

University of Wollongong

Research Online

---

Faculty of Engineering and Information  
Sciences - Papers: Part A

Faculty of Engineering and Information  
Sciences

---

1-1-2014

## Potential gains from using unit level cost information in a model-assisted framework

David G. Steel

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

Robert Graham Clark

*University of Wollongong*, [rclark@uow.edu.au](mailto:rclark@uow.edu.au)

Follow this and additional works at: <https://ro.uow.edu.au/eispapers>



Part of the [Engineering Commons](#), and the [Science and Technology Studies Commons](#)

---

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)

---

# Potential gains from using unit level cost information in a model-assisted framework

## Abstract

In developing the sample design for a survey we attempt to produce a good design for the funds available. Information on costs can be used to develop sample designs that minimise the sampling variance of an estimator of total for fixed cost. Improvements in survey management systems mean that it is now sometimes possible to estimate the cost of including each unit in the sample. This paper develops relatively simple approaches to determine whether the potential gains arising from using this unit level cost information are likely to be of practical use. It is shown that the key factor is the coefficient of variation of the costs relative to the coefficient of variation of the relative error on the estimated cost coefficients.

## Keywords

optimal allocation, optimal design, sample design, sampling variance, survey costs

## Disciplines

Engineering | Science and Technology Studies

## Publication Details

Steel, D. G. & Clark, R. Graham. (2014). Potential gains from using unit level cost information in a model-assisted framework. *Survey Methodology*, 40 (2), 231-242.

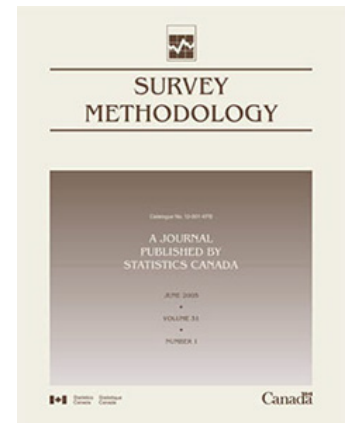
Catalogue no. 12-001-X  
ISSN 1492-0921

## Survey Methodology

# Potential gains from using unit level cost information in a model-assisted framework

by David G. Steel and Robert Graham Clark

Release date: December 19, 2014



Statistics  
Canada

Statistique  
Canada

Canada

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca).

You can also contact us by

**email** at [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca),

**telephone**, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- |                                                               |                |
|---------------------------------------------------------------|----------------|
| • Statistical Information Service                             | 1-800-263-1136 |
| • National telecommunications device for the hearing impaired | 1-800-363-7629 |
| • Fax line                                                    | 1-877-287-4369 |

## Depository Services Program

- |                  |                |
|------------------|----------------|
| • Inquiries line | 1-800-635-7943 |
| • Fax line       | 1-800-565-7757 |

## To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca), and browse by “Key resource” > “Publications.”

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.gc.ca](http://www.statcan.gc.ca) under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for  
Statistics Canada

© Minister of Industry, 2014

All rights reserved. Use of this publication is governed by the  
Statistics Canada Open Licence Agreement ([www.statcan.gc.ca/reference/copyright-droit-auteur-eng.htm](http://www.statcan.gc.ca/reference/copyright-droit-auteur-eng.htm)).

Cette publication est aussi disponible en français.

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

## Standard symbols

The following symbols are used in Statistics Canada publications:

- |                |                                                                                                                    |
|----------------|--------------------------------------------------------------------------------------------------------------------|
| .              | not available for any reference period                                                                             |
| ..             | not available for a specific reference period                                                                      |
| ...            | not applicable                                                                                                     |
| 0              | true zero or a value rounded to zero                                                                               |
| 0 <sup>s</sup> | value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded |
| p              | preliminary                                                                                                        |
| r              | revised                                                                                                            |
| x              | suppressed to meet the confidentiality requirements of the <i>Statistics Act</i>                                   |
| E              | use with caution                                                                                                   |
| F              | too unreliable to be published                                                                                     |
| *              | significantly different from reference category ( $p < 0.05$ )                                                     |

# Potential gains from using unit level cost information in a model-assisted framework

David G. Steel and Robert Graham Clark<sup>1</sup>

## Abstract

In developing the sample design for a survey we attempt to produce a good design for the funds available. Information on costs can be used to develop sample designs that minimise the sampling variance of an estimator of total for fixed cost. Improvements in survey management systems mean that it is now sometimes possible to estimate the cost of including each unit in the sample. This paper develops relatively simple approaches to determine whether the potential gains arising from using this unit level cost information are likely to be of practical use. It is shown that the key factor is the coefficient of variation of the costs relative to the coefficient of variation of the relative error on the estimated cost coefficients.

**Key Words:** Optimal allocation; Optimal design; Sample design; Sampling variance; Survey costs.

## 1 Introduction

Unequal unit costs have been reflected in sample designs by using simple linear cost models. In stratified sampling, a per-unit cost coefficient can sometimes be estimated for each stratum. The resulting allocation of sample to strata is proportional to the inverse of the square root of the stratum cost coefficients (Cochran 1977). In a multistage design the costs of including the units at the different stages of selection can be used to decide the number of units to select at each stage (Hansen, Hurwitz and Madow 1953).

While this theory is well established, unequal costs have not been used extensively in practice (Brewer and Gregoire 2009), perhaps because of a lack of good information on costs, and because of a focus on sample size rather than cost of enumeration. Groves (1989) argued that linear cost models are unrealistic, and that mathematical cost modelling can distract from more important decisions such as the mode of collection, the number of callbacks and how the survey interacts with other surveys conducted by the same organisation. Nevertheless, given the pressures on survey budgets, the final design should reflect costs and variance in a rational way, without being fixated on formal optimality.

Increasing use of computers in data collection is leading to more extensive and useful cost-related information on units on survey frames. In a programme of business surveys conducted by a national statistics institute, most medium and large enterprises will be selected in some surveys at least every year or two. This may provide information on costs for those businesses, for example some businesses may have required extensive follow-up or editing in a previous survey. Direct experience is less likely to be available for any given small business, but datasets of costs could be modelled to give predictions of likely costs.

---

1. David G. Steel, National Institute for Applied Statistics Research Australia, University of Wollongong, NSW Australia 2522. E-mail: [dsteel@uow.edu.au](mailto:dsteel@uow.edu.au); Robert Graham Clark, National Institute for Applied Statistics Research Australia, University of Wollongong, NSW Australia 2522. E-mail: [rclark@uow.edu.au](mailto:rclark@uow.edu.au).

Adaptive and responsive survey designs make use of paradata (process data) collected during a survey's operation, and auxiliary data known for the sampling frame (typically from administrative sources), to guide ongoing decisions. These may include the number of callbacks, which respondents to follow up, targeting of incentives, and choice of mode of collection for followup attempts (Groves and Heeringa 2006). In one example discussed by Groves and Heeringa (2006), interviewers designated non-respondents as having either low or high propensity to respond. The latter are less costly to convert to respondents, and a higher sampling fraction was assigned to them in a second phase of the survey. More recently, Schouten, Bethlehem, Beullens, Kleven, Loosveldt, Luiten, Rutar, Shlomo and Skinner (2012, Section 6) suggested that followup in the second phase of a survey should be designed to improve the R-indicator of non-response bias (defined in Schouten, Cobben and Bethlehem 2009; and in Schouten Shlomo and Skinner 2011). Peytchev, Riley, Rosen, Murphy and Lindblad (2010) argued that likely non-responders should be targeted with a different protocol from the very outset of a survey.

Thus, unequal unit costs can arise in practice, either for all units in advance of sampling, or for non-respondents who are to be targeted for followup. In either case, the collection and use of cost information incurs some expense and additional complexity. Moreover, effectively trading off cost and variance is only part of the picture, and response bias must also be considered. It is therefore important to understand whether the potential gains from using this information are worthwhile, particularly as any cost data is likely to be imperfect.

This paper develops relatively simple approximations to the gains arising from using unit level cost information in a model-assisted framework. Section 2 contains notation and some key expressions. Section 3 is concerned with the optimal design when cost parameters are known. Section 4 analyses the use of estimated unit costs, and Section 5 presents examples. Section 6 offers a discussion.

## 2 Notation and objective criterion

Consider a finite population,  $U$  containing  $N$  units, consisting of values  $Y_i$  for  $i \in U$ . A sample  $s \in U$  is to be selected using an unequal probability sampling scheme with positive probability of selection  $\pi_i = P[i \in s]$  for all units  $i \in U$ . A vector of auxiliary variables  $\mathbf{x}_i$  is assumed to be available either for the whole population, or for all units  $i \in s$  with the population total,  $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$ , also known. The auxiliary variables could consist of, for example, industry, region and size in a business survey, or age, sex and region in a household survey.

In the model-assisted approach (see for example Särndal, Swensson and Wretman 1992), the relationship between a variable of interest and the auxiliary variables is captured in a model, typically of the following form in single-stage surveys:

$$\left. \begin{aligned} E_M [Y_i] &= \beta^T \mathbf{x}_i \\ \text{var}_M [Y_i] &= \sigma^2 z_i \\ Y_i &\text{ independent of } Y_j \text{ for all } i \neq j \end{aligned} \right\} \quad (2.1)$$

where  $E_M$  and  $\text{var}_M$  denote expectation and variance under the model,  $\beta$  is a vector of unknown regression parameters,  $\sigma^2$  is an unknown variance parameter, and  $\mathbf{x}_i$  and  $z_i$  are assumed to be known for

all  $i \in U$ . Let  $E_p$  and  $var_p$  denote expectation and variance under repeated probability sampling with all population values held fixed.

The generalized regression estimator is a widely used model-assisted estimator of  $t_y$  :

$$\hat{t}_y = \sum_{i \in s} \pi_i^{-1} (y_i - \hat{\beta}^T \mathbf{x}_i) + \hat{\beta}^T \mathbf{t}_x \tag{2.2}$$

where  $\hat{\beta}$  may be a weighted or unweighted least squares estimate of the regression coefficients of  $y_i$  on  $\mathbf{x}_i$  using sample data. Estimators can also be constructed for nonlinear extensions to model (2.1), but in practice the linear model is almost always used.

The **anticipated variance** of  $\hat{t}_y$  is defined by  $E_M var_p [\hat{t}_y - t_y]$ , and is approximated by

$$E_M var_p [\hat{t}_y] \approx \sigma^2 \sum_{i \in U} (\pi_i^{-1} - 1) z_i \tag{2.3}$$

for large samples (Särndal et al. 1992, formula 12.2.12, p. 451) under model (2.1). Model-assisted designs and estimators should minimise  $E_M var_p [\hat{t}_y]$  subject to approximate design unbiasedness,  $E_p [\hat{t}_y] = t_y$ . Even if the model is incorrect, (2.2) remains approximately design-unbiased, although it will no longer have the lowest possible large sample anticipated variance. The anticipated variance has been used to motivate model-assisted sample designs in one stage (Särndal et al. 1992) and two stage sampling (Clark and Steel 2007; Clark 2009). One advantage of using the anticipated variance for this purpose is that it depends only on the selection probabilities and a small number of model parameters, which can be roughly estimated when designing the sample. In contrast,  $var_p [\hat{t}_y]$  typically depends on the population values of  $y_i$  and on joint probabilities of selection, both of which are difficult to quantify in advance.

The cost of enumerating a sample is assumed to be  $C = \sum_{i \in s} c_i$  where  $c_i$  is the cost of surveying a particular unit  $i$ . The values of  $c_i$  are usually assumed to be known. Typically  $c_i$  are also assumed to be constant for all units in the population, or constant within strata. With the generalization that  $c_i$  may be different for every unit  $i$ , the cost  $C$  depends on the particular sample  $s$  selected. The expected cost is  $E_p [C] = \sum_{i \in U} \pi_i c_i$ . The aim is to minimise the anticipated variance (2.3) subject to a constraint on the expected enumeration cost,

$$\sum_{i \in U} \pi_i c_i = C_f. \tag{2.4}$$

There will also be fixed costs that are not affected by the sample design and so do not have to be included here.

Some notation for population variances and covariances is needed. Consider the pairs  $(u_i, v_i)$ , and let  $S_{uv} = N^{-1} \sum_{i \in U} (u_i - \bar{u})(v_i - \bar{v})$  denote their population covariance, and  $S_u^2 = N^{-1} \sum_{i \in U} (u_i - \bar{u})^2$  denote the population variance of  $u_i$  ( $i = 1, \dots, N$ ). Let  $\bar{u}$  and  $\bar{v}$  be the population means of  $u_i$  and  $v_i$ . The population coefficient of variation of  $u_i$  is  $C_u = S_u / \bar{u}$ . The population relative covariance of  $(u_i, v_i)$  is  $C_{u,v} = S_{uv} / \bar{u} \bar{v}$ . A useful result is

$$\sum_{i \in U} u_i v_i = N \bar{u} \bar{v} (1 + C_{u,v}). \tag{2.5}$$

### 3 Optimal design with known cost and variance parameters

#### 3.1 Optimal Model-Assisted Design

The values of  $(\pi_i : i \in U)$  which minimise (2.3) subject to (2.4) are

$$\pi_i = C_f \frac{z_i^{1/2} c_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} c_j^{-1/2}} \propto z_i^{1/2} c_i^{-1/2} \quad (3.1)$$

and the resulting anticipated variance is

$$AV_{opt} = E_M \text{var}_p [\hat{t}_y] = \sigma^2 C_f^{-1} \left( \sum_{i \in U} c_i^{1/2} z_i^{1/2} \right)^2 - \sigma^2 \sum_{i \in U} z_i. \quad (3.2)$$

This can be easily derived using Lagrange multipliers or the Cauchy-Schwarz Inequality, and generalizes Särndal et al. (1992, Result 12.2.1, p. 452) to allow for unequal costs. Higher probability of selection is given to units which have higher unit variance or lower cost. However the square roots of  $z_i$  and  $c_i$  in (3.1) means that probabilities of selection do not vary dramatically in many surveys.

For the special case of stratified sampling where  $c_i = \bar{c}_h$  and  $z_i = \bar{z}_h$  for units  $i$  in stratum  $h$ , (3.1) becomes the usual optimal stratified allocation with  $\pi_i \propto \sqrt{\bar{z}_h / \bar{c}_h}$ , so that  $n_h \propto N_h \sqrt{\bar{z}_h / \bar{c}_h}$ .

It is assumed that the last term of (3.2), which represents the finite population correction, is negligible. Applying (2.5) gives:

$$AV_{opt} \approx \frac{\sigma^2 C_f^{-1} N^2 \bar{c} \bar{z} (1 + C_{\sqrt{c}, \sqrt{z}})^2}{(1 + C_{\sqrt{c}}^2)(1 + C_{\sqrt{z}}^2)} \quad (3.3)$$

where  $C_{\sqrt{c}}$  and  $C_{\sqrt{z}}$  refer to the population coefficients of variation of  $\sqrt{c_i}$  and  $\sqrt{z_i}$ , respectively. To make our results interpretable, we will assume that unit costs  $c_i$  and variances  $\sigma z_i$  are unrelated, so that  $C_{\sqrt{c}, \sqrt{z}} = 0$ . This assumption may not always be satisfied in practice, but any relationship between  $c_i$  and  $z_i$  will be specific to the particular example, and could be either positive or negative. To identify general principles, it makes sense to ignore any such relationship. In practice, it is often reasonable to also assume that  $C_{\sqrt{c}}$  and  $C_{\sqrt{z}}$  are small. A Taylor Series expansion then shows that  $C_c^2 \approx 4C_{\sqrt{c}}^2$  and  $C_z^2 \approx 4C_{\sqrt{z}}^2$ .

Putting these approximations together, (3.3) becomes

$$AV_{opt} = \frac{\sigma^2 C_f^{-1} N^2 \bar{c} \bar{z}}{\left(1 + \frac{1}{4} C_c^2\right) \left(1 + \frac{1}{4} C_z^2\right)}. \quad (3.4)$$

See the Appendix for details of these derivations.

#### Ignoring Costs

If the costs are ignored, then (3.1) suggests that  $\pi_i \propto z_i^{1/2}$ . To make comparisons for the same expected cost,  $C_f$ ,



$$\pi_i = C_f \frac{z_i^{1/2}}{\sum_{j \in U} z_j^{1/2} c_j} \tag{3.5}$$

with resulting anticipated variance

$$AV_{nocosts} = \sigma^2 C_f^{-1} \left( \sum_{i \in U} z_i^{1/2} \right) \left( \sum_{i \in U} c_i z_i^{1/2} \right) - \sigma^2 \sum_{i \in U} z_i. \tag{3.6}$$

Applying derivations similar to those used in Section 3.1,

$$AV_{nocosts} \approx \frac{\sigma^2 C_f^{-1} N^2 \bar{c} \bar{z}}{\left( 1 + \frac{1}{4} C_z^2 \right)}. \tag{3.7}$$

See Appendix for details. Comparing (3.7) and (3.4), we see that taking costs into account in the design results in dividing the anticipated variance by  $(1 + (1/4)C_c^2)$ .

### 4 The effect of using estimated cost parameters

In practice,  $c_i$  are not known precisely. Suppose that estimates  $\hat{c}_i = b_i c_i$  are used instead. Using the auxiliary variable and the estimated costs in the optimal probabilities implies  $\pi_i \propto z_i^{1/2} \hat{c}_i^{-1/2}$ . To make comparisons for the same expected costs,

$$\pi_i = C_f \frac{z_i^{1/2} \hat{c}_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j}.$$

The resulting anticipated variance is

$$AV_{ests} = \sigma^2 C_f^{-1} \left( \sum_{i \in U} \hat{c}_i^{1/2} z_i^{1/2} \right) \left( \sum_{j \in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j \right) - \sigma^2 \sum_{i \in U} z_i. \tag{4.1}$$

If we assume that the values of  $b_i$  are unrelated to the values of  $c_i$  and  $z_i$ , then

$$AV_{ests} = \sigma^2 C_f^{-1} \left( \sum_{i \in U} c_i^{1/2} z_i^{1/2} \right)^2 N^{-2} \left( \sum_{i \in U} b_i^{-1/2} \right) \left( \sum_{i \in U} b_i^{1/2} \right) - \sigma^2 \sum_{i \in U} z_i, \tag{4.2}$$

see Appendix for details. If the coefficient of variation of  $b_i$  is small, then a Taylor Series approximation gives  $N^{-2} \sum b_i^{-1/2} \sum b_i^{1/2} \approx 1 + (1/4)C_b^2$ . Applying this, and the same approximations as in Subsection 3.1, (4.2) becomes

$$AV_{ests} = \frac{\sigma^2 C_f^{-1} N^2 \bar{c} \bar{z} \left( 1 + \frac{1}{4} C_b^2 \right)}{\left( 1 + \frac{1}{4} C_c^2 \right) \left( 1 + \frac{1}{4} C_z^2 \right)}. \tag{4.3}$$

See Appendix for details.

Comparing (4.3) and (3.7), the effect of using estimated cost parameters rather than no costs at all is to multiply the anticipated variance by  $\left[1 + (1/4)C_b^2\right] / \left[1 + (1/4)C_c^2\right]$ . Therefore cost information is worth using provided  $C_b < C_c$ . The coefficient of variation of the error factors has to be less than that of the true unit costs over the population.

## 5 Examples of cost models

The key quantities determining the usefulness of the unit cost data are  $C_b$  and  $C_c$ . Optimal designs using unequal cost information are not very common, so there is relatively little literature on the typical values of these measures. Unequal costs may be driven by a variety of factors, including mode effects, geography and willingness to respond, and literature on these issues is helpful to give a rough idea of cost models that may apply in practice.

One reason why unequal per-unit costs may arise is the use of mixed mode interviewing. Different respondents may respond using different modes of collection, for example computer-assisted personal or telephone interviewing, mail or web questionnaires, or face to face interviewer (Dillman, Smyth and Christian 2009). This may be done to reduce cost or to improve response rate, however care must be taken that the approach does not introduce bias due to mode effects. Mode effects may consist of selection effects (which are generally not a problem) and measurement effects (which typically lead to bias), and the two are often hard to disentangle (Vannieuwenhuyze, Loosveldt and Molenbergs 2012). Cost savings from the use of mixed modes could potentially be magnified by incorporating mode costs into the sample design as described in this paper. Groves (1989, p. 538) compares per-respondent costs of telephone interviewing (\$38.00) and personal interviewing (\$84.90) of the general population. If the preference of all units on a frame was known, and half preferred each mode, this would imply  $C_c = 0.38$ . Greenlaw and Brown-Welty (2009) compared paper and web surveys, and found per-respondent costs of \$4.78 and \$0.64, respectively, in a survey of members of a professional association. In a mixed mode option, two thirds of respondents opted for the web option. If preferences are known in advance, then  $C_c = 0.76$ .

Another reason for varying costs is that some respondents are more difficult to recruit than others, requiring more visits or reminders. Groves and Heeringa (2006, Section 2.2) trialed a survey where interviewers classified non-respondents from the first approach as either likely or unlikely to respond. In subsequent follow-up, the first group had a response rate of 73.7% compared to 38.5% for the second group. This suggests that the per-respondent cost for the second group would be at least 1.9 times higher than the first group. (In fact, the ratio would be higher, because more follow-up attempts would be made for the difficult group.) If 50% of respondents are in both groups, then  $C_c = 0.31$ .

Geography is another source of differential costs in interviewer surveys. In the Australian Labour Force Survey, costs have been modelled as having a per-block component and a per-dwelling component (Hicks 2001, Table 4.2.1 in Section 4.2) depending on the type of area (15 types were defined). Assuming a constant 10 dwellings sampled per block, the net per-dwelling costs range from \$4.98 in Inner City Sydney and Melbourne to \$6.71 in Sparse and Indigenous areas. While this is a significant difference in costs across area types, the great majority of the population are in three area types (settled area, outer

growth and large town) where per-dwelling costs vary only between \$5.71 and \$6.07. As a result,  $C_c$  is estimated at a very small 0.054.

Table 5.1 shows the approximate percentage improvement in the anticipated variance from using estimated cost information for different values of  $C_c$  and  $C_b$ , some suggested by these examples. Negative values indicate that the design is less efficient than ignoring costs altogether. The table suggests that cost information is only worthwhile provided there is a fair variation in the unit costs, otherwise the benefit is very small, and can be erased when there is even small imprecision in the estimated costs. Mixed mode surveys have the most potential for exploiting varying unit costs in sample design, but the possibility of measurement bias would need to be carefully assessed in any such approach, using methods such as those in Vannieuwenhuyze, Loosveldt and Molenberghs (2010), Vannieuwenhuyze et al. (2012), Vannieuwenhuyze and Loosveldt (2013) and Schouten, Brakel, Buelens, Laan and Klaus (2013). It might even be possible to incorporate mode effects (or uncertainty about mode effects) into the optimal design via the variance model, and this may be the topic of future research. The findings made in this paper suggest that such an approach is worth considering.

**Table 5.1**

**Percentage improvement in anticipated variance from using estimated cost information compared to no cost information.**

Coefficient of Variation of Unit Costs ( $C_c$ ) (%)	Possible scenario	Coefficient of Variation of Error Factor ( $C_b$ ) (%)			
		0	10	25	50
5		0.1	-0.2	-1.5	-6.2
10	Interviewer travel due to remoteness	0.2	0.0	-1.3	-6.0
20		1.0	0.7	-0.6	-5.2
30	Response propensity	2.2	2.0	0.7	-3.9
40	Mixed mode (phone/personal int.)	3.8	3.6	2.3	-2.2
50		5.9	5.6	4.4	0.0
75	Mixed mode (paper/web self-complete)	12.3	12.1	11.0	6.8

## 6 Discussion

Incorporating unequal unit costs can improve the efficiency of sample designs. For the gains to be appreciable, the unit costs need to vary considerably. Even with no estimation error, a coefficient of variation of 50% may lead to a gain of only 6% in the anticipated variance. When this coefficient of variation is 75%, as can happen in a mixed mode survey, the reduction in the anticipated variance (or in the sample size for fixed precision) can be over 12%. Costs will be estimated with some error and this reduces the gain by a factor determined by the relative variation of the relative errors in estimating the costs at the individual level.

## Appendix

### A.1 Detailed derivations

**Lemma 1:** Let  $u_i$  be defined for  $i \in U$ . Let  $u_i = \bar{u} + \theta e_i$ , where  $\sum_{i \in U} e_i = 0$  and  $\theta$  is small. Then:

- $\sqrt{\bar{u}} = \sqrt{\bar{u}} - \frac{1}{8}\theta^2\bar{u}^{-3/2}S_e^2 + o(\theta^2)$ .
- $S_{\sqrt{\bar{u}}}^2 = \frac{1}{4}\theta^2\bar{u}^{-1}S_e^2 + o(\theta^2) = \frac{1}{4}\bar{u}^{-1}S_u^2 + o(\theta^2)$ .
- $N^{-2}\left(\sum_{i \in U} u_i^{1/2}\right)\left(\sum_{i \in U} u_i^{-1/2}\right) = 1 + \frac{1}{4}\theta^2\bar{u}^{-2}S_e^2 + o(\theta^2) = 1 + \frac{1}{4}C_u^2 + o(\theta^2)$ .
- $C_{\sqrt{\bar{u}}}^2 = \frac{1}{4}\theta^2\bar{u}^{-2}S_e^2 + o(\theta^2) = \frac{1}{4}C_u^2 + o(\theta^2)$ .

The notation  $o(C_u^2)$  can be used in place of  $o(\theta^2)$ , since  $C_u^2 = \theta^2 C_e^2$ . This will be done in the remainder of the Appendix.

**Proof:**

We start by writing  $\sqrt{\bar{u}}$  as a function of  $\theta$ :

$$\sqrt{\bar{u}} = N^{-1} \sum_{i \in U} \sqrt{u_i} = N^{-1} \sum_{i \in U} \sqrt{\bar{u} + \theta e_i}.$$

Call this  $g(\theta)$ , then differentiating about  $\theta = 0$  gives  $g(0) = \sqrt{\bar{u}}$ ,  $g'(0) = 0$  and

$$g''(0) = -\frac{1}{4}N^{-1}\bar{u}^{-3/2}\sum_{i \in U} e_i^2 = -\frac{1}{4}\bar{u}^{-3/2}S_e^2.$$

Hence

$$\sqrt{\bar{u}} = g(\theta) = g(0) + g'(0)\theta + \frac{1}{2}g''(0)\theta^2 + o(\theta^2) = \sqrt{\bar{u}} - \frac{1}{8}\theta^2\bar{u}^{-3/2}S_e^2 + o(\theta^2)$$

which is result a.

Result b is proven using result a:

$$\begin{aligned} S_{\sqrt{\bar{u}}}^2 &= N^{-1} \sum_{i \in U} (\sqrt{u_i})^2 - \left( N^{-1} \sum_{i \in U} \sqrt{u_i} \right)^2 \\ &= \bar{u} - \left( \sqrt{\bar{u}} \right)^2 \\ &= \bar{u} - \left( \sqrt{\bar{u}} - \frac{1}{8}\theta^2\bar{u}^{-3/2}S_e^2 + o(\theta^2) \right)^2 \\ &= \bar{u} - \left( \bar{u} + \frac{1}{64}\theta^4\bar{u}^{-3}S_e^4 - \frac{1}{4}\theta^2\bar{u}^{-1}S_e^2 + o(\theta^2) \right) \\ &= \frac{1}{4}\theta^2\bar{u}^{-1}S_e^2 + o(\theta^2) = \frac{1}{4}\bar{u}^{-1}S_u^2 + o(\theta^2). \end{aligned}$$

To derive c, we firstly write  $N^{-1} \sum_{i \in U} u_i^{-1/2}$  as a function  $g(\theta)$  of  $\theta$  and take a Taylor Series expansion:

$$\begin{aligned}
 N^{-1} \sum_{i \in U} u_i^{-1/2} &= N^{-1} \sum_{i \in U} (\bar{u} + \theta e_i)^{-1/2} \\
 &= g(\theta) = g(0) + g'(0)\theta + \frac{1}{2} g''(0)\theta^2 + o(\theta^2) \\
 &= \bar{u}^{-1/2} + 0\theta + \frac{1}{2} \frac{3}{4} \bar{u}^{-5/2} N^{-1} \sum_{i \in U} e_i^2 \theta^2 + o(\theta^2) \\
 &= \bar{u}^{-1/2} + \frac{3}{8} \bar{u}^{-5/2} S_e^2 \theta^2 + o(\theta^2)
 \end{aligned}
 \tag{A.1}$$

Note that  $N^{-1} \sum_{i \in U} u_i^{1/2} = \sqrt{\bar{u}}$ . Multiplying the expression for  $\sqrt{\bar{u}}$  in result a and (A.1) gives

$$\begin{aligned}
 N^{-2} \left( \sum_{i \in U} u_i^{1/2} \right) \left( \sum_{i \in U} u_i^{-1/2} \right) &= \left\{ \sqrt{\bar{u}} - \frac{1}{8} \theta^2 \bar{u}^{-3/2} S_e^2 + o(\theta^2) \right\} \left\{ \bar{u}^{-1/2} + \frac{3}{8} \bar{u}^{-5/2} S_e^2 \theta^2 + o(\theta^2) \right\} \\
 &= 1 + \frac{1}{4} \bar{u}^{-2} S_e^2 \theta^2 + o(\theta^2) \\
 &= 1 + \frac{1}{4} C_u^2 + o(\theta^2)
 \end{aligned}$$

which is result c.

For result d, firstly note that  $\sqrt{\bar{u}} = \sqrt{\bar{u}} + o(\theta)$  from result a, and so, from a first order Taylor Series,

$$\left( \sqrt{\bar{u}} \right)^{-2} = \left( \sqrt{\bar{u}} \right)^{-2} + o(\theta) = \bar{u}^{-1} + o(\theta).$$

Combining this with result b, we obtain

$$\begin{aligned}
 C_{\sqrt{u}}^2 &= S_{\sqrt{u}}^2 \left( \sqrt{\bar{u}} \right)^{-2} \\
 &= \left\{ \frac{1}{4} \theta^2 \bar{u}^{-1} S_e^2 + o(\theta^2) \right\} \left\{ \bar{u}^{-1} + o(\theta) \right\} \\
 &= \frac{1}{4} \theta^2 \bar{u}^{-2} S_e^2 + o(\theta^2) \\
 &= \frac{1}{4} C_u^2 + o(\theta^2)
 \end{aligned}$$

giving result d.

### Derivation of (3.3)

For the special case where  $u_i = v_i$ , (2.5) becomes

$$\sum_{i \in U} u_i^2 = N \bar{u}^2 (1 + C_u^2).
 \tag{A.2}$$

Applying (2.5),

$$\sum_{i \in U} c_i^{1/2} z_i^{1/2} = N \sqrt{c} \sqrt{\bar{z}} (1 + C_{\sqrt{c}, \sqrt{z}})
 \tag{A.3}$$

where  $\overline{\sqrt{c}} = N^{-1} \sum_{i \in U} \sqrt{c_i}$  and  $\overline{\sqrt{z}} = N^{-1} \sum_{i \in U} \sqrt{z_i}$ . Using (A.2), we can express  $\overline{\sqrt{c}}$  in terms of  $\bar{c}$ :

$$\bar{c} = N^{-1} \sum_{i \in U} c_i = N^{-1} \sum_{i \in U} (\sqrt{c_i})^2 = (\overline{\sqrt{c}})^2 (1 + C_{\sqrt{c}}^2). \quad (\text{A.4})$$

Similarly,

$$\bar{z} = (\overline{\sqrt{z}})^2 (1 + C_{\sqrt{z}}^2). \quad (\text{A.5})$$

Assuming the last term of (3.2) is negligible, applying (A.3), (A.4) and (A.5) gives (3.3).

#### Derivation of (3.4)

Lemma 1d implies that  $C_{\sqrt{c}}^2 = (1/4)C_c^2 + o(C_c^2) \approx (1/4)C_c^2$  and  $C_{\sqrt{z}}^2 = (1/4)C_z^2 + o(C_z^2) \approx (1/4)C_z^2$ . Result (3.4) follows from (3.3) by using these approximations, as well as assuming that  $C_{\sqrt{c}, \sqrt{z}} = 0$ .

#### Derivation of (3.7)

Firstly,  $\sum_{i \in U} c_i z_i^{1/2} = N\bar{c}\overline{\sqrt{z}}(1 + C_{c, \sqrt{z}})$ , from (2.5), where  $C_{c, \sqrt{z}}$  is the population relative covariance between the values of  $z_i^{1/2}$  and  $c_i$ . It is assumed that the values of  $c_i$  and  $z_i$  are unrelated, so that  $C_{c, \sqrt{z}} = 0$ . It is also assumed that the second term of (3.6) is negligible, corresponding to small sampling fraction. Hence (3.6) becomes:

$$AV_{nocosts} = \sigma^2 N^2 C_f^{-1} \bar{c} (\overline{\sqrt{z}})^2. \quad (\text{A.6})$$

From (A.5), and Lemma 1d, we have

$$(\overline{\sqrt{z}})^2 = \frac{\bar{z}}{1 + C_{\sqrt{z}}^2} \approx \frac{\bar{z}}{1 + (1/4)C_z^2}.$$

Substituting into (A.6) gives (3.7).

#### Derivation of (4.2)

Two terms in (4.1) will be simplified using (2.5). Firstly,

$$\begin{aligned} \sum_{i \in U} \hat{c}_i^{1/2} z_i^{1/2} &= \sum_{i \in U} b_i^{1/2} c_i^{1/2} z_i^{1/2} \\ &= N \left( N^{-1} \sum_{i \in U} b_i^{1/2} \right) \left( N^{-1} \sum_{i \in U} c_i^{1/2} z_i^{1/2} \right) + C_{\sqrt{b}, \sqrt{cz}} \end{aligned} \quad (\text{A.7})$$

where  $C_{\sqrt{b}, \sqrt{cz}}$  is the covariance between the population values of  $b_i^{1/2}$  and  $c_i^{1/2} z_i^{1/2}$ . Secondly,

$$\begin{aligned} \sum_{i \in U} z_i^{1/2} \hat{c}_i^{-1/2} c_i &= \sum_{i \in U} b_i^{-1/2} c_i^{1/2} z_i^{1/2} \\ &= N \left( N^{-1} \sum_{i \in U} b_i^{-1/2} \right) \left( N^{-1} \sum_{i \in U} c_i^{1/2} z_i^{1/2} \right) + C_{1/\sqrt{b}, \sqrt{cz}} \end{aligned} \quad (\text{A.8})$$

where  $C_{1/\sqrt{b}, \sqrt{cz}}$  is the covariance between the population values of  $b_i^{-1/2}$  and  $c_i^{1/2} z_i^{1/2}$ .

If we assume that the population values of  $b_i$  are unrelated to the values of  $c_i$  and  $z_i$ , so that  $C_{\sqrt{b}, \sqrt{cz}} = C_{1/\sqrt{b}, \sqrt{cz}} = 0$ , and substitute (A.7) and (A.8) into (4.1), then we obtain (4.2).

### Derivation of (4.3)

We can express (4.2) in terms of  $AV_{opt}$  which is defined in (3.2), assuming the last term of (3.2) is negligible, corresponding to small sampling fraction:

$$AV_{ests} \approx AV_{opt} N^{-2} \sum_{i \in U} b_i^{-1/2} \sum_{i \in U} b_i^{1/2} \quad (\text{A.9})$$

Lemma 1c implies that

$$N^{-2} \sum_{i \in U} b_i^{-1/2} \sum_{i \in U} b_i^{1/2} = 1 + \frac{1}{4} C_b^2 + o(C_b^2) \approx 1 + \frac{1}{4} C_b^2.$$

Substituting this, and (3.3), into (A.9) gives (4.3).

## References

- Brewer, K. and Gregoire, T.G. (2009). Introduction to survey sampling. In *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*, eds. Pfeffermann, D. and Rao, C.R., Amsterdam: Elsevier/North-Holland, pp. 9-37.
- Clark, R.G. (2009). Sampling of subpopulations in two-stage surveys. *Statistics in Medicine*, 28, 3697-3717.
- Clark, R.G. and Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 63-82.
- Cochran, W. (1977). *Sampling Techniques*. New York: Wiley, 3rd ed.
- Dillman, D., Smyth, J. and Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley and Sons, 3rd ed.
- Greenlaw, C. and Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost. *Evaluation Review*, 33, 464-480.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.
- Groves, R.M. and Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 439-457.
- Hansen, M., Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory Volume 1: Methods and Applications*. New York: John Wiley and Sons.

- Hicks, K. (2001). Cost and variance modelling for the 2001 redesign of the Monthly Population Survey. [www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.037](http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.037), *Australian Bureau of Statistics Methodology Advisory Committee Paper*.
- Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.
- Särndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80, 382-399.
- Schouten, B., Brakel, J.v.d., Buelens, B., Laan, J.v.d. and Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*.
- Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35 (1), 101-113.
- Schouten, B., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.
- Vannieuwenhuyze, J.T. and Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods and Research*, 42, 82-104.
- Vannieuwenhuyze, J.T., Loosveldt, G. and Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.
- Vannieuwenhuyze, J. T., Loosveldt, G., and Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, 80, 306-322.