THE IMPACT OF BIAS ON DUAL FRAME SURVEY DESIGN

James M. Lepkowski and Robert M. Groves, The University of Michigan

As the cost of personal visit or face-to-face household interviewing increases, surveys are being conducted more frequently by less expensive telephone survey methods. Telephone surveys use both a telephone frame for sample selection and a telephone mode of data collection, while the more traditional personal visit survey typically uses an area probability sample design and a face-to-face mode of data collection. Not only are telephone surveys found to be generally less costly than personal visit surveys, but they also offer the advantage that they may be conducted in a centralized location where closer supervision and monitoring of interviewer performance is possible than for a personal visit survey. On the other hand, telephone surveys have larger nonresponse rates (and presumably larger nonresponse bias) than do personal visit surveys. In addition, households without telephones are not covered in a telephone survey, a source of coverage error that can be sizeable for survey measures of interest.

An alternative to these single frame, single mode survey designs is to mix the two frames and modes of data collection in a single survey design. For example, a telephone survey can be conducted and supplemented with face-to-face interviews in nontelephone households to reduce noncoverage bias. Unless the costs of screening to identify nontelephone households is small, it is more efficient to conduct face-to-face interviews in both telephone and nontelephone households selected from an area frame. Whether screening or not, the central problem in designing a dual frame survey is to determine the best allocation of sample size between the two frames.

Hartley (1962) examined the more general problem of multiple frame survey designs, and for a specific type of post-stratified estimator, developed methods for determining the optimal allocation of the sample among multiple frames. Although Hartley applied the results to several different sample designs, he did not specifically address the problem of stratified multistage sample designs in one or both frames. Casady, Snowden, and Sirken (1982) applied Hartley's approach to a dual frame design in which telephone and area frames and stratified multistage sample designs in both frames are used.

Groves and Lepkowski (1982) investigated a similar dual frame sample design but used a much more detailed cost model for the dual frame design than had been considered by Casady, Snowden, and Sirken. Biemer (1983) examined the same dual frame design but extended the error model to include a consideration of both response errors and bias in the determination of the optimal allocation to the two frames. Biemer included bias in the telephone frame only, and allocated optimally within strata in the sample design.

In the investigation reported in this paper, a model is developed for survey errors which includes the added sources of error considered by Biemer as well as bias in the area frame. A cost model similar to that in Groves and Lepkowski (1982) is used to determine the optimal allocation given the more complete error structure, allocating across strata in the sample selection. The methods are then applied to the National Crime Survey to find the optimal allocation to the two frames for two different measures from the survey, and the effect of bias on the allocation is considered.

A Mean Square Error Model

Consider a sample design in which stratified multistage sample selection methods have been used in both frames. For example, in the area frame a multistage selection of segments of neighboring households is selected through successive stages such as counties and county-like units, enumeration districts, and segments. In the telephone frame, a random digit dialing procedure is used in which numbers are clustered by banks of 100 consecutive numbers. Let m_a and m_b denote the number of primary selections in the area and telephone frames, respectively, and let $\bar{n}_a$ and $\bar{n}_b$ denote the expected number of households in a cluster in each frame, respectively.

There are three domains of interest in the specification of an estimator from this design: nontelephone households from the area frame, telephone households from the area frame, and telephone households from the telephone frame. Let n_d and $\bar{v}_d$ denote the sample size and sample mean, respectively, for each domain where d = 11 for nontelephone households, d = 12 for telephone households from the area frame, and d = 2 for telephone households from the telephone frame. An estimator of the overall mean across the domains is

$$\bar{v} = \hat{p}\, \bar{v}_{11} + (1 - \hat{p})\, [\theta\, \bar{v}_{12} + (1 - \theta)\, \bar{v}_2]$$

where $\hat{p}$ denotes the proportion of households or persons which are in the nontelephone domain and θ denotes a mixing parameter which is chosen to minimize the mean square error of $\bar{v}$. This estimator is a weighted average of means from the telephone and nontelephone households, with the mean from the telephone households being a mixture of estimates from the two frames.

An allocation of sample sizes n_a and n_b is needed such that the overall error for $\bar{v}$ is minimized, where the overall error for $\bar{v}$ is referred to as the mean square error. For the k-th person in the a-th cluster of stratum h in the d-th domain, let y_{dihak} denote the sample value recorded by the i-th interviewer in that domain. Consider the response error model

$$y_{dihak} = \mu_d + \alpha_{d(i)} + \epsilon_{d(i)hak}$$

where μ_d denotes an overall mean for the d-th domain, $\alpha_{d(i)}$ denotes the deviation of responses for the i-th interviewer from the mean response for all interviewers in the d-th domain, and $\epsilon_{d(i)hak}$ is the sampling deviation for the (hak)-th person from the mean for the d-th domain. We assume that the response and sampling deviations are uncorrelated, the response deviations between the area and telephone frame households are uncorrelated, and that sampling deviations between area and telephone frame households are uncorrelated.

Using this conceptual response model, the mean square error for the estimator $\bar{v}$ was obtained as

$$MSE(\bar{v}) = E[\bar{v} - \bar{V}]^2 =$$

$$\hat{p}^2 ((\sigma_1^2\delta_{11} + \sigma_{r11}^2\delta_{r11}/n_{11}) + B_{11}^2)$$

$$+ (1 - \hat{p})^2\theta^2((\sigma_2^2\delta_{12} + \sigma_{r12}\delta_{r12})/n_{12} + B_{12}^2)$$

$$+ (1 - \hat{p})^2(1 - \theta)^2((\sigma_2^2\delta_2 + \sigma_{r2}^2\delta_{r2})/n_2 + B_2^2)$$

$$+ 2\hat{p}(1 - \hat{p})\theta[\rho_s\sigma_1\sigma_2\sqrt{\delta_{11}\delta_{12}/n_{11}n_{12}}$$

$$+ \rho_r\sigma_{r11}\sigma_{r12}\sqrt{\delta_{r11}\delta_{r12}/n_{11}n_{12}} + B_{11}B_{12}]$$

$$+ 2\hat{p}(1 - \hat{p})(1 - \theta)B_{11}B_{12} + 2(1 - \hat{p})^2\theta(1 - \theta)B_{12}B_2$$

$$+ Var(\hat{p})\{E(\bar{v}_{11}) - [\theta E(\bar{v}_{12}) + (1 - \theta)E(\bar{v}_2)]\}$$

where

- $\bar{V}$ = the population mean value,
- σ_1^2 = the element variance for nontelephone households,
- σ_2^2 = the element variance for telephone households,
- δ_d = the sampling design effect for households in the d-th domain,
- ρ_s = the sampling correlation between telephone and nontelephone households from the area frame,
- σ_{rd}^2 = the simple response variance for households in the d-th domain,
- δ_{rd} = the correlated response design effect for households in the d-th domain,
- ρ_r = the response correlation between telephone and nontelephone households in the d-th domain,
- B_d = the bias (i.e., the deviation of the sampling distribution mean from the population mean) for the d-th domain, and
- d = 11, 12, or 2.

The effects of the stratified multistage sample design, including the use of a ratio mean, are summarized in terms of the sampling design effects δ_d. The effects of the interviewer or the correlated response variance are summarized in terms of the response deviation design effects δ_{rd}. The sampling and response correlations arise because telephone and nontelephone households in the area frame were selected from the same primary selections and interviewed by the same interviewers.

The expression for the mean square error is composed of separate pieces that reflect the contribution of various domains and errors, and contains the unknown parameter θ which can be chosen to minimize the mean square error at any given allocation. The expression for the value of θ which minimizes the mean square error is somewhat lengthy and is not repeated here (derivation of the mean square error and the optimal value of θ can be obtained from the authors). The problem is thus to find the allocation of sample size between the area and telephone frames which minimizes the mean square error under the constraints of a cost model.

A Dual Frame Cost Model

A simple cost model which includes only per unit costs for interviews in the two frames could have been developed, but a simple cost model would ignore aspects of the dual frame survey administrative structure which could effect the allocation. A more detailed cost model was used to make key features of the administrative structure explicit in the model and hence subject to manipulation in a simulation exercise, of which this investigation is one part.

The cost model used in this investigation is

$$C = C_0 + c_r \cdot n_r + c_t \cdot n_t + c_{ta} \cdot n_{ta}$$

$$+ c_{pt} INT[n_a/w_a + 1] + c_{tt} INT[w_t/c_b] + c_a \cdot n_2/40$$

$$+ c_a \cdot n_a + c_b \cdot n_2$$

where

- C = the total survey budget,
- C_0 = indirect costs for survey administration,
- c_{ro} = the cost per regional office in the area frame,
- n_{ro} = the number of regional offices,
- c_t = the cost per telephone facility of 50 interviewers,
- n_t = the number of telephone facilites,
- c_{ta} = the cost of expanding a telephone facility to add 10 interviewer carrels,
- n_{ta} = the number of expansions needed beyond beyond the basic facilities,
- c_{pt} = the cost to train a personal visit interviewer,
- n_a = $n_{11} + n_{12}$, the annual number of interviews in the area frame,
- w_a = the annual number of completed interviews for a personal visit interviewer,
- c_{tt} = the cost of training a telephone interviewer,
- INT[·] = the integer portion of the argument [·],

w_t = the annual number of completed interviews for a telephone interviewer,

c_a = the remaining cost per completed personal interview,

c_b = the remaining cost per completed telephone interview, and

n_2 = the annual number of interviews in the telephone frame.

The full complexity of the cost model is not shown here since the number of regional offices, the number of telephone facilities, and the number of expansions to telephone facilities are functions of the sample size. The remaining cost per completed telephone interview (c_b) is a function of the response rate, increasing as the response rate desired is increased. In addition, the use of step functions such as INT[·] makes the finding of an explicit solution for the allocation quite difficult.

Although the complexity of the cost model creates problems for finding exact solutions to the allocation problem, it offers the advantage of explicit types of cost which can be more directly measured for a particular survey. Since the cost estimates that can be obtained often are subject tc uncertainty, the detailed cost model also allows sensitivity analysis for specific cost components to identify those which need to be more carefully estimated because they influence the allocation dramatically when changed only slightly.

Determination of an Optimal Allocation

Given the mean square error model and the detailed cost model, the problem is to determine which allocation of sample sizes to the two frames (i.e., values of n_a and n_b) provides the smallest mean square error. An explicit closed form solution was not sought because of the complexity of the cost model. Instead, an algorithm was developed which obtained the mean square error for a range of discrete allocations, the allocation with the smallest mean square error being identified as the approximate optimum. The algorithm begins with a small sample size for the telephone frame, solves for the sample size in the area frame which meets the total budget available for the survey, and then computes the mean square error.

A computer program was written to solve the allocation problem using this algorithm and input of values for the error and cost parameters needed in the mean square error and cost models. In the next section, the results of applying the cost and error models and the minimization algorithm to a large federal survey are presented.

An Application to the National Crime Survey

The National Crime Survey (NCS) is a large national survey of the civilian noninstitutionalized population of the U.S. designed to provide estimates of the extent and nature of victimization occurring to the population. The survey collects information monthly from a sample of persons about victimizations occurring during the previous six months. The survey employs a rotating panel design in which panels of housing units are visited every six months over a three year period for a total of seven visits. The NCS is a continuing survey operation, and the rotating panel design replaces panels that have been visited seven times with new previously unvisited panels.

Each month approximately 14,000 housing units are visited and a total of 12,000 completed household interviews and 25,000 completed person interviews are obtained. The NCS employs more than 500 interviewers located in more than 350 primary areas across the U.S. The annual budget for the survey is more than seven million dollars.

A dual frame design has previously been suggested and investigated for the NCS (Groves and Lepkowski, 1982). Extensive work was done to estimate the parameters in the mean square error and cost models described in previous sections. (The values of parameters in the mean square error model for two types of crime are not presented in this paper but may be obtained from the authors.) Estimation of these parameter values involved detailed study of NCS estimates, sampling errors, response errors, and costs. Since numerous assumptions were made to derive values for some of the parameters, a sensitivity analysis has been conducted to identify parameters which effect the allocation substantially when changed across reasonable sets of alternative values. Only findings from an investigation of the importance of bias to the allocation are presented here. The results are presented for two types of crime routinely reported by the NCS.

Figure 1

Allocation to Telephone Frame as Telephone Frame Bias Increases from 0 to 10 Percent and Area Frame Bias Is 0 Percent of the Estimated Rate, Total Personal Crimes

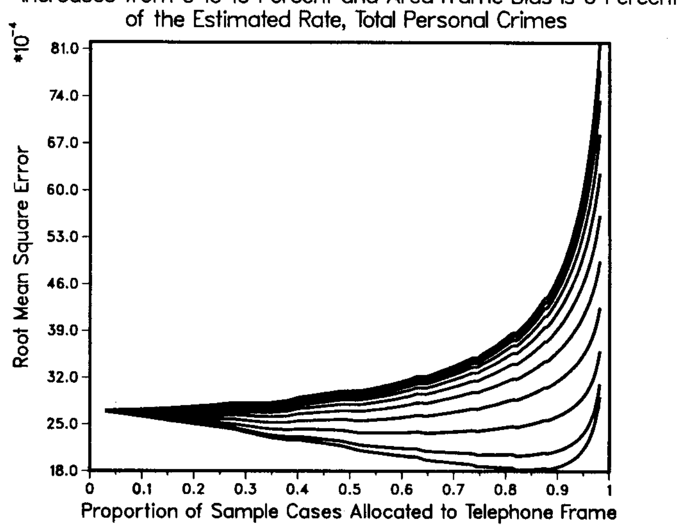


Figure 1 presents for the Total Personal Crimes victimization rate the square root of the mean square error which would be achieved for a fixed survey budget of $5.5 million under different allocations to the telephone frame and under different assumptions about the amount of bias in each domain. The lowest of the eleven curves on the figure represents the allocation problem when there is no bias in any of the three domains. As the allocation of sample size to the telephone frame increases to about 82 percent of the total sample size, the root mean square error decreases.

The optimal allocation of sample size to the two frames thus occurs when aproximately 82 percent of the sample is allocated to the telephone frame. At the optimal allocation the root mean square error is about one-third less than that achieved when none of the sample is allocated to the telephone frame. The irregularities in the curve are the result of the use of step functions in the cost model.

The remaining curves in Figure 1 represent the allocation problem when bias occurs in the results from the telephone frame but no bias is present for estimates from nontelephone households or from telephone households selected from the area frame. The next to the last curve is the root mean square error when there is bias equivalent to one percent of the estimated victimization rate for the telephone households, a relative bias of one percent for the telephone frame. The next highest curve represents the root mean square error under different telephone frame allocations when there is a two percent relative bias for telephone households selected from the telephone frame. Each subsequent curve represents another one percent increase in the relative bias.

As the amount of relative bias increases in the telephone frame, the optimal allocation to the telephone frame decreases from 82 percent when there is no bias to zero percent when there is 10 percent bias. The root mean square error curves become somewhat more horizontal for lower allocations to the telephone frame as the amount of bias increases.

In Table 1 the specific optimal allocations derived from the algorithm described previously are given for the bias assumptions in Figure 1 and, since zero bias in the area frame is an unrealistic assumption, several other sets of assumptions about bias. The first column provides the optimal allocations when there is no bias in the area frame domains and increasing relative bias in the telephone frame. Once the amount of relative bias reaches nine percent in the telephone frame, the optimal solution is to allocate none of the sample to the telephone frame. This represents, however, a sizeable difference in relative bias among the three domains.

TABLE 1

OPTIMAL SAMPLE ALLOCATIONS TO TELEPHONE FRAME UNDER ALTERNATIVE RELATION BIASES IN AREA AND TELEPHONE DOMAINS, TOTAL PERSONAL CRIMES

% Bias in Telephone Frame	% Bias in Area Frame (%BIASA1 = %BIASA2)										
	0	1	2	3	4	5	6	7	8	9	10
0	82.6										
1	82.1	82.6									
2	64.6	74.5	82.6								
3	34.6	34.6	34.6	82.6							
4	34.6	0.0	0.0	0.0	82.6						
5	14.0	0.0	0.0	0.0	0.0	82.6					
6	11.4	0.0	0.0	0.0	0.0	0.0	82.6				
7	8.7	0.0	0.0	0.0	0.0	0.0	0.0	82.6			
8	5.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	82.6		
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	82.6	
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	82.6

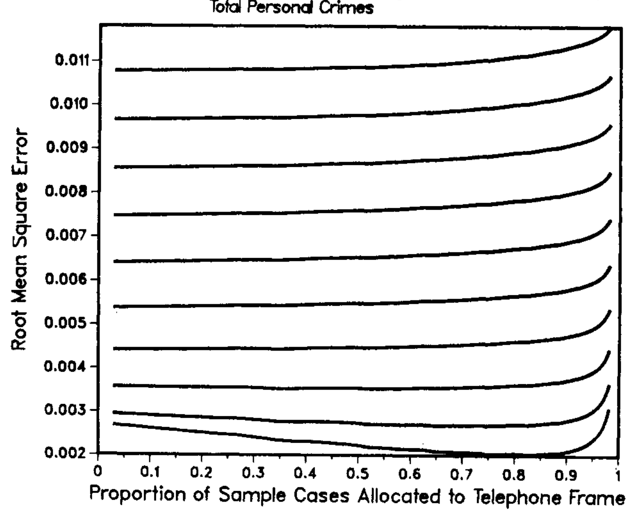
The subsequent columns in Table 1 demonstrate what happens as increasing amounts of relative bias are added to the area frame. It is assumed that the relative bias is the same for telephone and nontelephone households selected from the areas frame, but since the victimization rates differ in the two domains (in particular, nontelephone households have approximately 50 percent higher victimization rates than telephone households), the amount of bias will differ between the two domains. In addition, it is assumed that the relative bias in the telephone frame is always no smaller than that in the area frame domains.

As long as the amount of relative bias is equal in the two frames, the optimal allocation remains at 82 percent. However, as the relative bias increases for the area frame domains, smaller increases in the telephone frame relative bias above the area frame relative bias leads to zero percent optimal allocations. Thus, when there is no bias in the area frame, approximately nine percent bias is needed in the telephone frame to achieve a zero percent optimal allocation; when there is three percent bias in the area frame only a one percent increase in bias to four percent for the telephone frame leads to a zero percent optimal allocation.

Examination of the optimal allocation alone fails to reveal an important effect of bias on the optimal allocation problem, the flattening of the root mean square error curve. Figure 2 presents the root mean square error curves when the relative biases in the two frames differ by exactly one percent (i.e., the set of optimal allocations corresponding to the first off-diagonal optimal allocations in Table 1). The lowest curve presents the root mean square error when there is no bias in the area frame and one percent bias in the telephone frame; the next curve presents root mean square error when there is one percent bias in the area frame domains and two percent bias in the telephone frame; and, so on, until there is nine percent bias in the area frame and ten percent bias in the telephone frame.

Figure 2

Allocation to Telephone Frame for Area Frame Bias of 0 TO 9 Percent of the Estimated Rate and Telephone Frame Bias 1 Percent Larger than Area Frame Bias, Total Personal Crimes

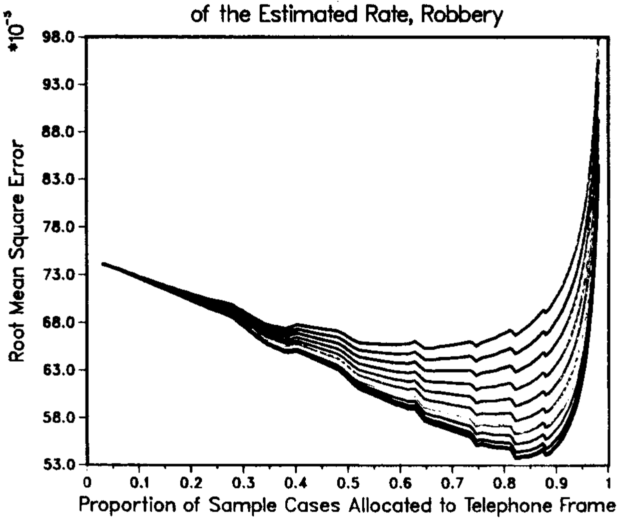


The optimal allocations for these curves rapidly declines to zero percent starting when there is three percent bias in the area frame and four percent in the telephone frame. But the flat shape of the curves suggests that an allocation which departs from the optimum even substantially will not result in large losses in root mean square error. For example, for the highest curve (i.e., nine percent area frame and ten percent telephone frame bias) the smallest root mean square error is achieved when no sample size is allocated to the telephone frame. However, if 80 percent of the sample were allocated to the telephone frame, the root mean square error would increase by only about five percent.

The multipurpose nature of the NCS, reporting on several types of crime and presenting results for numerous subclasses of the population, forces a consideration of other types of crime in the allocation investigation as well. Figure 3 presents for the Robbery victimization rate the root mean square error for allocations to the telephone frame when there is no bias in the area frame domains and increasing amounts of bias in the telephone frame. As for Figure 1, the lowest curve presents root mean square error when there is no bias in any of the three domains, the next highest curve presents root mean square error when there is one percent bias in the telephone frame and none in the area frame domains, and so on.

Figure 3

Allocation to Telephone Frame as Telephone Frame Bias Increases from 0 to 10 Percent and Area Frame Bias Is 0 Percent of the Estimated Rate, Robbery



Since the Robbery victimization rate has much different error properties than that for Total Personal Crimes, the root mean square error curves and the optimal allocations are different than those observed for Total Personal Crimes. When there is no bias in any of the three domains (i.e., the lowest curve), the optimal allocation is achieved when nearly 83 percent of the sample is allocated to the telephone frame. Even when there is ten percent bias in the telephone frame and no bias in the area frame domains (i.e., the highest curve in Figure 3), the optimal allocation is still substantially to the telephone frame at approximately 65 percent.

Table 2 presents the optimal allocations for Robbery under the same set of bias assumptions as illustrated for Total Personal Crimes in Table 1. The optimal allocations for Robbery are all markedly higher than those obtained for Total Personal Crimes. In no case is the optimal allocation equal to zero, and the lowest optimal allocation is 65 percent. The optimal allocations for Robbery are clearly less effected by bias than were those for Total Personal Crimes.

TABLE 2

OPTIMAL SAMPLE ALLOCATIONS TO TELEPHONE FRAME UNDER ALTERNATIVE RELATION BIASES IN AREA AND TELEPHONE DOMAINS, ROBBERY

% Bias in Telephone Frame	% Bias in Area Frame (%BIASA1 = %BIASA2)										
	0	1	2	3	4	5	6	7	8	9	10
0	82.6										
1	82.6	82.6									
2	82.1	82.1	82.6								
3	82.1	82.1	82.1	82.6							
4	82.1	82.1	82.1	82.1	82.6						
5	82.1	82.1	82.1	82.1	82.1	82.6					
6	82.1	82.1	82.1	82.1	82.1	82.1	82.6				
7	74.5	82.1	82.1	82.1	82.1	82.1	82.1	82.6			
8	74.5	74.5	74.5	74.5	82.1	82.1	82.1	82.1	82.6		
9	74.5	74.5	74.5	74.5	74.5	74.5	82.1	82.1	82.1	82.6	
10	64.6	64.6	64.6	64.6	74.5	74.5	74.5	82.1	82.1	82.1	82.6

Figure 4

Allocation to Telephone Frame for Area Frame Bias of 0 To 5 Percent of the Estimated Rate and Telephone Frame Bias 5 Percent Larger than Area Frame Bias, Robbery

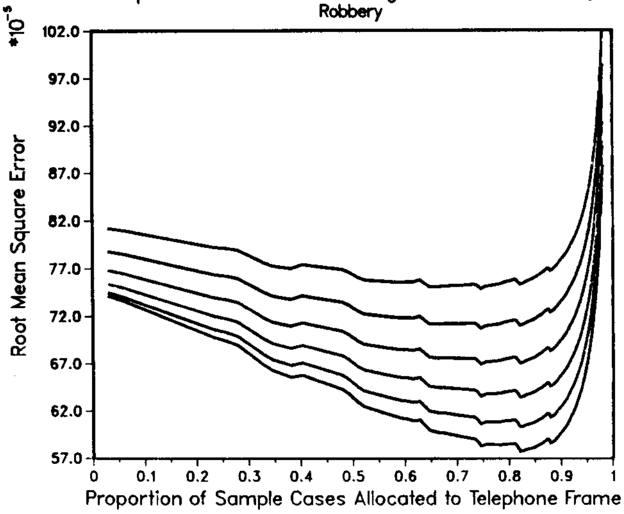


Figure 4 presents the root mean square error curves when the bias in the telephone frame is five percent larger than that in the area frame domains. The bottom curve presents the root mean square error when there is no bias in the area frame domains but five percent bias in the telephone frame, while the highest curve presents root mean square error when there is five percent bias in the area frame domains and ten percent in the telephone frame. The curves are, as before, flattened by increasing amounts of bias, but the optimal allocations do not decline substantially as the amount of bias in the area frame increases. Departures from the optimum will not lead to substantial increases in the root mean square error for curves based on larger amounts of bias in the area frame. Thus, a zero percent allocation for the highest curve where the optimal allocation is approximately 75 percent will have about an 8.5 percent increase in root mean square error over that obtained at the optimum.

Discussion

The dual frame error and cost models and the determination of an optimal allocation to the frames described here utilize parameters for which, in many cases, there is some uncertainty about the value. Bias in each of the three domains is a largely unknown parameter, and it has been demonstrated that the allocation is quite sensitive to changes in bias for at least one type of crime measured in the NCS. Before an allocation for a dual frame design for the NCS can be determined, relatively accurate estimates of bias must be obtained for each domain.

Although increasing bias to the telephone frame while holding the amount of bias constant in the area frame leads to lower optimal allocations to the telephone, the investigation has also shown that bias tends to "flatten" the root mean square error curves. That is, allocations which depart from the optimum have smaller losses in root mean square error when there is more bias in the model. This feature of the dual frame allocation problem would allow a design to be chosen which was not far from the optimum in terms of root mean square error but which could provide substantially improved levels of precision because a large proportion of the interviews are obtained from less expensive telephone interviews (i.e., sample sizes would be larger for the same fixed budget). Thus, for Total Personal Crimes, where an optimal allocation of zero percent to the telephone frame seems to be indicated by these results, an allocation of 80 percent to the telephone would result in root mean square errors somewhat larger than at the optimum, but the sampling variance would be considerably reduced (by approximately one-third) for the same fixed budget. On the other hand, for Robbery, where optimal allocations were at least 65 percent in the problems investigated here, a zero percent allocation to the telephone frame would lead to some losses in root mean square error and to substantial losses in precision. But departures from the optimum for either Robbery or Total Personal Crimes could be made to find a compromise solution.

Finally, the difficulty of multipurpose allocation for the NCS has been illustrated by presenting findings from two different types of crime. For one, Total Personal Crimes, bias has a large effect on the optimal allocation, while for the other, the effect of bias is relatively small. In order to determine a compromise allocation for the overall survey design, the different estimates may be weighted according to a variety of weighting schemes (e.g., inversely proportional to variance, arbitrary weights assigned based on relative importance). The choice of a compromise solution can be made easier, though, by recognizing that the root mean square error curves are flatter in shape with increasing amounts of bias. Departures from the optimal allocation will not lead to large losses or gains in root mean square error.

References

Groves, Robert M. and Lepkowski, James M. "Dual Frame, Mixed Mode Survey Designs," submitted to Journal of the American Statistical Association, February, 1984.

Casady, Robert, Snowden, Cecelia, and Sirken, Monroe. "A Study of Dual Frame Estimators for the National Health Interview Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1981.

Hartley, H. "Multpile Frame Surveys," Proceedings of the Social Statistics Section, American Statistical Association, 1962.

Biemer, Paul. "Optimal Dual Frame Design: Results of a Simulation Study," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1983.