag{4}$$

Hence $\hat{Y}_k^{sq} = \mathbf{A}_\mathbf{k}'\alpha$ and $var\left(\hat{Y}_k^{sq}\right) = \mathbf{A}_\mathbf{k}'\mathbf{V}\mathbf{A}_\mathbf{k}$, where $\mathbf{A}_\mathbf{k}'$ is the kth row vector of $\mathbf{A}'$.

From (4) we may express $Var\left(\hat{Y}_1^{sq}\right) = N^2 S_1^2 / n_1^*$ and $Var\left(\hat{Y}_2^{sq}\right) = N^2 S_2^2 / n_2^*$, where $n_1^* = n^{(13)} + n^{(2)}\rho^2 / \left(1 + n^{(2)}(1-\rho^2)/n^{(3)}\right)$ and $n_2^* = n^{(23)} + n^{(1)}\rho^2 / \left(1 + n^{(1)}(1-\rho^2)/n^{(3)}\right)$.

These effective sample sizes show how the BLUE exploits, for the estimation of Y_k, the sample that does not collect information on y_k. These expressions show that $\hat{Y}^{sq}$ reduces to $\hat{Y}^{tp}$ when $n^{(2)} = 0$ and $\hat{Y}^{sq}$ reduces to $\hat{Y}^{se}$ when $\rho = 0$.

3. Optimal Split Questionnaire Design for $K=2$

Two common ways to define the optimal sample design and estimator are: minimising a variance function subject to a fixed cost constraint; and minimising a cost function subject to a fixed variance constraint (Bethel 1989; Chromy 1987).

We assume we are at the survey design stage and are evaluating which of the above designs and estimators to apply. This involves determining the optimal allocation for each design and associated estimator. This in turn means comparing the efficiency of $\hat{Y}^{tp}$, $\hat{Y}^{se}$, $\hat{Y}^{sq}$, and $\hat{Y}^{sp}$.

While this section focuses on two data items, a data item here could represent some linear combination of data items within a module; the corresponding cost function and covariance matrix could then be readily defined at the module level.

3.1. Minimise Variance for Fixed Cost

Rahim and Currie (1993) consider minimising $Z = \sum_{k=1}^{2} I_k CV\left(\hat{Y}_k\right)^2$, where $CV\left(\hat{Y}_k\right)^2 = Var\left(\hat{Y}_k\right)/Y_k^2$, subject to $C \leq C_B$, where C_B is the survey's budget, I_K is the measure of relative importance assigned to $\hat{Y}_k$ such that $I_1 + I_2 = 1$.

It is easy to show that the percentage gain of $\hat{Y}^{sq}$ over $\hat{Y}^{sp}$, measured by the size of Z, is a function of $\tilde{\mathbf{n}} = (\tilde{n}^{(1)}, \tilde{n}^{(2)}, \tilde{n}^{(3)})$, where $\tilde{n}^{(j)} = n^{(j)}/n$ is the proportion of the sampled units allocated to pattern j, and the design parameters ρ, $\tilde{c}_o = c_o/(t^{(3)})$, $c_r = c^{(1)}/c^{(2)}$, and $\phi_r = \phi_1/\phi_2$, where $\phi_k = I_k CV\left(\hat{Y}_k\right)^2$. The term $\tilde{c}_o$ is the proportion of the total unit cost of collecting y_1 and y_2 that is fixed and c_r is the ratio of the marginal costs of collecting y_1 and y_2.

It is useful to consider what values the design parameters may take in practice. First consider c_r. The marginal cost of collecting data item k would often correspond to the average time taken by the respondent to provide the information. Average times are routinely estimated by survey organisations during questionnaire development to manage respondent burden. While the absolute time taken to respond is influenced by many factors, such as the complexity of the question, the important design parameter here is the ratio of the time taken to provide the information on y_1 and y_2. So if the time taken to provide the information on the data items is approximately the same then $c_r = 1$.

Next consider $\tilde{c}_o$ and the situation where cost equates to respondent burden, as is often the case in an establishment survey where information on income and expenditure is collected. It is easy to see that most burden would result from providing the information on the data items. It may be reasonable in this situation to assume that $\tilde{c}_o < 10\%$. If the survey involves adding supplementary data items to an existing survey, then the cost of the supplementary survey is limited to the marginal cost of collecting the supplementary data items; this implies that $\tilde{c}_o = 0$. However, for surveys involving face-to-face household interviews and where cost is measured in dollars spent by the survey organisation, a substantial proportion of unit cost ($t^{(3)}$) would not depend upon the data items collected.

For example, at the Australian Bureau of Statistics about half a survey's budget is often spent on interviewer travel to selected households, implying that $\tilde{c}_o > 50\%$.

Also, high correlation between items is often observed in practice, especially for economic items. Lastly, it is clear that ϕ_r could vary widely from 1. For example, $\phi_r = 1$ if $I_1 = 0.5$ and $\bar{y}_1 = \bar{y}_2$ where y_k is a dichotomous variable so $CV(\hat{Y}_1) = CV(\hat{Y}_2)$; however, if instead $I_k = 0.66$ then $\phi_r = 2$.

To illustrate the performance of an SQD and other designs in a range of situations, Table 3 gives the percentage reduction in Z for each estimator with its optimal allocation (i.e., choice of $\tilde{\mathbf{n}}$) relative to $Var\left(\hat{Y}_k^{sp}\right)$ for different values of the design parameters.

It is easy to show that the optimal allocation for $\hat{Y}^{sq}$ is found by maximising

$$\frac{1 - \left(\phi_r/\tilde{n}_1^* + 1/\tilde{n}_2^*\right)\left[1 - (1 - \tilde{c}_0)(\tilde{c}^{(1)}\tilde{n}^{(2)} + \tilde{c}^{(2)}\tilde{n}^{(1)})\right]}{(\phi_r + 1)} \tag{5}$$

where $\tilde{n}_k^*$ has the same form as n_k^* except that $n^{(j)}$ is replaced with $\tilde{n}^{(j)} = n^{(j)}/n$ and $\tilde{c}^{(2)} = (1 + c_r)^{-1}$ and $\tilde{c}^{(1)} = c_r(1 + c_r)^{-1}$.

For simplicity the solution to this problem for $\hat{Y}^{sq}$ was found by a grid search, where $\tilde{n}^{(1)}$, $\tilde{n}^{(2)}$ and $\tilde{n}^{(3)}$ were allowed to range between 0 and 1 with the constraint that $\sum_j \tilde{n}^{(j)} = 1$. For consistency, the solutions for $\hat{Y}^{tp}$ and $\hat{Y}^{se}$ were found in the same way after substituting $\tilde{n}^{(2)} = 0$ and $\rho = 0$, respectively, into (5). However, it is easy to show

Table 3. Percentage reduction in Z relative to $\hat{Y}^{sp}$

Scenario 1:	$c_r = 1$, $\phi_r = 1$, $\rho = 0.8$				
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_o = 20\%$	$\tilde{c}_o = 10\%$	$\tilde{c}_o = 5\%$	$\tilde{c}_o = 0\%$
$\hat{Y}^{tp}$	3.1 (1.2)	5.4 (3.6)	8.1 (5.6)	9.7 (7.7)	10.2 (7.7)
$\hat{Y}^{se}$	0	0	0	0	0
$\hat{Y}^{sq}$	7.0 (3.7)	10.8 (9.0)	14.9 (13.8)	17.5 (15.3)	19.2 (18.8)
Scenario 2:	$c_r = 1$, $\phi_r = 1$, $\rho = 0.6$				
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_o = 20\%$	$\tilde{c}_o = 10\%$	$\tilde{c}_o = 5\%$	$\tilde{c}_o = 0\%$
$\hat{Y}^{tp}$	0	0	1.6	1.7	1.8
$\hat{Y}^{se}$	0	0	0	0	0
$\hat{Y}^{sq}$	0	2.1	5.4	7.2	9.0
Scenario 3:	$\tilde{c}_o = 10\%$, $\phi_r = 1$, $\rho = 0.8$				
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{Y}^{tp}$	18.3	14.1	10.2	8.8	8.2
$\hat{Y}^{se}$	0	0	0	0	0
$\hat{Y}^{sq}$	19.1	16.5	15.4	15.1	14.9
Scenario 4:	$\tilde{c}_o = 10\%$, $\phi_r = 1$, $\rho = 0.8$				
Method	$\phi_r = 2$	$\phi_r = 1.5$	$\phi_r = 1.1$	$\phi_r = 1.05$	$\phi_r = 1$
$\hat{Y}^{tp}$	13.6 (11.9)	10.7 (10.4)	9.4 (9.4)	8.7 (8.7)	8.4
$\hat{Y}^{se}$	1.0	0	0	0	0
$\hat{Y}^{sq}$	16.7 (15.0)	15.5 (14.9)	15.0 (14.9)	14.9 (14.9)	14.9
Scenario 5:	$\tilde{c}_o = 10\%$, $\phi_r = 2$, $\rho = 0.8$				
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{Y}^{tp}$	12.8	9.0	11.0	12.6	13.6
$\hat{Y}^{se}$	0	0	0	1.0	1.0
$\hat{Y}^{sq}$	15.5	15.1	15.8	16.5	16.7

that the optimal allocation for $\hat{Y}^{tp}$ is

$$\begin{aligned} \tilde{n}^{(1)} &= 1 - \sqrt{1-R}\, when\, 0 \le R \le 1 \\ &= 0\, otherwise \\ \tilde{n}^{(3)} &= 1 - \tilde{n}^{(1)} \end{aligned} \tag{6}$$

where

$$R = \left[\frac{c_r + \tilde{c}_o}{1 + c_r}(\phi + 1) - (\rho^2 + \phi)\right]^{1/2} \left[\left(\frac{c_r + \tilde{c}_o}{1 + c_r} - 1\right)(\rho^2 + \phi)\right]^{-1/2}$$

and for $\hat{Y}^{se}$ the optimal allocation is given by (6) with

$$R = \left[\frac{c_r + \tilde{c}_o}{1 + c_r}(\phi + 1) - \phi\right]^{1/2} \left[\left(\frac{c_r + \tilde{c}_o}{1 + c_r} - 1\right)\phi\right]^{-1/2}.$$

Table 3 shows that reductions in Z when using $\hat{Y}^{sq}$ instead of $\hat{Y}^{tp}$ are appreciable when $\rho = 0.8$, $\phi_r = 1$, $c_r = 1$ and $\tilde{c}_0$ is small. For example, in Scenario 1 when $\tilde{c}_0 = 10\%$, the values of Z for $\hat{Y}^{sq}$ and $\hat{Y}^{tp}$ were 14.9% and 8.1% smaller than those for $\hat{Y}^{sp}$, respectively. Scenario 2 shows that if ρ is reduced from 0.8 to 0.6, the benefit of using $\hat{Y}^{sq}$ instead of $\hat{Y}^{tp}$ is reduced and the gains from using $\hat{Y}^{sq}$ instead of $\hat{Y}^{sp}$ are only positive when the fixed cost per unit is small (i.e., $\tilde{c}_0 \le 30\%$). These scenarios illustrate the general point that the lower the value of $\tilde{c}_0$ and the larger the value of ρ, the larger the potential gains under $\hat{Y}^{sq}$ and $\hat{Y}^{tp}$ relative to $\hat{Y}^{sp}$.

Scenarios 3–5 fix $\tilde{c}_0 = 10\%$ and $\rho = 0.8$ and vary ϕ_r and c_r. Scenarios 3 and 4 show that as ϕ_r or c_r increases from 1 to 2 the superiority of $\hat{Y}^{sq}$ over $\hat{Y}^{tp}$ reduces. Scenario 5 shows that when one of the data items has great importance and is costly to collect compared with the other data item, $\hat{Y}^{sq}$ consistently outperforms $\hat{Y}^{tp}$. When $\phi_r = 2$, the gains of $\hat{Y}^{sq}$ are between 3–6 percentage points higher than $\hat{Y}^{tp}$ as c_r ranges from 1 to 2. Scenarios 3–5 illustrate that the interaction between ϕ_r and c_r affects the size of the gains of $\hat{Y}^{sq}$ relative to $\hat{Y}^{tp}$.

Table 3 readily allows us to determine the gains arising from using an SQD under a range of values of the design parameters. In the scenarios considered, SQDs showed gains over the alternative designs.

The zeros in Table 3 mean that the optimal allocation for the estimator is a single phase allocation (i.e., $n = n^{(3)}$). This was often the case for $\hat{Y}^{se}$. However, in Scenario 4 when $\phi_r = 2$ $\hat{Y}^{se}$ had a Z value that was 1.0% smaller than that for $\hat{Y}^{sp}$.

In general optimal allocation requires assuming population parameters, referred to here as design parameters, at the design stage. The sensitivity of the optimum to errors in the design parameters was investigated by Cochran (1977) (see Section 5A 1) in the case of optimal allocation to strata.

Table 3 shows the the sensitivity of the optimum to errors in the design parameters ρ and ϕ_r for Scenarios 1 and 4, respectively. For Scenario 1 the numbers in parentheses show the gains when the allocation is based on $\rho = 0.6$, when in fact the correct population value is $\rho = 0.8$. For example, in Scenario 1 and $\tilde{c}_0 = 0\%$ the gain of using $\hat{Y}^{sq}$ instead of $\hat{Y}^{sp}$ is 19.2% if the SQD allocation was based on $\rho = 0.8$ and reduces marginally to 18.8% if the

allocation was based on $\rho = 0.6$. The gains of $\hat{Y}^{sq}$ tend to be less sensitive to errors in ρ than $\hat{Y}^{tp}$.

For Scenario 4 the numbers in parentheses show the gains when the allocation is based on $\phi_r = 1$, when in fact the correct population value varies. For example, in Scenario 4 the gain of using $\hat{Y}^{sq}$ instead of $\hat{Y}^{sp}$ is 15.5% if the allocation was correctly based on $\phi_r = 1.5$ and reduces marginally to 14.9% if the allocation was incorrectly based on $\phi_r = 1$. The gains due to both $\hat{Y}^{sq}$ and $\hat{Y}^{tp}$ are very insensitive to errors in ϕ_r.

3.2. Minimise Cost for Fixed Variance

An alternative objective is to minimise C given $Var(\hat{Y}_k)/Y_k^2 \leq v_k^2$ for $k = 1, 2$, where v_k represents the target coefficient of variation of $\hat{Y}_k$. It is easy to show that the percentage reduction in C of $\hat{Y}^{sq}$ over $\hat{Y}^{sp}$ is a function of $\tilde{\mathbf{n}}$ and the design parameters ρ, $\tilde{c}_o$, c_r, and $L = q_2/q_1$, where $q_k = CV(y_k)^2/v_k^2$ is the sample size required to meet the variance constraint for data item k under a single phase design and $CV(y_k) = S_k^2/\bar{Y}_k^2$. What are the likely values of L? If $L = 1$ then the effective sample size, n_k^*, required to meet the variance constraints is the same for both y_1 and y_2; this would occur, for example, if y_1 and y_2 were dichotomous variables, $\bar{y}_1 = \bar{y}_2$ and $v_1 = v_2$; if instead, $v_1 = \sqrt{2}v_2$ then $L = 2$.

It is easy to show that the optimal allocation for $\hat{Y}^{sq}$ is found by maximising

$$= 1 - \left(1 - (1 - \tilde{c}_0)(c_r(1 + c_r)^{-1}\tilde{n}^{(2)} + (1 + c_r)^{-1}\tilde{n}^{(1)})\right)/(\tilde{n}_1^* max\,(L^{-1}, 1)) \tag{7}$$

subject to $\tilde{n}_2^*/\tilde{n}_1^* \geq L^{-1}$.

For simplicity the solution to this problem for $\hat{Y}^{sq}$ was found by a grid search, where $\tilde{n}^{(1)}$, $\tilde{n}^{(2)}$ and $\tilde{n}^{(3)}$ were allowed to range between 0 and 1 with the constraint that $\sum_j \tilde{n}^{(j)} = 1$ and $\tilde{n}_2^*/\tilde{n}_1^* > L^{-1}$. For consistency, the solutions for $\hat{Y}^{tp}$ and $\hat{Y}^{se}$ were found in the same way after substituting $\tilde{n}^{(2)} = 0$ and $\rho = 0$ respectively into (7). However, it is easy to show that the optimal allocation for $\hat{Y}^{tp}$ is

$$\begin{aligned} \tilde{n}^{(1)} &= 1 - \sqrt{1 + H} \text{ if } -1 < H \leq 0 \text{ and } \tilde{n}_2^* \leq L^{-1} \\ &= \frac{1 - L^{-1}}{1 - \rho^2 L^{-1}} \text{ otherwise} \\ \tilde{n}^{(3)} &= 1 - \tilde{n}^{(1)} \end{aligned} \tag{8}$$

where

$$H = \frac{1 - \rho^2 - (1 - \tilde{c}_0)\tilde{c}^{(1)}}{\rho^2(1 - \tilde{c}_0)\tilde{c}^{(1)}}$$

For $\hat{Y}^{se}$ the optimal allocation is

$$\begin{aligned} \tilde{n}^{(1)} &= max\,(0, 1 - L^{-1}) \\ \tilde{n}^{(3)} &= 1 - \tilde{n}^{(1)} \end{aligned} \tag{9}$$

Table 4 compares the relative size of C for each of the estimators under their optimal allocation. Again, to illustrate the performance of the SQD and the other designs in a range

Table 4. Percentage reduction in C, subject to $Var(\hat{Y}_k)/\hat{Y}_k^2 < v_k^2$, relative to $\hat{Y}^{sp}$

Scenario 6: $c_r = 1$, $L = 1$, $\rho = 0.8$					
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_o = 20\%$	$\tilde{c}_o = 10\%$	$\tilde{c}_o = 5\%$	$\tilde{c}_o = 0\%$
$\hat{Y}^{tp}$	0	0	0.8	1.3	1.5
$\hat{Y}^{se}$	0	0	0	0	0
$\hat{Y}^{sq}$	5.5	10.0	14.5	16.8	19.0
Scenario 7: $c_r = 1$, $L = 1$, $\rho = 0.6$					
Method	$\tilde{c}_o = 30\%$	$\tilde{c}_o = 20\%$	$\tilde{c}_o = 10\%$	$\tilde{c}_o = 5\%$	$\tilde{c}_o = 0\%$
$\hat{Y}^{tp}$	0	0	0	0	0
$\hat{Y}^{se}$	0	0	0	0	0
$\hat{Y}^{sq}$	0	1.0	5.0	7.0	9.0
Scenario 8: $\tilde{c}_o = 10\%$, $L = 1$, $\rho = 0.8$					
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{Y}^{tp}$	5.0	3.0	1.4	0.9	0.8
$\hat{Y}^{se}$	0	0	0	0	0
$\hat{Y}^{sq}$	14.5	14.5	14.5	14.5	14.5
Scenario 9: $\tilde{c}_o = 10\%$, $\tilde{c}_r = 1$, $\rho = 0.8$					
Method	$L = 2$	$L = 1.5$	$L = 1.1$	$L = 1.05$	$L = 1$
$\hat{Y}^{tp}$	30.8	23.7	8.0	5.3	0.8
$\hat{Y}^{se}$	19.5	13.2	6.4	1.0	0
$\hat{Y}^{sq}$	33.2	27.3	17.5	15.5	14.5
Scenario 10: $\tilde{c}_o = 10\%$, $L = 2$, $\rho = 0.8$					
Method	$c_r = 2$	$c_r = 1.5$	$c_r = 1.1$	$c_r = 1.05$	$c_r = 1$
$\hat{Y}^{tp}$	19.5	24.0	29.1	30.0	30.8
$\hat{Y}^{se}$	12.3	15.4	18.9	19.4	19.5
$\hat{Y}^{sq}$	23.7	27.5	31.8	32.5	33.2

of situations, Table 4 considers different values of L and the same variations to values of $\tilde{c}_0$, c_r, and ρ as in Table 3.

Table 4 shows that for the paramater values considered the gains of $\hat{Y}^{sq}$ relative to its alternatives are greatest when ρ is 0.8, $L = 1$, $c_r = 1$ and $\tilde{c}_0$ is small. For example, in Scenario 6, $\hat{Y}^{sq}$ costs between 5.5% and 19.0% less than $\hat{Y}^{tp}$ and $\hat{Y}^{se}$. (The optimal allocations under Scenario 6 for $\hat{Y}^{sq}$ were all $\tilde{\boldsymbol{n}} = (36\%, 36\%, 28\%)'$.) Scenario 7 shows that as ρ decreases from 0.8 to 0.6 the relative gains of $\hat{Y}^{sq}$ over the alternatives are reduced.

Scenarios 6 and 7 again illustrate the general point that the lower the value of $\tilde{c}_0$ and the larger the value of ρ, the larger the potential gains from using $\hat{Y}^{sq}$. Interestingly, in these scenarios the values of $\tilde{c}_0$ and ρ had only a marginal impact on the efficiency of $\hat{Y}^{tp}$.

Scenarios 8–10 fix $\tilde{c}_0 = 10\%$ and $\rho = 0.8$ and vary L and c_r. When $L = 1$, Scenario 8 shows that as c_r ranges from 1 to 2, the efficiency of $\hat{Y}^{sq}$ over $\hat{Y}^{tp}$ and $\hat{Y}^{se}$ decreases but remains significant. If $c_r = 1$, Scenario 9 shows that as L increases from 1 to 2 the efficiency of $\hat{Y}^{sq}$ relative to $\hat{Y}^{tp}$ decreases. Scenario 10 shows that when $L = 2$, the gains of $\hat{Y}^{sq}$ relative to $\hat{Y}^{tp}$ are marginal as c_r goes from 1 to 2.

Scenarios 9 and 10 shows that as L increases from 1 the relative efficiency of $\hat{Y}^{se}$ relative to $\hat{Y}^{sp}$ also increases.

Tables 3 and 4 show that when minimising cost and variance the gains of $\hat{Y}^{sq}$ relative to the alternatives depend upon the same factors, noting that ϕ_r and L both measure some concept of relative importance/variability of the two data items. The gains of an SQD

relative to the other designs were larger when minimising cost under a variance constraint; the variance constraint seems more rigid than a cost constraint, illustrating the flexibility of $\hat{Y}^{sq}$ through its use of all three patterns and its exploitation of correlation between the data items.

4. Split Questionnaire Design for the BLUE and Arbitrary *K*

4.1. BLUE for Arbitrary K

This section describes SQDs using BLUE for arbitrary K and notes that the two-phase and the simple estimator are special cases. We empirically go on to evaluate an SQD and the multi-phase design for $K = 4$.

For a given K, there are $J = \sum_{p=1}^{K} {}^{K}C_p$ possible patterns. We define the number of data items assigned to pattern j as $g^{(j)}$ and the set of data items assigned to pattern j as $u^{(j)}$. For example, in Table 2 $g^{(3)} = 2$ and $u^{(3)} = (y_1, y_2)$. We define the pattern that collects all K data items by $j = J$ such that $g^{(J)} = K$ and $u^{(J)} = (y_1, \ldots, y_K)$.

Allocation for an SQD involves choosing $n^{(j)}$, which is the number of sample units from which the set of data items $g^{(j)}$ is collected, for $j = 1, \ldots, J$. The case of $K = 4$ is shown in Table 5. The cost function for such a design is $C = c_0 n + \sum_{j=1}^{J} n^{(j)} c^{(j)} = c_0 n + \sum_{j=1}^{J} n^{(j)} \sum_{k \in g^{(j)}} c_k$, where c_k is the marginal cost of collecting y_k and we assume that $c^{(j)} = \sum_{k \in g^{(j)}} c_k$. An alternative form is $C = \sum_{j=1}^{J} n^{(j)} t^{(j)}$, where $t^{(j)} = c_0 + \sum_{k \in g^{(j)}} c_k$ is the total cost (fixed cost plus marginal cost) per unit that is allocated to pattern j.

The BLUE is a linear combination of the $M = \sum_{j=1}^{J} g^{(j)}$ Horvitz-Thompson estimates, $\hat{Y}_k^{(j)}$, $k \in u^{(j)}, j = 1, 2, \ldots, J$ that can be calculated from the J different patterns. For example, if $K = 4$ then $J = 15$ and $M = 32$ (see Table 5).

Table 5. Patterns for $K = 4$

Pattern	y_1	y_2	y_3	y_4	Marginal cost	Sample size
1	X				$c^{(1)}$	$n^{(1)}$
2		X			$c^{(2)}$	$n^{(2)}$
3			X		$c^{(3)}$	$n^{(3)}$
4				X	$c^{(4)}$	$n^{(4)}$
5	X	X			$c^{(5)}$	$n^{(5)}$
6	X		X		$c^{(6)}$	$n^{(6)}$
7	X			X	$c^{(7)}$	$n^{(7)}$
8		X	X		$c^{(8)}$	$n^{(8)}$
9		X		X	$c^{(9)}$	$n^{(9)}$
10			X	X	$c^{(10)}$	$n^{(10)}$
11	X	X	X		$c^{(11)}$	$n^{(11)}$
12	X	X		X	$c^{(12)}$	$n^{(12)}$
13	X		X	X	$c^{(13)}$	$n^{(13)}$
14		X	X	X	$c^{(14)}$	$n^{(14)}$
15	X	X	X	X	$c^{(15)}$	$n^{(15)}$

The BLUE of $\mathbf{Y}$ is given by (1) and its variance by (4) where: $\mathbf{V}$ is the $M \times M$ block diagonal matrix with diagonal elements $\boldsymbol{V}^{(j)} = N^2\mathbf{F}^{(j)}/n^{(j)}$, where $\mathbf{F}^{(j)}$ is the $g^{(j)} \times g^{(j)}$ population variance-covariance matrix with elements $S_{kk'}$ where both k' and k belong to the set $u^{(j)}$; α is an M column vector of estimates $Y_k^{(j)}$ ordered by pattern j and data item k; $\mathbf{W}$ is an $M \times K$ matrix with 1 in position (m, k) if the mth element in $\boldsymbol{\alpha}$ is an estimate of Y_k and zero otherwise and where $m = 1, \ldots, M$. The matrix $\mathbf{A}'$ is a $K \times M$ matrix of composite factors. Hence $\hat{Y}_k^{sq} = \mathbf{A}_\mathbf{k}'\alpha$ and $Var\left(\hat{Y}_k^{sq}\right) = \mathbf{A}_\mathbf{k}'\mathbf{V}\mathbf{A}_\mathbf{k}$, where $\mathbf{A}_\mathbf{k}'$ is the kth row vector of $\mathbf{A}'$. The multi-phase regression estimator (Sitter 1997), $\hat{Y}^{mp}$, for arbitrary K is given by (1) except that the data item patterns are restricted to following a monotonic pattern. For arbitrary K, there are $K!$ different monotonic patterns, each with K patterns. The simple estimator, $\hat{Y}^{se}$, for arbitrary K is given by (1) with the off diagonals of $\mathbf{V}$ set to zero.

4.2. *Optimal Allocation for Arbitrary K*

For arbitrary K the optimal allocation must be found for the vector $\boldsymbol{n} = (n^{(1)}, \ldots, n^{(J)})'$. In this section we define the allocation problems. Solving these problems requires only standard linear programming. In the empirical study below for $K = 4$, Newton's method was found to converge quickly.

4.2.1. Deciding Which Patterns to Exclude

The number of possible patterns, J, increases very quickly with the number of data items, K, since $J = 2^K - 1$. For example, if $K = 10$ then $J = 1,023$. In practice only a subset of patterns can be considered for an optimal SQD. Next we consider one way to rank the relative efficiency of the J patterns. If only J_o patterns are to be considered, where $J_o < J$, then the J_o patterns with the highest rank can be used in the optimal allocation algorithm.

To motivate this approach, an approximation to the effective sample size of $\hat{Y}_k^{(sq)}$ is $n_k^* = \sum_j n^{(j)}R_k^{(j)}$, where $R_k^{(j)} = 1 - S_{k\cdot u^{(j)}}^2 S_k^{-2}$, $S_{k\cdot u^{(j)}}^2$ is the variance of the residuals from the regression of y_k against the variables in $u^{(j)}$. Of course $R_k^{(j)} = 1$ if y_k is observed in pattern j (i.e., if $y_k \in u^{(j)}$). If y_k is not observed in pattern j then $R_k^{(j)}$ will be close to 1 if the observed data items in pattern j, $u^{(j)}$, explain much of the variation in y_k. In contrast, if observed data items in pattern j are uncorrelated with y_k then $R_k^{(j)}$ will be zero. This approximation of the effective sample size is valid only when the pattern that collects all K data items, namely $j = J$, is assigned a nonzero allocation (i.e., $n^{(J)} > 0$). (As compared with the effective sample size expressions n_1^* and n_2^* in Section 2.3.2, this crude approximation performs well when $n^{(2)}/n^{(3)}$ and $n^{(1)}/n^{(3)}$ are small, respectively.)

When minimising variance subject to fixed cost the relative efficiency measure for pattern j is

$$E_{min\ C}^{(j)} = \frac{\Sigma_k I_k R_k^{(j)}}{c^{(j)}}$$

which is pattern j's unit contribution to the effective sample size for population total k, $R_k^{(j)}$, weighted by importance, I_k, and the inverse of the pattern cost, $c^{(j)}$, and then summed over over all k.

In an analagous way, when minimising cost subject to fixed variance the relative efficiency measure for pattern j is

$$E^{(j)}_{min\ V} = \frac{\Sigma_k \tilde{L}_k R^{(j)}_k}{c^{(j)}}$$

This approach worked well in scenarios 14–16 (see Section 4.3) where $J = 15$. The patterns assigned a nonzero allocation by the optimal allocation algorithm, which considered all J patterns, were those with the highest values of $E^{(j)}_{min\ V}$.

To ensure that the approximation of the effective sample size is valid, pattern $j = J$ should always be considered by the optimal allocation algorithm, regardless of its value of $E^{(j)}_{min\ V}$ or $E^{(j)}_{min\ C}$.

For other practical reasons, such as respondent burden and simplicity in the form design, the survey design, we may exclude some patterns. Such restrictions are investigated in Section 4.3.

4.2.2. Minimise Variance for Fixed Cost

In the context of multi-variate allocation in stratified single-phase designs, Rahim and Currie (1993) formulated the problem of minimising a variance function subject to fixed cost constraints as finding the value of **n** that minimises

$$Z = \sum_{k=1}^{K} I_k CV(\hat{Y}_k)^2 \tag{10}$$

subject to the constraint that $C_B > C = \sum_{j=1}^{J} t^{(j)} n^{(j)}$, where $CV(\hat{Y}_k) = Var(\hat{Y}_k)^{1/2}/Y_k$, $Var(\hat{Y}_k)$ is as described in Section 4.1, and I_k is the measure of importance assigned to $\hat{Y}_k$.

4.2.3. Minimise Cost for Fixed Variance

In the context of multi-variate allocation in stratified single-phase designs Kokan and Khan (1967); Bethel (1989), and Chromy (1987) define the optimisation problem by finding the value of **n** that minimises C subject to the constraint

$$V(\hat{Y}_k) < \nu_k^2 Y_k^2 \tag{11}$$

for all k, where $V(\hat{Y}_k)$ is as described in Section 4.1, and ν_k is the maximum value that $CV(\hat{Y}_k)$ may take in order to meet the design constraints.

The problem of allocation in an SQD with BLUE, given by (1), is defined by minimising C subject to (11).

4.3. Empirical Evaluation

To illustrate the gains of an SQD, consider a hypothetical survey with four data items. Accordingly, there are 15 patterns (see Table 5). We let $CV(y_k) = 0.65$ and the correlation

matrix be

$$\rho = \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.8 \\ 0 & 0 & 0.8 & 1 \end{pmatrix}$$

We assume that we are at the survey design stage and wish to evaluate the estimators, $\hat{Y}^{sq}$, $\hat{Y}^{se}$, and $\hat{Y}^{mp}$, defined in Section 4.1 and their associated optimal allocations. We do not consider $\hat{Y}^{sp}$ because it was shown to be suboptimal. To determine the optimal allocations we will apply the algorithms in Section 4.2.2. All $J = 15$ patterns are considered unless specified otherwise. The optimal allocation for $\hat{Y}^{mp}$ was found by applying the algorithm to all possible monotonic patterns and selecting the one that was optimal.

First, we consider a number of scenarios where the objective is to minimise Z with the constraint that $C_B < 250$. Table 6 gives the minimum value of Z and the associated optimal allocation for $\hat{Y}^{se}$, $\hat{Y}^{mp}$, and $\hat{Y}^{sq}$ using slightly different parameters. Note that values of $n^{(j)}$ not in Table 6 were zero for the allocations under all scenarios.

Under Scenario 11, the design parameters are given by $I_k = 0.25$, $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$– that is, the marginal cost of collecting each data item is 1 cost unit. In this scenario, the value of Z for $\hat{Y}^{sq}$ is 8.0% smaller than that for $\hat{Y}^{mp}$. Scenario 12 considers the impact on the results of Scenario 11 if we increase the fixed cost per unit from 0.25 to 1, which is equal to the marginal cost associated with collecting each data item; as a result the relative gain of $\hat{Y}^{sq}$ over $\hat{Y}^{mp}$ reduces to 5.7%.

Scenario 13 considers the impact on the results of Scenario 11 if instead $I_1 = 0.35$, $I_2 = 0.15$, $I_3 = 0.2$, and $I_4 = 0.3$. In Scenario 13 $\hat{Y}^{sq}$ achieved a gain of 11.8% over $\hat{Y}^{mp}$. This illustrates that the combination of patterns not available to $\hat{Y}^{mp}$ was exploited by $\hat{Y}^{sq}$.

Table 6. Optimal allocation: minimise variance for fixed coat

	Allocation							
	$n^{(3)}$	$n^{(6)}$	$n^{(7)}$	$n^{(8)}$	$n^{(9)}$	$n^{(13)}$	$n^{(15)}$	$Z \times 10^4$
Scenario 11: $c_o = 0.25$ and $c^{(j)} = g^{(j)}$, $I_k = 0.25$								
$\hat{Y}^{se}$	7	0	11	0	0	0	47	109
$\hat{Y}^{mp}$	0	0	38	0	0	0	38	100
$\hat{Y}^{sq}$	0	26	6	6	27	0	23	92
Scenario 12: $c_o = 1$ and $c^{(j)} = g^{(j)}$, $I_k = 0.25$								
$\hat{Y}^{se}$	0	0	0	0	0	0	50	130
$\hat{Y}^{mp}$	0	26	0	0	0	0	34	123
$\hat{Y}^{sq}$	0	18	4	4	18	0	22	116
Scenario 13: $c_o = 0.25$ and $c^{(j)} = g^{(j)}$, $I_1 = 0.35$, $I_2 = 0.15$, $I_3 = 0.2$, $I_4 = 0.3$								
$\hat{Y}^{se}$	0	0	10	0	0	7	47	109
$\hat{Y}^{mp}$	0	0	16	0	0	0	50	102
$\hat{Y}^{sq}$	0	17	31	8	10	0	21	90

We considered the impact on the results for Scenario 11 if the cost of collecting information on data item y_k is k (e.g., $c_7 = 1 + 4$ where pattern $j = 7$ means y_1 and y_4 are collected). The result was that $\hat{Y}^{sq}$ and $\hat{Y}^{mp}$ were equally efficient.

Suppose that, due to respondent burden, we restrict the number of data items that may be collected from a unit to be at most 2 (i.e., $j < 11$) or 3 (i.e., $j < 15$). As a result, the minimum values of $Z \times 10^{-4}$ for $\hat{Y}^{sq}$ became 96 and 95 respectively– still 4% more efficient than $\hat{Y}^{mp}$ under Scenario 11, where such a restriction *cannot* be imposed. In Scenarios 11, 12, and 13, $\hat{Y}^{se}$ is less efficient than $\hat{Y}^{mp}$ and $\hat{Y}^{sq}$.

Next we consider a number of scenarios where the objective is to minimise C for fixed variance. Table 7 gives the minimum value of C and the associated optimal allocation for $\hat{Y}^{se}$, $\hat{Y}^{mp}$, and $\hat{Y}^{sq}$ under slightly different constraints. Again, values of $n^{(j)}$ not in Table 6 were zero for the allocations under all scenarios.

Under Scenario 14, the design constraints are given by $v_k = 0.05$, $c_0 = 0.25$ and $c^{(j)} = g^{(j)}$. For this scenario, under $\hat{Y}^{sq}$ the design cost is $C = 943$, or 14.4% smaller than the cost for $\hat{Y}^{mp}$. This scenario highlights how the multi-phase design is restrictive. The symmetry in the design constraints and correlation matrix implies that the optimal allocation for y_1 and y_2 will be a mirror image of the optimal allocation for y_3 and y_4: symmetry means the optimal allocation is unchanged if y_1 is collected instead of y_2 and vice versa and y_3 is collected instead of y_4 and vice versa. As can be seen from Table 7, the optimal allocations for $\hat{Y}^{sq}$ under Scenarios 14,15, and 16 are approximately symmetric. However, the multi-phase allocation can only be symmetric in this way if it reduces to the single-phase allocation.

When we restricted the number of data items that may be collected from a unit to be 2 in Scenario 14, the design cost was $C = 1,035$, which was only 10.6% larger than without this restriction and still 6% smaller than $\hat{Y}^{mp}$.

Scenario 15 considers the impact on the results for Scenario 14 when c_0 is increased from 0.25 to 1. The relative efficiency of $\hat{Y}^{sq}$ over $\hat{Y}^{mp}$ was reduced to 9.5%. We also considered the impact of changing the cost parameters of Scenario 14 so that the cost of collecting information on data item y_k is k. The result was that $\hat{Y}^{sq}$ and $\hat{Y}^{mp}$ were equally efficient.

Table 7. Optimal allocation: minimise cost for fixed variance

	Allocation									
	$n^{(1)}$	$n^{(2)}$	$n^{(5)}$	$n^{(6)}$	$n^{(7)}$	$n^{(8)}$	$n^{(9)}$	$n^{(12)}$	$n^{(15)}$	C
Scenario 14: $c_o = 0.25$ and $c^{(j)} = g^{(j)}$, $v_k = 0.05$										
$\hat{Y}^{se}$	0	0	0	0	0	0	0	0	261	1,109
$\hat{Y}^{mp}$	0	0	0	0	21	0	0	0	248	1,101
$\hat{Y}^{sq}$	0	0	0	81	6	6	79	0	132	943
Scenario 15: $c_o = 1$ and $c^{(j)} = g^{(j)}$, $v_k = 0.05$										
$\hat{Y}^{se}$	0	0	0	0	0	0	0	0	261	1,305
$\hat{Y}^{mp}$	0	0	0	0	0	0	0	0	261	1,305
$\hat{Y}^{sq}$	0	0	0	62	4	4	62	0	157	1,181
Scenario 16: $c_o = 0.25$ and $c^{(j)} = g^{(j)}$, $v_1 = v_2 = 0.05$, $v_3 = v_4 = 0.06$										
$\hat{Y}^{se}$	0	0	78	0	0	0	0	0	183	953
$\hat{Y}^{mp}$	94	0	0	0	0	0	0	63	146	942
$\hat{Y}^{sq}$	49	49	0	0	0	0	0	0	183	900

Scenario 16 considers the impact on the results of Scenario 14 of setting $v_3 = v_4 = 0.06$. Under this scenario, the cost for $\hat{Y}^{sq}$ was 900, which is 4.5% smaller than the corresponding cost for $\hat{Y}^{mp}$.

In all scenarios $\hat{Y}^{sq}$ was never less efficient and sometimes appreciably more efficient than the alternatives. It is easy to see that the optimal allocation for the single-phase design would not change across Scenarios 14–16, highlighting its inefficiency as a general approach to sample design.

4.3.1. Design Parameters

It is easy to show that the minimum set of parameters that are needed to measure the gains of SQD relative to $\hat{Y}^{sp}$ when minimising variance for fixed cost are $\tilde{n}^{(j)}$, $\tilde{c}_o = c_o/(c^J + c_0)$, $\tilde{c}_k = c_k/c^J$ $\rho = (\rho_{kk'})$, and $\tilde{\phi}_k$, where $\phi_k = I_k CV(\hat{Y}_k)^2$, $\tilde{\phi}_k = \phi_k/\phi$ and $\phi = \sum_k \phi_k$.

Similarly, it is easy to show that the parameters that are needed to fully describe the gains of an SQD relative to $\hat{Y}^{sp}$ when minimising cost for fixed variance are $\tilde{n}^{(j)}$, $\tilde{c}_0$, $\tilde{c}_k$, and $\tilde{L}_k = q_k/q_{k'}$, where $q_{k'} = CV(y_{k'})^2/v_{k'}^2$ and k' denotes one of the K constraints.

Expressing the design parameters as ratios and proportions, rather than absolute values, makes the results of Section 4.2 more general. For example, the relative sizes of Z for $\hat{Y}^{sq}$, $\hat{Y}^{se}$ and $\hat{Y}^{mp}$ under Scenario 11 extend to all designs with the following design parameters $\tilde{c}^0 = 12.5\%$, $\tilde{c}_k = 0.25$, and $\tilde{\phi}_k = 0.25$.

5. Discussion

The SQD and associated BLUE provide a general approach for a multivariate survey which gives maximum flexibility in meeting survey objectives. The algorithms developed enable us to compare a range of designs that include optimal multi-phase designs, restricted and unrestricted SQD.

SQDs showed appreciable gains (up to 19%) relative to the multi-phase designs in many scenarios. The size of the gains depends upon specific costs parameters associated with the design, the variance objectives of the survey and the correlation between the survey's data items.

SQDs can be used in cases where the maximum number of data items collected from a sample unit is restricted to be less than the number of population totals we wish to estimate. The main reason for such a restriction is to reduce respondent burden in order to improve response rates or to increase the number of data items that can be collected from the sample while controlling the burden on each respondent. When the fixed cost per selected unit is small, this restriction has only marginal impact on the efficiency of an SQD. This flexibility is not available to the multi-phase approach.

With the replacement of pen-and-paper interviewing by computer-assisted interviewing (CAI) in recent years comes the potential for more sophisticated questionnaire designs. To implement an SQD as described here, the choice of which data items to collect is made prior to the interview: this results in survey data being Missing Completely at Random (MCR) (see Little 1988). An interesting extension of this article is to make the choice of which data items to collect dependent upon the answers to the data items, potentially leading to further gains.

It is worthwhile mentioning some limitations of this article. Firstly, while the focus here has been on estimation of totals, survey data is also used for analysis purposes

(e.g., estimation of regression coefficients). An extension of this work would be to also incorporate a design constraint on estimates of regression coefficients so analysts' requirements are not compromised.

Secondly, we consider only simple random sampling, whereas many surveys involve stratification, clustering, and unequal probabilities of selection.

Thirdly, the analysis of survey data collected by an SQD would be more complicated than if the survey data were collected by an SPD. However, methods such as Multiple Imputation (Rubin and Little 1987) and the EM algorithm (Rubin and Little 1987) could be applied directly to the analysis of data collected by an SQD.

Finally, it is well-known that different routings through a questionnaire can sometimes affect response values (for discussion see Lyberg et al. 1997, Chapter 5). Care should be taken to ensure this issue is addressed when considering an SQD.

6. References

Adiguzel, F. and Wedel, M. (2008). Split Questionnaire Designs for Massive Surveys. Journal of Marketing Research, 25, 608–617.

Bell, P. (2001). Comparison of Alternative Labour Force Survey Estimators. Survey Methodology, 27, 53–63.

Bethel, J. (1989). An Optimal Allocation Algorithm for Multivariate Surveys. Survey Methodology, 15, 47–57.

Chromy, J. (1987). Design Optimization with Multivariate Objectives. Proceedings of the American Statistical Association, Survey Research Section, 194–199.

Cochran, W.C. (1977). Sampling Techniques. New York: John Wiley and Sons.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91, 490–498.

Fuller, W.A. (1990). Analysis of Repeated Surveys. Survey Methodology, 16, 167–180.

Gelman, A., King, G., and Liu, C. (1998). Not Asked and Not Answered: Multiple Imputation for Multiple Surveys. Journal of the American Statistical Association, 93, 846–857.

Kokan, A.R. and Khan, S. (1967). Optimal Allocation in Multivariate Surveys: An Analytical Solution. Journal of the Royal Statistical Society, Series B, 29, 115–125.

Little, R.J.A. (1988). Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values. Applied Statistics, 37, 23–38.

Lyberg, L., Biemer, P., Collins, M., Leeuw, E., de Dippo, C., Schwarz, N., and Trewin, D. (eds) (1997). Survey Measurement and Process Control. New York: John Wiley and Sons.

McLaren, C.H. and Steel, G.D. (2001). Rotation Patterns and Trend Estimation for Repeated Surveys Using Rotation Group Estimates. Statistica Neerlandica, 55, 220–237.

Merkouris, T. (2004). Combining Independent Regression Estimators from Multiple Surveys. Journal of the American Statistical Association, 99, 1131–1139.

Munger, G. and Lloyd, B.H. (1988). The Use of Multiple Matrix Sampling for Survey Research. Journal of Experimental Education, 56, 187–191.

Raghunathan, T.E. and Grizzle, J.E. (1995). A Split Questionnaire Design. Journal of the American Statistical Association, 90, 54–63.

Rahim, M.A. and Currie, S. (1993). Optimizing Sample Allocation for Multiple Response Variables. Proceedings of the American Statistical Association, Survey Research Methods Section, 346–351.

Rao, J.N.K. (1996). On Variance Estimation with Imputed Survey Data. Journal of the American Statistical Association, 91, 499–506.

Renssen, R.H. and 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. and Little, R.J.A. (1987). Statistical Analysis of Missing Data. New York: John Wiley and Sons.

Särndal, C., Swensson, B., and Wretman, J. (1992). Model Assisted Sampling. New York: Springer-Verlag.

Shao, J. and Sitter, R.R. (1996). Bootstrap for Imputed Survey Data. Journal of the American Statistical Association, 91, 1278–1288.

Shoemaker, D.M. (1973). Principles and Procedures of Multiple Matrix Sampling. Ballinger U.S.A.

Sitter, R.R. (1997). Variance Estimation for the Regression Estimator in Two-Phase Sampling. Journal of the American Statistical Association, 12, 780–787.

Srivastava, M.S. and Carter, E.M. (1986). The Maximum Likelihood Method for Non-Response in Sample Surveys. Survey Methodology, 12, 61–72.

Wretman, J. (1994). Estimation in Sample Surveys with Split Questionnaires. Research Report, University of Stockholm, 3, 1–11.

Received September 2006
Revised January 2009