INTERVIEWER ASSIGNMENTS WHICH MINIMIZE THE EFFECT OF NON-SAMPLING ERRORS

H. O. Hartley and Howard Monroe, Texas A&M University

1. Introduction

The importance of non-sampling, or measurement errors has long been recognized (for the numerous references see e.g., the comprehensive papers by Hansen, Hurvitz and Bershad (1961) and Bailar and Dalenius (1970). The abbreviated list of references given below obviously cannot claim to be exhaustive and even the papers listed cannot all be discussed in this short note. Briefly the various models suggested for such errors assume that a survey record (recorded content item) differs from its "true value" by a systematic bias, B, and various additive error contributions associated with various sources of errors such as interviewers, coders, etc. The important feature of these models is that the errors made by a specified error source (say a particular interviewer) are usually "correlated". These correlated errors contribute additive components to the total mean square error of a survey estimate which do not decrease inversely proportional to the overall sample size but only inversely proportional to the number of interviewers, coders, etc. Consequently, the application of standard text book formulas for the estimation of the variances of survey estimates may lead to serious underestimates of the real variability which should incorporate the non-sampling errors.

In this note we are concerned with the estimation of the "total variances" (including non-sampling components) of target parameter estimates. It is an almost trivial but important observation that the estimates of these "total variances" do not require separate estimates of the "elementary" non-sampling variance components if certain finite population corrections can be ignored. A number of references to this fact in special simple survey situations can be found in the early literature but it has been comprehensively exploited by Hartley and Rao (1978) and Hartley and Biemer (1978) who show that total variance estimates can be made for practically all survey designs with the help of component of variance estimation techniques.

The interviewer and coder assignments recommended by Hartley and Biemer had the aim of ensuring the estimability of all variance components required to estimate the variances of target parameters. In this paper we raise the question of optimizing the interviewer and coder assignments in order to minimize the effect of their correlated error variance components on the variances of the target parameters. Indeed the literature on the sampling errors of survey designs almost exclusively stresses the point that the design of the survey should result in estimators of minimum variances. Only secondary considerations are devoted to the estimability of variances (a case in point is the often used practice of one primary unit per stratum).

We confine ourselves here to a discussion of the assignment of interviewers and develop a restricted randomisation of interviewer assignments. It is common place knowledge that the restrictions which practical considerations impose will vary from survey to survey. Thus in certain surveys with (say) regional primaries it is vital to restrict an interviewer work-load entirely to within a single primary but the splitting of the primary work load into two or more groups of secondaries to be allocated to different interviewers is feasible. In other surveys reasons of economy dictate that the total work-load of a primary must be carried by a single interviewer (as, for example, if schools are used as primaries in an educational survey). We are here treating the latter case but our method is not restricted to it.

2. Model Assumptions for Non-Sampling Errors

Hartley and Biemer adopt "additive error models" (also used in the more recent literature) in which the error made by (say) a particular interviewer are correlated through an additive error term. They assume that the true content item of the t^{th} respondent interviewed by interviewer i and coded by coder c has the following additive non-sampling errors.

$$\begin{aligned} &\text{Interviewer error} = b_i + \delta b_t \\ &\text{Coder error} \quad = c_c + \delta c_t \qquad (1) \\ &\text{Respondent error} = \delta r_t \end{aligned}$$

where

b_i = error variable contributed by i^{th} interviewer common to all units, t, interviewed by i^{th} interviewer

c_c - error variable contributed by c^{th} coder common to all units, t, coded by c^{th} coder.

δb_t, δc_t, and δr_t = elementary interviewer, coder, and respondent errors afflicting the content item of unit t (respondent t).

They further assume that the b_i, c_c, δb_t, δc_t are random samplesfrom infinite populations with zero mean and variances σ_b^2, σ_c^2, $\sigma_{\delta b}^2$, $\sigma_{\delta c}^2$ and δr_t a (nonobserved) error with zero mean sampled from the finite population of respondents by the survey design implemented.

3. The type of survey covered.

Similarly to Hartley and Biemer the type of survey here covered is a two stage stratified survey. However, it is necessary to modify the component of variance estimation slightly. Denote by η_{hps} the true content item for secondary s of primary p in stratum h. Denote by y_{hps} the corresponding recorded content item. Then we clearly have $y_{hps} = \eta_{hps}$ + total error in (1) and if we replace the unit label, t, by the triple subscript hps this equation can be written in the form

$$y_{hps} = \bar{\eta}_{h..} + (\bar{\eta}_{hp.} - \eta_{h..}) + b_i + c_c + (\eta_{hps} - \bar{\eta}_{hp.}) + \delta r_{hps} + \delta b_{hps} + \delta c_{hps} \qquad (2)$$

where the $\bar{\eta}_{hp.}$ and $\bar{\eta}_{h..}$ are respectively the primary and strata population means of the true content items. We now combine the terms in the second and third lines of (2) and denote them by e_{hps}. We confine ourselves to the special case when secondaries are drawn with equal probability and the secondary fpc's are negligible. Then the e_{hps} are random samples from infinite populations with variances $\sigma_e^2(h,p)$ (say) and the variance of the target parameter estimates only depend on $\sigma_e^2(h,p)$. In this case the variance components σ_b^2, σ_c^2 and σ_e^2 can be estimated by standard methods of component of variance estimation. The method is analogous to that of Hartley and Biemer (1979) Appendix 1.

One of our estimability conditions (see Section 4) will stipulate that all secondaries in a primary are handled by one interviewer and one coder. Consequentially to invoke the components of variance estimation procedure it is convenient to average (2) over the secondary units and obtain

$$\bar{y}_{hp.} = \eta_{h..} + b_i + c_c + \delta_{hp} + \bar{e}_{hp.} \qquad (3)$$

where $\delta_{hp} = (\bar{\eta}_{hp.} - \bar{\eta}_{h..})$. Finally, in order to give a concise description of the formulas it is convenient to rewrite (3) in matrix motation

$$y = X\bar{\eta} + U_b b + U_c c + \sum_h W_h (\delta_h + e_h) \qquad (4)$$

where y, $\bar{\eta}$, b, c, δ_h and e_h are the vectors of the terms in (3) and X, U_b, U_c, and W_h are the corresponding design matrices. Notice that the W_h consists of identity matrices for the primaries in other strata.

4. Estimability conditions

Hartley and Biemer (p. 258) give sufficient conditions for the estimability of all components of a variance for the case in which it is feasible for two different interviewers to be allocated to different secondaries in the same primary. In this paper we consider the situation in which (because of practical limitations) only one interviewer must carry the whole work load in a primary. Sufficient conditions for the estimability of all components of variance are then as follows:

(i) The sample contains at least two primaries per stratum and two secondaries per primary.

(ii) All secondaries in a primary are interviewed by the same interviewer and coded by the same coder.

(iii) In at least one stratum there are at least two primaries entirely interviewed by different interviewers but coded by the same coder.

(iv) In at least one stratum there are at least two primaries entirely coded by different coders.

The above are sufficient conditions, however a more reliable estimate of σ_b^2 based on more interviewer contrasts is obtained if (iii) is replaced by the more restrictive condition.

(iii)' If the number of primaries in stratum h is n_h then in all strata there must be at least two primaries interviewed by the same interviewer and the remaining n_h-2 primaries (if any) by n_h-2 different interviewers.* There must be at least one stratum with $n_h \geq 3$. The condition (iii)' will certainly provide more within stratum interviewer contrasts for the estimation of σ_b^2 but it is difficult to assess the reduction in the variance of $\hat{\sigma}_b^2$ through replacing (iii) by (iii)'.

5. The variance of the estimate of target parameters.

The majority of target parameters - including population totals and means - which are computed from sample survey data are linear functions of the y_{hps}. Since sampling within primaries is with equal probabilities, the discussion is confined to estimators of the form

$$\hat{Y} = \gamma'\bar{y} \qquad (5)$$

where $\bar{y}$ is the vector of primary means and γ is a coefficient vector which may depend upon the set P of primaries in the sample. It is easily shown that $\hat{Y}$ is unbiased for the target parameter if $\gamma'\dot{\eta}$ is unbiased where $\dot{\eta}$ is the vector of true primary means, $\bar{\eta}_{hp}$.

Let G be the set of sampled primaries P and the interviewer-coder work assignments. The variance of $\hat{Y}$ is composed of two components as follows:

$$\text{Var } \gamma'\bar{y} = \underset{G}{\text{Var}}\, E|_G \gamma'\bar{y} + \underset{G}{E}\, \text{Var}|_G \gamma'\bar{y} \qquad (6)$$

where $\text{Var}|_G$ and $E|_G$ denote the variance and expectation, respectively, given the set G and Var_G and E_G denote the variance and expectation over all possible sets G.

An unbiased estimator of (6) is derived in Biemer (1978) by estimating each of the two components separately. The resulting estimation formula is

$$\text{Var } \gamma'\bar{y} = \bar{y}'\Omega\bar{y} - \text{tr } \Omega\hat{\Sigma} + \gamma' U_b U_b' \gamma \hat{\sigma}_b^2$$
$$\gamma' U_c U_c' \gamma \hat{\sigma}_c^2 + \sum_h \gamma' W_h \hat{D}_h W_h' \gamma \tag{7}$$

where

Ω = a constant matrix for given set P which is directly provided by standard finite population sampling theory without non-sampling errors and satisfies $E(\dot{\eta}'\Omega\dot{\eta}) = \text{Var } \gamma'\dot{\eta}$,
G

$$\hat{\Sigma} = U_b U_b' \hat{\sigma}_b^2 + U_c U_c' \hat{\sigma}_c^2 + \Sigma\, W_h \hat{D}_h W_h' , \tag{8}$$

$\hat{\sigma}_b^2, \hat{\sigma}_c^2$ are computed as in Hartley and Biemer Appendix 1 and

$$\hat{\sigma}_e^2(j,p) = \sum_{s=1}^{m(h,p)} \frac{(y_{hps} - \bar{y}_{hp.})^2}{m(h,p)-1} \tag{9}$$

where the m(j,p) are the number of secondaries in the primary labeled (j,p) and

$$\hat{D}_h = \text{diag}\left\{ \frac{\sigma_e^2(h,p)}{m(h,p)} \right\} \tag{10}$$

6. The optimization of the interviewer assignment.

By taking expectations of (7) the component of variance of Var $\gamma'\bar{y}$ which depends on σ_b^2 is given by

$$\text{Comp}_b\{\text{Var } \gamma'\bar{y}\} = \{\gamma' U_b U_b' \gamma\}\sigma_b^2 = (U_b'\gamma)'(U_b'\gamma)\sigma_b^2 \tag{11}$$

and this will be minimized (whatever the values of any of the variance components) if the sum of squares $(U_b'\gamma)'(U_b'\gamma)$ is minimized. Denoting the the elements of γ by γ_{hp} the $U_b'\gamma$ are the "interviewer totals" of the γ_{hp} . The latter are the "jack up factors" to be applied to the $\bar{y}_{hp}$ and are therefore predetermined by the survey design. For example, if both stages are with equal probability $\gamma_{hp} = N_h M_{hp}/n_h$. This minimization is to be constrained by the estimability condition (iii)'. Moreover as a practical consideration we would normally prefer to restrict the optimisation by assigning each interviewer "approximately" an equal number of primaries, a concept which is discussed below. Since the total sum of squares of the γ_{hp} is given, the minimization of the between interviewer sum of squares is of course equivalent to the maximization of the within interviewer sum of squares. In order to satisfy the first part of condition (iii)' optimally we pool for each stratum h the largest and the smallest of the γ_{hp}, denote these totals by $\dot{\gamma}_{hp}$ with a p-label corresponding to (say) the smaller of the two p and will assign the <u>total</u> $\dot{\gamma}_{hp}$ to some interviewer. The remaining γ_{hp} are kept separate $\dot{\gamma}_{hp}$ totals. The number of $\dot{\gamma}_{hp}$ is therefore given by $\nu = \Sigma\, (n_h - 1)$. The second condition of (iii)' now stipulates that all $\dot{\gamma}_{hp}$ with the same stratum index h must all be allocated to different interviewers. We now specify the "approximately even" allocation to mean that each interviewer's quota of assigned $\dot{\gamma}_{hp}$ should only fractionally differ from their average. Thus if

$$k = \left[\frac{\sum_h (n_h - 1)}{I} \right] \tag{12}$$

then

$$k \leq \text{interviewer allocation of primaries} \leq k+1 \tag{13}$$

Denote by $\dot{\gamma}(h,i)$ <u>that</u> $\dot{\gamma}_{hp}$ in stratum h which is allocated to interviewer i then the task is to find <u>that</u> allocation which minimizes

$$S^2 = \sum_i (\sum_h \dot{\gamma}(j,i)\}^2 = \sum_i \dot{\gamma}(.,i)^2 \tag{14}$$

Introduce

$$H = \#\text{ of strata} \tag{15}$$

then it is convenient for the algorithm to construct an H·I double array of $\dot{\gamma}$ (h,i) in which h = 1, ..., H and i=1, ..., I.

In each stratum h the I - n_h + 1 "excess positions" all filled by "marked dummies" $\dot{\gamma}(h,i) = 0$ (see Fig. 1 below in which an example allocation of example $\dot{\gamma}(h,i)$ is exhibited for H = 10 strata and I = 5 interviewers).

Figure 1.

<u>Initial Interviewer Assignment</u>

h=	Stratum										Total
i=	1	2	3	4	5	6	7	8	9	10	$\dot{\gamma}(.,i)$
1	4	0	7	0	0	0	3	0	0	4	18
2	0	4	3	0	0	0	0	3	0	2	12
3	0	2	0	0	0	3	2	0	1	0	8
4	0	1	0	0	2	1	0	2	0	0	6
5	0	1	0	1	0	1	0	1	0	0	4
$(n_h - 1)$ =	1	4	2	1	1	3	2	3	1	2	

In the above example $S^2 = 18^2 + 12^2 + 8^2 + 6^2 + 4^2 = 584$, a high value in spite of the fact that precisely k = 4 values of $\dot{\gamma}(h,i)$ are allocated to each interviewer. The algorithm will commence with an "initial arrangement" such as given in Fig. 1 satisfying (iii)' and (13) constructed as shown below. The "improving algorithm" will then attempt all possible double swaps of $\dot{\gamma}(h,i)$ with $\dot{\gamma}(h,j)$ and $\dot{\gamma}(\ell,j)$ with $\dot{\gamma}(\ell,i)$ which reduce S^2 and at the same time preserve the above conditions. It is programmed as a quadruple loop in h, ℓ; i, j and no double swap is made if

(a) either $\dot{\gamma}(h,i)$ or $\dot{\gamma}(\ell,j) = 0$

(b) if $h \neq \ell$ and if either $\dot{\gamma}(\ell,i) \neq 0$ or $\dot{\gamma}(h,j) \neq 0$

(c) if
$$\frac{1}{2}\Delta S^2 = \{\dot{\gamma}(.,i)-\dot{\gamma}(.,j)-\dot{\gamma}(h,i)+\dot{\gamma}(\ell,j)\}(\dot{\gamma}(h,i)-\dot{\gamma}(\ell,j)) \geq 0 \tag{16}$$

Clearly with the above conditions $\Delta S^2 < 0$ with every legitimate double swap. If and when for a full quadruple loop no legitimate swap was found with $\Delta S^2 < 0$ the algorithm will terminate usually at the global minimum of S^2. However it is not possible to prove this in general. In the above example the algorithm did reach the global minimum as shown in Fig. 2.

Figure 2.
Terminal Interviewer Assignment

	Stratum										
i= \ h=	1	2	3	4	5	6	7	8	9	10	$\dot{\gamma}(.,i)$
1	0	4	3	1	2	0	0	0	0	0	10
2	4	1	0	0	0	1	0	3	0	0	9
3	0	1	0	0	0	0	3	2	0	4	10
4	0	2	0	0	0	3	2	0	0	2	9
5	0	0	7	0	0	1	0	1	1	0	10

The terminal value of $S^2 = 462\sigma_b^2$ is the minimum value of Comp (Var $\gamma'y$) . The arrangement of Fig. 1, (although balanced with regard to interviewer load) would have resulted in a variance 26.4% in excess of the optimum.

We now turn to the construction of the "initial arrangement" satisfying both (iii)' and (13) and at the same time giving the global minimum a finite probability to be selected. There are $\binom{I}{n_h-1}$ ways of selecting n_h-1 interviewers out of the I interviewers for stratum h and $C = \prod_h \binom{I}{n_h-1}$ possible selections of interviewers to fill the "positions" in the strata columns. Everyone of these selections has a probability of 1/C to be selected but most of them will violate (13). Denote by k_i the number of assignments to interviewer i then if (13) is violated we have $(k_{max} - k_{min}) > 1$. Denote by i_{max} and by i_{min} two interviewers with k_{max} and k_{min} assignments respectively, then there must be a stratum in which i_{max} is assigned a position but i_{min} is not. Swap this assignment from i_{max} to i_{min}. This will result in an arrangement in which the number of interviewers with k_{max} assignments and the number of interviewers with k_{min} assignments are both reduced by 1. Continue the swapping process until the number of interviewers with k_{max} assignments and/or the number of interviewers with k_{min} assignments is zero. The new value of $k_{max}-k_{min}$ will then be reduced by at least 1. Continue the process until $k_{max}-k_{min} \leq 1$ and hence (13) is satisfied.

The swapping procedure is illustrated in Figure 3 below in which the initial assignments are $k_1 = 6$, $k_2 = 5$, $k_3 = k_4 = k_5 = 3$ and three swaps are made transferring sequentially the positions $\boxed{x}^-$, $\textcircled{x}^-$ and $\diamond x^-$ to positions $\boxed{x}^+$, $\textcircled{x}^+$ and $\diamond x^+$.

Figure 3.

Illustration of Swapping Procedure

	Stratum										Swap #			
											0	1	2	3
h= \ i=	1	2	3	4	5	6	7	8	9	10	k_i	k_i	k_i	k_i
1	x	x	$\diamond x^-$			x	$\boxed{x}^-$	x			6	5	5	4
2		x	x		x	x		$\textcircled{x}^-$			5	5	4	4
3		x		x			$\boxed{x}^+$		x		3	4	4	4
4		x					x	$\textcircled{x}^+$		x	3	3	4	4
5			$\diamond x^+$			x		x		x	3	3	3	4

This procedure will increase the probability of $\frac{1}{C}$ for any particular legitimate assignment to be selected to $\frac{1}{C} + q$ with $q \geq 0$. The above procedure of selecting an "initial arrangement" followed by the improving algorithm will result in a procedure in which the global minimum of S^2 is reached with a finite probability.

All the above operations are automated in a computer program written by one of us (H.M.) covering up to 50 strata and up to 10 interviewers. This is available on request.

References

Bailar, B.A. and Dalenius, T. (1970). "Estimating the response variance components of the U.S. Bureau of the Census survey model." Sankhya Series B, 341-360.

Battese, G.E., Fuller, W.A., and Hickman, R.D. (1976). "Estimation of reponse variance from intervier re-interview surveys." Journal Indian Society of Agricultural Statistics, 28, 1-14.

Biemer, P. P. (1978). "The estimation of non-sampling variance components in sample surveys," unpublished Ph.D. dissertation, Institute of Statistics, Texas A&M University.

Cochran, W. G. (1968). "Errors of measurements in statistics." Technometrics 10, 637-666.

Fellegi, I.P. (1974). "An improved method of estimating the correlated response variance". Journal of American Statistical Assn., 1969, 496-501.

Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961). "Measurement errors in censuses and surveys." Bull. International Stat. Inst., 38, 359-374.

Hansen, M.H., Hurwitz, W.N., Marks, E.S. and Mauldin, W.P. (1951). "Response errors in surveys", JASA, 46, 147-90.

Hanson, R.H. and Marks, E.S. (1958). "Influence of the interviewer on the accuracy of survey results", JASA, 53, 635-55.

Hartley, H.O. and Rao, J.N.K. (1978). "The estimation of non-sampling variance components in sample surveys," in Survey Sampling and Measurement, New York: Academic Press.

Hartley, H.O. & Biemer, P.P. (1978) "The estimation of non-sampling variances in current surveys", Proc. Survey Research Methods of ASA.

Koch, G.G. (1973). "Some survey designs for estimating response error model components", Tech. Report #5 (21U-730) Research

Koch, G.G. (1971). "A reponse error model for a simple interviewer structure situation" Tech. Report #4 (SU-618) Research Triangle Institute.

Mahalanobis, P.C. (1946). "Recent experiments in statistical sampling in the Indian Statistical Institute", JRSS, 109, 325-70.