Optimal Dual Frame Sample Design: Results of a Simulation Study

Paul P. Biemer, U.S. Bureau of the Census

1. INTRODUCTION

One promising solution to the problem of noncoverage of the population without telephones for random-digit-dial (RDD) sampling is the "dual frame survey" method. This method combines information collected in a centralized telephone survey utilizing a telephone frame and an areal survey utilizing an area/list frame to form cost efficient estimators of population parameters. The areal survey provides the missing nontelephone population component for the telephone survey.

Some issues of major concern to the Bureau of the Census relate to the quality of the dual frame estimates. Recent RDD studies conducted by the Bureau have demonstrated that RDD response rates are considerably lower than for area/list surveys. Since interviewer workload sizes are much larger in centralized telephone interviewing than in decentralized interviewing, telephone interviewer variance could be a problem. And, the yet unknown nonresponse biases and mode (or method) effects raise doubts about whether the quality of the data will be enhanced or diminished with dual frame.

These issues suggest that the goal of dual frame survey design should be to minimize the total mean square error (MSE), and not simply the variance, of a survey estimator within the constraints of the budget. Biemer (1983) provides a general methodology for minimizing the MSE in the design of dual frame surveys, applicable to the complex sampling schemes typically used in Bureau current surveys. The key estimation and variance formulae are summarized in Section 2 of this paper.

Although the concept of a dual frame survey is simple, the sample design issues can be very complex. The major objective of this work is to provide some insight to the survey designer on some key elements of optimal dual frame design. Since the complexity of the optimization formulas preclude most analytical investigations, simulation was used to study the effects of a few of the many parameters which are input into dual frame design optimization. To begin our analysis, the MSE optimization procedure described in Biemer (1983) is applied for data from the Current Population Survey (CPS). To better understand results of this application, a number of simulations are conducted. Some issues addressed in the paper are:

- The importance of nonsampling biases in the decision to adopt a dual frame survey approach.
- The performance of four possible dual frame estimators for the CPS data set.
- The role of telephone availability, cost, and sampling and nonsampling error components on dual frame allocation decisions.
- The loss in relative efficiency of dual frame estimators over subpopulations which are less accessible by the telephone.

2. FORMULAS

We shall refer to the area/list frame survey as Survey A and the telephone frame survey as Survey B. The objective of the dual frame survey is to estimate the mean, $\overline{Y}$, of some characteristic y for a population of M elements.

Survey A: The sample design is a stratified two-stage design where the secondary sampling units are area segments of dwellings. The segments are selected by an equal probability without replacement method (EPSEM) while the primary units are selected with unequal probabilities. We assume that the within primary sampling fractions are such that the overall probability of selecting a dwelling in a stratum is the same for all dwellings in the stratum. Interviewer assignments are composed of segments which are randomly selected within a primary. Each interviewer is assigned approximately the same number of dwellings.

Survey B: The frame is a list of telephone numbers. For simplicity, it is assumed that each population element may be linked to, at most, one telephone number. The sample is selected completely independently of Survey A using either a stratified or unstratified two-stage design (such as Waksberg's (1978) random-digit-dialing method). The stratified design considered assumes the same strata definitions as for Survey A. The secondary sampling units are telephone numbers sampled with EPSEM without replacement within each primary. Like Survey A, we assume that the within primary sampling fractions are such that the overall probability of selecting a telephone number linked to a dwelling (referred to as a residential telephone number) is the same for all residential telephone numbers in a stratum. Interviewer assignments are made up of telephone numbers which are randomly assigned without regard to primary or stratum boundaries.

Notation: Let D_1 refer to the elements in the population which belong only to Frame A and let D_2 denote elements belonging to both frames (i.e., elements linked to residential telephone numbers). Four estimators of $\overline{Y}$ are considered. These estimators and their variances are derived and discussed in detail in Biemer (1983). The notation here is somewhat simplified.

The following table defines the symbols used in the mean square error (MSE) formulae. The subscript h indicates stratum h (h=1,...,L) while absence of the subscript indicates the symbol is defined for the entire population or sample.

NOTATION

Description of Symbol	D_1	D_2	Total
Proportion of elements in domain	T_{1h}	T_{2h}	1
	T_1	T_2	1
Stratum weight			W_h
Population mean	$\bar{Y}_{1h}$	$\bar{Y}_{2h}$	$\bar{Y}_h$
Total nonsampling bias: Survey A	B_{1h}	B_{2h}	
Survey B		B_{Bh}	B_B
Population mean plus nonsampling bias: Survey A	$\bar{X}_{1h}$	$\bar{X}_{2h}$	
Survey B		$\bar{X}_{Bh}$	$\bar{X}_B$
Survey A: between primary sampling variance	S^2_{1h}	S^2_{2h}	S^2_{Ah}
within primary primary srs variance	σ^2_{1h}	σ^2_{2h}	σ^2_{Ah}
within primary design effect	δ_{1h}	δ_{2h}	δ_{Ah}
Survey B: srs variance		σ^2_{Bh}	σ^2_B
design effect (all stages)		δ_{Bh}	δ_B
Interviewer correlation coefficient: Survey A			ρ_{Ah}
Survey B			ρ_B
Ratio sample mean per element: Survey A	$\bar{x}_{1h}$	$\bar{x}_{2h}$	$\bar{x}_{Ah}$
Survey B		$\bar{x}_{Bh}$	$\bar{x}_B$
Number of sample primary units: Survey A			n_{Ah}
Survey B		n_{Bh}	n_B
Average number of elements per primary: Survey A			$\bar{\bar{m}}_A$
Survey B			$\bar{\bar{m}}_B$
Interviewer assignment size: Survey A			q_A
Survey B			q_B

Let $\theta_h (0 < \theta_h < 1)$ denote a constant, suggested by Hartley (1962), which is to be optimized. For the case where T_{2h} is known approximately, say from previous Census data, defined the "separate" estimator of $\bar{Y}$ as

$$\bar{x}_s = \sum_h^L W_h(T_{1h}\,\bar{x}_{1h} + \theta_h\, T_{2h}\,\bar{x}_{2h} + (1-\theta_h)\, T_{2h}\,\bar{x}_{Bh}). \qquad (2.1)$$

For some populations, T_{2h} may not be known exactly and other estimators of $\bar{Y}$ may be preferred which may have smaller MSE than $\bar{x}_s$. It may be possible to use information from data banks on the telephone population or combine data from other ongoing surveys to estimate T_{2h}. If no information outside the survey is available, an estimator of T_{2h} which is sometimes used (e.g., see Casady, Snowden and Sirken (1981)) is $t_{2h} = m_{2h}/m_{Ah}$ and (2.1) becomes, for $t_{1h} = 1 - t_{2h}$

$$\bar{x}'_s = \sum_h^L W_h(t_{1h}\,\bar{x}_{1h} + \theta_h\, t_{2h}\,\bar{x}_{2h} + (1-\theta_h)\, t_{2h}\,\bar{x}_{Bh}) \qquad (2.2)$$

In our analyses, only the simplest forms of the MSE's derived by Biemer will be investigated. For known T_{2h},

$$MSE(\bar{x}_s) \simeq [\sum_h^L W_h(b_{Ah}(\theta_h) + b_{Bh}(\theta_h))]^2 + \sum_h^L W_h^2[n_{Ah}^{-1}\, v_{Ah}(\theta_h) + n_{Bh}^{-1}\, v_{Bh}(\theta_h)] \qquad (2.3)$$

where

$$b_{Ah}(\theta_h) = T_{1h}\, B_{1h} + \theta_h\, T_{2h}\, B_{2h}$$

$$b_{Bh}(\theta_h) = (1-\theta_h)\, T_{2h}\, B_{Bh}$$

$$v_{Ah}(\theta_h) = \theta_h\, S^2_{Ah} + (1-\theta_h)\, T_{1h}\, S^2_{1h} - \theta_h(1-\theta_h)\, T_{2h}\, S^2_{2h} + \bar{\bar{m}}_A^{-1}[T_{1h}\,\delta_{1h}\,\sigma^2_{1h} + \theta_h\, T_{2h}\,\delta_{2h}\,\sigma^2_{2h} + (T_{1h} + \theta_h\, T_{2h})^2 \frac{\rho_{Ah}}{1-\rho_{Ah}}\,\sigma^2_{Ah}\, q_A],$$

and

$$\bar{\bar{m}}_B\, v_{Bh}(\theta_h) = (1-\theta_h)^2\, [T^2_{2h}\,\delta_{Bh}\,\sigma^2_{Bh} + T_2^2 \frac{\rho_B}{1-\rho_B}\,\sigma^2_{Bh}\, q_B/L].$$

When T_{2h} is estimated by t_{2h}, then $MSE(\bar{x}'_s)$ is approximately given by (2.3) with $v_{Ah}(\theta_h)$ replaced by

$$v'_{Ah}(\theta_h) = v_{Ah}(\theta_h) + \bar{\bar{m}}_A^{-1}\, u_{Ah}(\theta_h) \qquad (2.4)$$

where

$$u_{Ah}(\theta_h) = T_{1h}T_{2h}\, \{\Phi_{Ah}(\bar{X}_{1h} - \theta_h\bar{X}_{2h} - (1-\theta_h)\bar{X}_{Bh})^2 - (\Phi_{Ah} - \phi_{Ah})(\bar{X}_{1h} - \theta_h\,\bar{X}_{2h})^2\}.$$

ϕ_{Ah} is the Survey A average within primary design effect and Φ_{Ah} is the total design effect associated with t_{2h} in stratum h.

Another class of estimator considered in the analysis is the so-called combined estimator, denoted by $\bar{x}_c$. Here we suppose the Survey B sample is not explicitly stratified and therefore uses only one θ-parameter. For known T_{2h},

$$\bar{x}_c = \sum_h^L W_h(T_{1h}\,\bar{x}_{1h} + \theta\, T_{2h}\,\bar{x}_{2h}) + (1-\theta)\, T_2\,\bar{x}_B \qquad (2.5)$$

where $\bar{x}_B$ is the ratio mean of all observations in the Survey B sample. We denote the estimator $\bar{x}_s$ by $\bar{x}_s'$ when T_{2h} is estimated by t_{2h} and T_2 by $t_2 = \sum_h W_h\, t_{2h}$.

The MSE of $\bar{x}_c$ is given by

$$MSE(\bar{x}_c) = [\sum_h^L W_h\, b_{Ah}(\theta) + b_B(\theta)]^2 + \sum_h^L W_h^2\, n_{Ah}^{-1}\, v_{Ah}(\theta) + n_B^{-1}\, v_B(\theta) \qquad (2.6)$$

where $b_{Ah}(\theta) = b_{Ah}(\theta_h)$ and $v_{Ah}(\theta) = v_{Ah}(\theta_h)$ defined in (2.3) with $\theta_h = \theta$, h=1,...L,

$$b_B(\theta) = (1-\theta)\, T_2\, B_B$$

and

$$\bar{\bar{m}}_B\, v_B(\theta) = (1-\theta)^2\, T_2\, \sigma_B^2(\delta_B + \frac{\rho_B}{1-\rho_B} q_B).$$

For $MSE(\bar{x}_c')$, replace $v_{Ah}(\theta)$ by $\dot{v}_{Ah}(\theta)$

where

$$\dot{v}_{Ah}(\theta) = v_{Ah}(\theta) + \bar{\bar{m}}_A\, \dot{u}_{Ah}(\theta), \qquad (2.7)$$

$$\dot{u}_{Ah}(\theta) = T_{1h}T_{2h}\, \{\Phi_{Ah}(\bar{X}_{1h} - \theta\,\bar{X}_{2h} - (1-\theta)\bar{X}_B)^2 - (\Phi_{Ah} - \phi_{Ah})(\bar{X}_{1h} - \theta\,\bar{X}_{2h})\}^2$$

and ϕ_{Ah} and Φ_{Ah} are defined as before in (2.4).

3. DATA SET

An important objective of this investigation was to explore feasible ranges of the dual frame design parameters in order to determine the relative effects of each for dual frame optimization. The basic data set was provided by the Current Population Survey since this information was the most complete and readily available. However, it was not our intention to assess the relative efficiency of dual frame designs for CPS.

A simple cost model was used for design optimization. In 1982, the average CPS interviewer workload was about 50 dwellings and currently the average number of interviewer assignments per PSU (primary sampling unit) is about 2. These parameters, which determine q_A and $\bar{\bar{m}}_A$, were held fixed in our analyses. Thus, for Survey A, only n_{Ah}, the number of PSUs per stratum, is to be optimized.

In current RDD studies at the Census Bureau, the Waksberg (1978) sampling method is used with the within primary sampling quota equal to six residential telephone numbers. This parameter, which determines $\bar{\bar{m}}_B$, was held fixed as well. Then only n_{Bh}(or n_B) need be optimized for Survey B.

Our cost model is

$$C = C_o + \sum_h [113C_{Ah}\, n_{Ah} + 6C_{Bh}\, n_{Bh}]$$

for the separate estimator and

$$C = C_o + \sum_h 113C_{Ah}\, n_{Ah} + 6C_B\, n_B$$

for the combined estimator where C is the total survey budget, C_o is the fixed survey cost and C_{Ah} and C_{Bh} (C_B) are the average variable cost per unit for Survey A and Survey B households, respectively. For simplicity we assumed that the total fixed costs, C_o, would not change over current levels for a dual frame survey. Thus, we only need to know C_{Ah}, C_{Bh} (C_B), and $VC = C - C_o$ in order to optimize n_{Ah} and n_B.

Our study was concerned with the accuracy of dual frame estimators of population proportions. The population target parameter was the CPS monthly unemployment rate. For nontelephone and telephone households, the assumed rates were P_1 = 15% and P_2 = 6%, respectively. (This relative difference is consistent with available data.) These were the same in all strata, as were the within primary design effects $\delta_{Ah} = 1.33$ and $\delta_{Bh} = 1.25$. The latter was computed from a recent RDD experiment for the CPS. The usual srs formulas were used to compute σ_A^2, σ_1^2 and σ_2^2 and the domain design effects δ_{1h} and δ_{2h} were obtained using δ_A and the formulas in Biemer (1983, Appendix A).

We assumed that PSUs are stratified by state in both Survey A and Survey B for the separate estimator while, for the combined estimator, only Survey A was explicitly stratified. Table 1 summarizes the optimization parameters that varied across strata. The telephone coverage rates, T_{2h}, are proportions of households with telephones from the 1980 Census. The between PSU variance components, S_{Ah}^2, were obtained from CPS as the percent of S_{Ah}^2 to $S_{Ah}^2 + \delta_{Ah}\sigma_{Ah}^2/\bar{\bar{m}}_A$, denoted by $BPSU_h$. We also assumed that $S_{1h}^2 \approx S_{2h}^2 \approx S_{Ah}^2$.

Variable costs per household by state (C_{Ah}) were synthetically estimated for Survey A using available data on regional costs per household, interviewer time and mileage by state, and CPS state workloads. The stratum weights, W_h are based on recent CPS data on civilian labor force by state.

There is little information available on biases for area/list or telephone frame surveys. To simulate the effects of biases which may vary by state for each frame, it was assumed that biases are proportional to Survey A nonresponse rates (NR_h). Unless it is stated otherwise, nonsampling biases and interviewer correlations are zero in the analyses and ϕ_A and Φ_A in (2.5) and (2.7) are unity. The other parameter values in the optimization are summarized in Table 1, column 2 to 6.

Table 1.--CPS Data Set and Optimum Telephone Allocation for TBIAS = 0% and 5%.

Stratum	W_h	T_{2h}	$BPSU_h$	C_{Ah}	NR_h	$ALLOC_h$ TBIAS= 0%	$ALLOC_h$ TBIAS= 5%
US	100%	93%	10%	$14	4%	23%	3%
AL	1.5	87	9	17	3	43	10
AK	.2	84	8	20	6	54	8
AZ	1.4	90	2	17	5	45	0
AR	1.0	87	9	20	3	51	39
CA	1.0	95	2	9	5	0	0
CO	1.6	94	6	16	4	58	0
CT	1.4	97	0	15	4	60	0
DE	.3	95	0	20	3	69	41
DC	.3	96	0	19	7	71	7
FL	4.5	91	1	11	4	0	0
GA	2.4	88	6	15	4	34	0
HI	.4	95	.02	22	4	68	43
ID	.4	93	7	19	2	62	48
IL	4.9	95	2	11	7	0	0
IN	2.3	93	5	13	4	33	0
IA	1.3	97	11	14	2	62	38
KS	1.0	95	10	17	4	62	0
KY	1.6	87	7	17	3	41	0
LA	1.8	90	3	18	4	49	5
ME	.5	93	3	13	2	44	0
MD	2.0	96	4	15	4	60	0
MA	2.6	96	0	10	4	5	0
MI	3.9	96	2	11	4	13	0
MN	2.0	97	7	13	2	59	0
MS	1.0	84	8	19	3	42	26
MO	2.0	95	5	13	2	42	0
MT	.4	93	9	18	3	60	31
NE	.7	96	9	19	4	70	42
NV	.5	91	3	21	4	58	22
NH	.5	94	0	14	3	51	0
NJ	3.2	95	0	11	7	11	0
NM	.5	86	5	20	5	45	0
NY	7.0	92	1	8	6	0	0
NC	2.7	89	2	14	4	23	0
ND	.3	96	14	18	3	70	53
OH	4.5	94	.8	11	5	0	0
OK	1.4	92	5	17	5	54	0
OR	1.3	93	4	18	4	59	0
PA	4.8	96	2	10	4	0	0
RI	.4	95	0	17	3	61	29
SC	1.3	87	10	17	4	44	0
SD	.3	94	14	17	3	64	41
TN	2.0	89	7	17	3	47	8
TX	6.8	90	6	11	4	0	0
UT	.7	95	3	17	4	59	9
VT	.2	93	0	16	2	59	20
VA	2.4	93	13	15	4	50	0
WA	1.9	94	4	18	4	61	7
WV	.7	90	6	17	4	46	4
WI	2.1	97	11	14	2	63	0
WY	.3	92	7	24	4	66	44

4. ANALYSIS

The results of the dual frame survey optimization procedure for separate estimation are also given in Table 1 in the last two columns. The MSE of $\bar{x}_s'$ was minimized subject subject to total variable costs, VC, set at about $1 million which is 1/12 the estimated total annual variable cost for CPS. $ALLOC_h$, the optimal allocation of sample to the telephone survey, is reported in the next-to-last column for the case of zero nonsampling biases. For example, for estimating monthly unemployment, seven states would not use the telephone frame ($ALLOC_h$ = 0) while 27 states would allocate at least 50% to RDD. The total telephone sample allocation is 23%. (Please note that our optimization procedure is not intended to satisfy state minimum precision requirements.)

Now assume a small differential bias between the two modes, say 5% of the proportion to be estimated. In the last column, TBIAS, the telephone survey bias parameter, is 5% while ABIAS, the area survey bias parameter, 0. Now $ALLOC_h$ = 0 for 29 states and only 1 state would allocate as much as 50% to the telephone survey. Nationally, only 3% of the sample would be allocated to RDD.

This example illustrates how widely $ALLOC_h$ can vary between states as well as the potential impact of telephone bias. In the remainder of the paper, we will investigate through simulation (1) the role each major input parameter plays in determining $ALLOC_h$ for a stratum; (2) the gains in efficiency using the separate rather than the combined dual frame estimator; (3) the potential loss in precision using $\bar{x}_s'$ and $\bar{x}_c'$ when telephone domain sizes, T_{2h}, are not known; (4) the importance of nonsampling errors in dual frame survey design; finally, we illustrate (5) the loss in efficiency of dual frame estimators relative to the area/list frame estimator for subpopulation or domains of analysis.

Discussion of Table 2: Using the data of Table 1, in Table 2 we compare the relative efficiency of the four estimators of Section 2. Here, our measure of efficiency is the reduction in MSE of the estimator from the minimum MSE for the single frame design. Even though the reduction is small for all cases, it is at least twice as great for the separate estimators than for the combined estimators for both values of TBIAS. It also appears that, for these data, the effect of estimating T_{2h} in each stratum by t_{2h} (for $\bar{x}_s'$ and $\bar{x}_c'$) is a relatively small loss in efficiency over the case of T_{2h} known ($\bar{x}_s$ and $\bar{x}_c$).

Table 2.--Performance of $\bar{x}_s$, $\bar{x}_s'$, $\bar{x}_c$, and $\bar{x}_c'$ Relative to the Optimum Area/List Frame Estimator

	MSE Reduction TBIAS= 0%	MSE Reduction TBIAS= 5%	Telephone Allocation TBIAS= 0%	Telephone Allocation TBIAS= 5%
$\bar{x}_s$	7%	2%	24%	3%
$\bar{x}_s'$	6%	1%	23%	3%
$\bar{x}_c$	3%	1%	31%	1%
$\bar{x}_c'$	2%	0%	29%	1%

Discussion of Table 3: To produce this table, the variable cost of the dual frame survey, VC, was minimized for fixed variance. The variance was set at the minimum attainable by the area/list frame estimator for a budget of $1 million. The cost savings reported is the difference between $1 million and the minimum dual frame cost, divided by $1 million.

Two things are illustrated by this table: 1) the cost of estimating T_{2h} and 2) the rate at which decreasing telephone costs translate into total survey savings. Considering (1), we note that the absolute difference between $\bar{x}_c$ and $\bar{x}_{\hat{c}}$ is small, only two percentage points. Regarding (2), the data indicate that reducing telephone costs from $11 to $5 resulted in an added 28% savings (33%-5%). One could summarize by saying that, for the ranges considered, reducing telephone cost per household by $1 resulted in added savings of about 4% or about 50 - 60 cents per survey household.

Table 3.--CPS Dual Frame Variable Cost Savings by Telephone Cost/HH for $\bar{x}_c$ and $\bar{x}_{\hat{c}}$

Telephone Cost/HH	Cost Savings (% of variable costs)				
	$5	$7	$9	$11	$13
$\bar{x}_c$	33%	22%	12%	5%	<1%
$\bar{x}_{\hat{c}}$	31%	20%	10%	3%	0%

Discussion of Table 4: When the variances of the separate and combined estimators were compared in Table 2, we noted some improvement with the separate estimator. Table 4 suggests when gains in accuracy could be expected. The impact of six input parameters on the relative improvement of $\bar{x}_{\hat{s}}$ over $\bar{x}_{\hat{c}}$ is illustrated in this table. Assuming two strata, we computed the MSE and $ALLOC_h$ (h=1,2) for the optimal separate and combined estimators when one stratum was set to the low value of a parameter and the second stratum to the high value. These values were obtained from the CPS data set (Table 1). All other parameters except the one being investigated, were identical and set to their midrange values or, in the case of nonsampling bias, to zero. The last column of Table 4 is $[MSE(\bar{x}_{\hat{c}}) - MSE(\bar{x}_{\hat{s}})]/MSE(\bar{x}_{\hat{s}})$ with optimal allocations to the telephone survey as indicated.

For the ranges considered, the largest improvement is observed when area survey costs or area survey bias are varied while the smallest improvement is observed for the differing between PSU variances. In this example, the strata weights, W_h, were equal (.5). The difference between separate and combined MSE's were smaller when one stratum was substantially smaller than the other.

Table 4.--Illustration of the Improvement of the Separate Estimator over the Combined Estimator for Two Strata

Parameter	$ALLOC_h$ for $\bar{x}_{\hat{s}}$ St. 1 (low)	$ALLOC_h$ for $\bar{x}_{\hat{s}}$ St. 2 (high)	$ALLOC_h$ for $\bar{x}_{\hat{c}}$ Both Strata	Improvement
Telephone Coverage	32%	77%	52%	5%
Btwn PSU Variance	42%	53%	46%	2%
Cost/hh (area)	0%	63%	39%	10%
Cost/hh (tele.)	70%	8%	46%	9%
Bias (area)	48%	70%	70%	10%
Bias (tele.)	48%	2%	3%	6%

Discussion of Table 5: The values of $ALLOC_h$ reported in Table 1 for $\bar{x}_{\hat{s}}$ range from 0% to 68%. The question addressed in this analysis is how the particular configuration of input parameters for a stratum affect the optimal telephone sample allocation. We assumed one stratum with all parameters, except the particular parameter being investigated, equal to the midpoint of their ranges in the CPS data set or, in the case of bias, to zero. Reported in columns 2 and 3 are the optimal allocations computed at each extreme value. In column 4 are the resulting changes in $ALLOC_h$ divided by the parameter range. For example, for every percentage point increase in telephone bias, the telephone allocation decreased by about 5 percentage points. Telephone bias appears to be the most important cause of fluctuations in $ALLOC_h$ while between PSU variance is the least important.

Table 5.--Sensitivity Analyses for the CPS Data Set

Parameter [range of variation]	Telephone Allocation At low	Telephone Allocation At high	Change per unit increase
Telephone bias [0%, 10%]	47%	1%	-5
Cost/hh [$8, $28]	1%	62%	3
Telephone Coverage [80%, 98%]	30%	76%	3
Betw. PSU Variance [0%, 20%]	40%	52%	1

Discussion of Table 6: The optimizations in the preceding analysis were carried out assuming the target population parameter to be estimated is a proportion of the total population. Assuming the optimal dual frame allocation can be achieved for this objective, efficiency may still be lost for estimators of proportions for demographic subpopulations. We focus here on only two factors which may account for this loss in efficiency: 1) the proportion of the subpopulation reachable by telephone may be substantially less than the population as a whole and 2) the estimator of T_{2h} for the subpopulation used in the estimator $\bar{x}_S^{\xi}$ or $\bar{x}_C^{\xi}$ may be less precise for domains which cluster geographically.

In our analyses we took into account the changes in sample size and within primary design effects in the subpopulation estimator variance formulas. Except as indicated, all parameters which were varied in the previous analyses were set to midrange values; all biases and interviewer correlations were set to zero. The design effects ϕ_A and Φ_A associated with the estimator t_2 were varied together from $\phi = 0$ (corresponding to the case were T_2 is known for the subpopulation) to $\phi = 10$ (corresponding to a highly clustered subpopulation such as Hispanics or Blacks). Table 6 gives the efficiency of the dual frame relative to the area/list frame as a function of T_2 and ϕ. The size of the subpopulation for the table was 20%; however, this was varied from 5% to 80% with trivial effects on the results.

The table indicates that the dual frame estimator lost efficiency over the single frame estimator in the range 70% - 85% telephone availability (89% was assumed for the total population dual frame optimization). As expected, the loss is more severe the more geographically clustered is the subgroup. Further, the estimator $\bar{x}_C^{\xi}$ seems to be rather robust to small fluctuations in ϕ.

Table 6.--Efficiency of Dual Frame Estimators of Proportions for Subpopulations

Domain Telephone Coverage	Var (dual frame) / Var (single frame)			
	$\phi = 0$[1]	$\phi = 1.0$	$\phi = 5.0$	$\phi = 10.0$
60%	1.22	1.25	1.38	1.54
70%	1.12	1.15	1.27	1.42
80%	1.0	1.03	1.13	1.26
90%	.87	.88	.95	1.03
100%	.71	.71	.71	.71

[1]The design effects, ϕ_A and Φ_A, associated with t_2, the estimator of telephone population coverage from Survey A, were both set to ϕ. $\phi = 0$ corresponds to estimation with T_2 known.

5. CONCLUSIONS

The results of this study are based upon a somewhat restricted range of the dual frame design parameters, viz., CPS cost and error data for national estimates of monthly unemployment rate. Therefore, caution must be exercised in extending these results for general dual frame survey design. However, there are a number of observations that can be made.

1. Survey bias appears to be the most important factor in determining the allocation of sample in dual frame survey design. In a number of other runs of this procedure for the CPS data set, error and cost parameters where set at their extreme values in various combinations. In each instance for TBIAS = 5%, telephone allocation rarely rose above 4% of the sample for any of the four estimators. This would suggest that perhaps a more general statement on the effect of telephone sampling bias is possible. Given the survey objective of minimizing MSE of an estimator for fixed cost, a telephone survey bias as small as 5% could practically eliminate the telephone survey in the dual frame design for making national estimates in large scale surveys.

2. The separate estimator can be a significant improvement over the usual combined estimator in stratified dual frame surveys. However, in this particular study, the degree of improvement was not dramatic - between 5% to 10% increase in efficiency.

3. The loss in precision as a result of using $\bar{x}_S^{\xi}$ or $\bar{x}_C^{\xi}$ (for unknown T_{2h}) instead of $\bar{x}_S$ or $\bar{x}_C$ (when T_{2h} is known) was small for estimators of total population characteristics. However, as Table 6 illustrates, the loss is greater for the estimators over subpopulations with low telephone availability. This would indicate that moderate gains in efficiency could be realized by reducing the variance of t_{2h} in the estimators.

4. The loss in relative efficiency of dual frame estimates over subpopulations increases as telephone availability for the subpopulation decreases and as geographic clustering of the subpopulation increases. For our data set, subpopulations with telephone coverage rates of 85% or higher were estimated with about the precision of the area/list survey. Populations having less than 85% telephone coverage, could suffer substantial losses in estimation precision compared to the current levels.

ACKNOWLEDGEMENTS

The author wishes to thank Ms. Vicki Horton for her patience and diligence in typing the final as well as earlier drafts of the paper.

REFERENCES

Biemer, P.P. (1983). "Methodology for Optimal Dual Frame Sample Design." (Submitted)

Casady, R.J., Snowden, C.B., Sirken, M.G. (1981). "A Study of Dual Frame Estimators for the National Health Interview Survey", Proceedings of the ASA, Section on Survey Research Methods.

Hartley, H.O. (1962). "Multiple Frame Surveys," Proceedings of the ASA, Social Statistics Section.

Waksberg, J. (1978). "Sampling Methods for Random Digit Dialing", JASA, 70, 40 - 46.