Reordering the Darkness: Application of Effort and Unit Nonresponse in the Survey of Consumer Finances[1]

Arthur B. Kennickell
Federal Reserve Board, Mail Stop 153, Washington, DC 20551

Key words: Nonresponse, interviewers

Nonresponse may be addressed before, after, or during data collection. Before data collection, interviewers may be trained in techniques to contact respondents and gain their cooperation; it may also be possible to deal with some statistical inefficiencies induced by nonresponse by stratifying on characteristics associated with nonresponse. After data collection is completed, many surveys use post-stratification to adjust for important known deviations of the interviewed population from the target population.

During the field period, additional application of effort might increase response rates. But straightforward dedication simply to reducing nonresponse may lead to the application of substantial effort without accompanying confidence of reducing nonresponse biases. If contact and response propensities were sufficiently understood and adequate data were available rapidly enough during the field period, one might be able to make more efficient use of resources by targeting specific cases or types of cases.

Unfortunately, it is not straightforward to understand either contact or response propensities—not least because virtually everything we know from the field is affected by the application of effort. "Effort" is a complex product of the decisions of interviewers, their managers, survey organizations, and survey sponsors, each of whom may face a different set of objectives, constraints, and incentives. Where effort is only partially observed, as in the case of most of interviewers' work, complicated patterns of "accidental" deviations from optimal behavior and even shirking may arise (see Kennickell, 2000a).

This paper focuses on nonresponse and the application of effort in the 2001 Survey of Consumer Finances (SCF), using case-level administrative records along with information about the neighborhoods of the sample addresses. It follows on earlier related work in Kennickell (1999a and 1999b); recent work reported in Groves et al. (2003) deals with related problems of monitoring and management of field effort. The first section of this paper gives background on the SCF. The next section presents a simple behavioral model of survey organizations. The third section describes the level of effort applied in the 2001 SCF and a model of the likelihood of continued application of effort to cases. The fourth section argues for the development of a contact strategy to facilitate better estimation of and control for the respondents' role in nonresponse while allowing more control over costly field operations. A final section makes recommendations for the future.

I. Background on the SCF

The SCF is designed as a survey of households' finances and it is conducted on a triennial basis by the Federal Reserve Board in cooperation with the Statistics of Income Division (SOI) of the IRS. The SCF data used in this paper derive from the 2001 survey.[2] The data for this survey were collected by NORC at the University of Chicago, between May and December of 2001. Cases were largely completed in-person, but 34.6 percent of cases preferred to be interviewed by telephone. The dual-frame sample for the survey consists of a multistage area-probability sample (AP) and a list sample (LS) intended to oversample relatively wealthy households. The 2001 sample contained about 10,000 observations approximately equally split between AP and LS cases.

The response rate (adjusted for ineligible units) for the AP sample was 68.1 percent. Unlike AP cases, LS cases were given the option of returning a postcard to refuse to participate in the survey in advance of their being approached by an interviewer; 13.2 percent of the sample did so. In addition, a relatively small number of cases were deleted from the sample during a review designed to eliminate members of "*Forbes* 400" and a few other very unusual people.[3] Leaving aside the deleted cases, the LS response rate was 35.4 percent. By far, the largest category of nonresponse was "final stopped work"—21.7 percent of the eligible AP cases and 51.4 percent of the LS eligible cases; these are cases that, in theory, remained eligible for further work at the end of the field period.

II. A model of case management

Through the process of case management, cases are resolved into completed or refused status as interviewers try to present information to respondents through a number of actions ("attempts") and respondents arrive at a determination of their willingness to participate. As work continues, the set of cases with censored outcomes—those remaining at risk to be completed or refused in a future attempt—shrinks. Several factors may complicate this process. First, it is not always possible to reach respondents to provide them with information. Second, there is no firm definition of what constitutes a final refusal. For example, some "refusal converters" are renowned for their ability to persuade people who have repeatedly refused other interviewers. Third, the application of effort is generally a dynamic decision process that is a function of many factors. Some cases may not be worked to the point that an unambiguous resolution is reached, probably because those cases are ones perceived to be either too "expensive."

As a stylized framework, consider the following simple model. Let the population distribution of the characteristics of interest be given by $\mathcal{R}^{\bullet}$; for a sample of size N from the target population let the distribution be given by $\mathcal{R}$. To avoid needless complications, assume that $\mathcal{R}$ and $\mathcal{R}^{\bullet}$ are identical. Assume that the scientific goal of the data collection process is to minimize, to the degree possible, the generalized distance between the distribution of characteristics of the final set of participants and that of the full sample by targeting effort to cases still at risk at each point in the data collection period. Nearly every survey has some formal or informal targeting to direct effort, even if it is only high-order response rate goals. For expository convenience, suppose that *a priori* there are *K* observable discrete categories $\mathcal{R}^K$ (where N^k is the number of elements in the *k*th cell and $f^k=N^k/N$) on which we want the population and the cases interviewed to coincide in terms of proportions.

As effort is applied to case *i* in the field at a given step *t*, the case may resolve as complete ($C_t^i=1$) or refused ($C_t^i=2$), or the ultimate outcome may remain censored ($C_t^i=3$). Let effort on case *i* at point *t+1* be given by W_{t+1}^i, which is taken to equal one if effort is applied and zero otherwise; for simplicity the cost of effort is normalized to equal one per application. From the point of view of the surveyor at step *t* planning effort at *t+1*, the probability function for the set of outcomes for a case *i* given the application of additional effort at *t+1* is expressed as $\pi(C_{t+1}^i \mid W_{t+1}^i, W_t^{i*}, I_t^i)$, where W_t^{i*} summarizes all effort from the beginning of the field period to step *t*, and I_t^i denotes the information available to the survey agents about the case at step *t*.

The total cost over the field period (t=1 to T) of all efforts over all cases (i=1 to N) must be within a budget W^*. There may also be additional managerial constraints that limit the number and distribution of interviewers available, specific contractual obligations, etc.; these will be ignored here. For each of the categories $\mathcal{R}^k$ at step *t*, there are n_t^k survey participants, a number which may deviate from the desired proportion in each group.

With the information available at step *t*, the problem is to choose a vector of effort $\mathbf{W}_{t+1}$ to be applied in the next period so that in expectation (denoted E_t) over the whole sequence of potential efforts over the remaining field period, the constraints are satisfied:

$$\left(n_t^k + \mathrm{N}_{t+1}^k\right) \Big/ \sum_{k=1}^{K} \left(n_t^k + \mathrm{N}_{t+1}^k\right) = f^k, \quad \forall k$$

$$\textit{where } \mathrm{N}_{t+1}^k \equiv E_t\left[\sum_{i=1}^{N^k}\sum_{\tau=t+1}^{T} (W_\tau^i = 1) * (C_\tau^i = 1)\right]$$

$$\textit{and } \sum_{k=1}^{K}\sum_{i=1}^{N^k}\sum_{\tau=1}^{T} E_t(W_\tau^{ik} = 1) \le W^*$$

In this example, the expected response rate and the expected length of the field period are endogenously determined at every decision point.[4] The optimal choice to proceed on any given case is a complex dynamic function of the entire sequence probabilities of success on all observations still at risk at *t*. But if the expectation of these probabilities can be calculated for each case and actions across cases are independent, then an obvious decision rule applies: at point *t*, choose to exert effort on the N_{t+1}^k cases in each group with the lowest expected costs relative to the likelihood of completion. In early stages, where I_t^i contains little information, it may be rational to apply effort to "learn" about response propensities of various "types" of cases.

Typical practice of survey organizations that aspire to scientific practice may be less deterministic, and constrained in ways beyond what the model supposes, but at least some key aspects must tend to carry through over time if analysts insist on representative data and if the organization survives economically. Four points from this stylized framework have important practical implications for the analysis of the call records of a survey. First, the relatively easy (likely) cases should tend to be approached and, on average, interviewed first. Second, the distribution of effort across cases is endogenously determined. Third, over the field period, the cases remaining at risk should become increasingly dense in cases that would ultimately refuse if pursued.

A key difference between the assumptions of the model and what is desired in practice is that the equivalent of $\mathcal{R}^k$ is generally very difficult to define—at least in part because the measures of ultimate interest are very often not observable *a priori*, so that proxies must be used. In the case of the SCF, the most common proxies have been response rates in PSUs and at least a minimum level of effort in all areas for the AP sample, and special targeting of respondents by stratum for the list sample.

Three very important practical factors are omitted in the model. First, because the incentives and constraints faced by the different players—interviewers, managers, survey organizations, and sponsors—are not always the same, their views of the optimal application of effort may also differ. Second, actual effort applied in the field is not directly observed by anyone other than the interviewer and perhaps the respondent, and most of what is known is filtered through the interviewer. As a result, the ability to make adjustments to effort is potentially limited by the scope of the instruments managers have to influence interviewers' behavior directly or indirectly. Third, even when the incentives are aligned and some information on effort is observed, it is still often quite difficult, even in principle, to process attempt-level records and related data into a form that could be used by the managers to guide interviewers in achieving a project's goals. Each of these points is deserving of a full treatment in a separate paper.

III Case records in the SCF

SCF interviewers are required to maintain "call records" on all actions taken on each observation in their assignment. Managers, "locaters," and other specialists also may record such data. In addition, cases may be transferred among interviewers. Generally, interviewers enter their call records into their computer based on notes they record while they are working using a paper "face sheet" generated for each case. The primary incentive for interviewers to enter their call records fully and correctly is that this information is used by their managers to judge interviewers' productivity—those who do not enter call records are assumed not to be working.

The information entered into a call record includes the following: the date and time of the operation noted; whether the action was taken in person, by telephone, or by mail; whether the interviewer interacted with the respondent, some other person, or no one; and a working "disposition code" describing the operation or its outcome.

Given the May-to-December field period, approximately 210 days is the longest than any observation could have remained "in play," defined here as the days elapsed between the earliest call record of any sort other than an initial mailing, and the last one. The median case remained in play for over three months, but there is a long right tail of the distribution that runs to the length of the entire field period. For completed and refused LS cases, the distribution of time in play is shifted upward from the distributions for those response groups in the AP sample; in contrast, the distribution for censored LS cases lies below that for the censored AP cases. However, this relatively pure time measure does not give a clear sense of the amount of effort that was actually expended over the period.

Unfortunately, the data in the call records needed for a deeper investigation are flawed, most importantly in that the standards for describing events in terms of disposition codes and other information were not uniform across all interviewers or their immediate managers. In some cases the data recorded may even be seen as internally inconsistent—for example, a case where the disposition code suggests that a respondent refused, but there was no record of a contact with a person. In other cases, there may be multiple reports of a set of events that might better be treated analytically as a single event—for example, an interviewer who made a large number of stops at a house over the course of a day of other work in the neighborhood. In general, for this analysis a record was taken to be any type of "attempt" to contact the respondent, where the record type indicated that the information related to any field event (other than a simple update of an address or a comment) or an appointment, and where the action described was made in person, by telephone, or by mail. Because this definition is relatively loose, it may overstate the level of effort actually applied. Unfortunately, there did not appear to be obvious alternative mechanical definitions that were not clearly overly restrictive, and the important findings appear to be robust to simple perturbations of the definition.

The effort expended on SCF cases tended to be fairly concentrated. For example among the AP sample, 5 percent of the cases accounted for 18 percent of total attempts, and 20 percent for 47 percent of the total; this sort of disproportion also holds over cases viewed separately by final dispositions. The effort measure has a long right-hand tail.

For any type of attempt to reach the respondent, the distributions for refusals and censored cases are shifted to the right of the distribution for the completed cases. This difference serves to indicate both that some completed cases were relatively easy to convince—over 20 percent of those who ultimately agreed did so within three attempts of any sort—and that even observations that gave strong signs of refusing were pursued. Overall, the distributions of attempts for final refusals and censored cases are very similar to each other. By sample type, the clearest difference is the tendency for a larger number of attempts to be needed to secure an interview in the list sample than in the AP sample.

It is almost impossible to determine, even by close examination of the traces of information remaining for individual cases, how likely the censored cases would have been to be completed had additional effort been applied. The formal model presented above suggests that the censored cases should become increasingly like the marginal refusals as the field period progresses, though overall the two groups may differ. Examination of the disposition code recorded in the call records for the last step taken before attempts were suspended indicates that about 70 percent of censored AP cases and almost half of the list sample cases were behaving in such a way that a permanent refusal was imminent. Substantial fractions also appear to have been difficult to locate or contact. About 1 percent of AP cases and about 1.5 percent of list sample cases had started some phase of the interview process but broke off the interview and could not be rescheduled to complete it; from the available evidence, it is doubtful that many of these suspensions were made during the actual main interview, but the call record data are insufficient to make any finer discrimination.

For the final refusals, the recent prior case history is, unsurprisingly, heavily weighted toward various degrees of refusal. In contrast, an examination of previous call entries for completed cases shows a dominant pattern of appointments and other indications of cooperation. However, 33.8 percent of completed AP cases refused at some point during their evolution, versus 70.7 percent of ultimately censored cases.

A probit model for each of the two samples was used to search for other systematic differences between the groups of cases that ultimately refused or were censored.

The independent variables represent the aspects of the sample design; regions of the country; and characteristics of the sample address and surrounding neighborhood, some drawn from interviewers' observations and others from census tract-level data matched to the sample data. According to these models, there were some significant differences between the two response status groups. There were significant regional differences for both samples. AP cases in large apartment buildings were more likely to be censored than to be recorded as final refusals; those living in buildings with a locked lobby or doorman, in neighborhoods with relatively high incomes or with larger proportions of non-Hispanic minorities were more likely to remain censored. For the LS cases, the observations in the sample strata more likely to be wealthy were more likely to be censored, as were those who lived in a building with a doorman, cases where the sample address was not observed, and those who lived in neighborhoods with larger proportions of non-Hispanic minorities. LS cases in neighborhoods with larger proportions of Hispanics were less likely to have censored outcomes. None of the differences have an obvious explanation, but the significance of so many factors indicates the presence of some underlying decision structure that may have varied across field managers and interviewers.

If the application of effort to cases were either random or independent of the expected outcomes, the empirical hazard rates for cases at risk being completed or permanently refused at each application of effort might be used to estimate the expect cost of a given response rate and the length of a field period. But as is no doubt clear at this point, the choice and the outcome are interrelated—the choice to apply effort comes before an interaction with the respondent, and this choice is affected by the subjective probability of completing an interview. One might model jointly the decision to pursue a case and the likelihood of its completion. However, because all the variables that are available for analysis might well enter into both processes (other than the outcome, there is no systematically available information that became available only after each attempt was made), such a model is not statistically identified. Nonetheless, useful information may still be gained by closer examination of effort and nonresponse.

If response probabilities are well assessed and effort is allocated rationally and without constraint on the distribution of cases within monitored outcome grous, one would expect that as more effort is devoted to a sample, the relatively easy cases would be interviewed early, the very resistant cases would refuse firmly, and the remaining cases would become increasingly rich in those that are inclined to refuse; consequently, an increasingly large fraction of cases pursued should ultimately refuse firmly. But at the surface, the data show a different pattern. Over the course of attempts during the field period, the proportion of all cases at risk at each point that ultimately refuses is roughly constant between about 12 and 15 percent; this result also holds separately for the AP and LS cases. At the same time, the proportion of cases ultimately completed declines gradually as the rate for cases that are ultimately censored rises; throughout, AP cases have a lower fraction of ultimately censored cases and a correspondingly higher completion rate than LS cases.

Overall, it appears that continued effort yields an increasing share of ultimate non-interviews (refused and ultimately censored cases) along with a declining payoff in terms of completed cases. The unexpectedly flat refusal rates for the first three measures may reflect reluctance of interviewers and mangers to "give up" on cases, even when the probability of success appears low; among other things, they may think that some of those cases might be "needed" later to meet production quotas.

The choice whether to continue exerting effort on a case is clearly a key factor in the determination of outcomes and the distribution of cases within outcomes. One way of extracting systematic information about the choice to continue effort is to frame the decision as a hazard model. In such a model, the unit events are the elements of the sequences of decision across all cases remaining at risk whether to expend further effort on the case, or to leave the outcome permanently censored. Once a case is completed, refused or permanently censored, it adds no further elements. The choice element is to pursue a case further or to allow the outcome to be permanently censored. Because the model has only two choices, estimation may be performed using a simple logit model of the stacked sequences of decisions (table 1).

A separate model was estimated for each sample type using interviewer observations, census tract-level data, and case administration data derived from the call records. The case administration variables include information specific to each decision point: the number of days a case was in play as of the previous attempt, the number of prior attempts made, an indicator variable for whether contact had ever been made previously, the number of prior contacts, and an indicator for whether the working dispositions codes record any prior refusal by the respondent to participate. Both models show strongly that more days in play, greater numbers of contacts, and a prior refusal lower the frequency with which cases were followed. For the AP cases, the positive effect of the number of prior attempts on the likelihood of continued attempts, which probably captures the repeated calls necessary to make initial contacts, is quickly offset by the negative effect of days in play; the fact that the number of attempts is not a significant factor for the LS cases suggests that there were deeper differences either in the way in which such cases were worked or in the reactions of respondents.

Even with the administrative controls, other variables

also show evidence of significantly different applications of effort across cases.[5] There were strong, but different, geographic effects for the two samples. For the AP cases, those living in mobile homes were less likely and those in apartment buildings were more likely to be followed than those living in single-family homes; those living in areas in areas with larger Hispanic populations were more likely to be followed than those living in other neighborhoods, but the converse was true for cases in neighborhoods with higher fractions of people with limited skills in speaking English; cases in neighborhoods of moderately widely-spaced houses were more likely to be pursued than either cases in more densely or sparsely built neighborhoods. Cases in areas with higher levels of income or higher proportions of people aged 65 and older were less likely to be pursued. One might expect that barriers to contacting the respondent would have a substantial effect, but only the presence of a "gatekeeper" (typically, an employee of the respondent, rather than a literal gatekeeper) has a significant deterring effect on following AP cases.

For the LS cases, there are significant differences in the likelihood of following cases according to their sample stratum, with the strata most likely to be wealthy having the lowest propensity to be followed. Such cases living in a house in worse condition than others in their neighborhood were less likely to be pursued; where the interviewer did not observe the neighborhood, cases were more likely to be followed. Where there was a doorman, LS cases were less likely to be pursued. Those living in neighborhoods with higher fractions of people aged 65 and older and those with higher fractions of Hispanics were more likely to be followed; those in neighborhoods with higher fractions of owner-occupied housing or higher fractions of all types of minorities were less likely to be followed.

Although the sketchy data available for respondents and nonrespondents make it very difficult to coax out a clear structural interpretation of the decision making process in pursuing cases, the models do suggests that there were systematic patterns in the allocation of effort. If variations in effort are not offset by opposite variations in the frequency with which respondents are persuaded to complete an interview, then the distribution of outcomes would be skewed away from the population distribution.

A simple probit model of case completion using all observations and the same non-administrative variables shows that some of the systematic effects in the application of effort remain, but there are also other effects that more likely reflect the difficulty of contacting and persuading respondents. For the AP cases, two key factors on which the models agree are lower effort and response among respondents who have a "gatekeeper" or who live in neighborhoods that have relatively high median incomes. For the LS cases, the key agreements are lower effort and response among cases in the strata most likely to be wealthy and among those living in neighborhoods with higher proportions of minorities.

IV. Alternative case management strategies

The ultimate goal of survey field operations is to collect data that represent the target population as efficiently and with as little bias as possible. Unfortunately, it is generally highly unlikely that every respondent selected will agree to participate. In the absence of specific guidance, interviewers and their managers will perform an 'implicit stratification" of the sample through their decisions to apply effort to the set of cases available to be worked throughout the field period. Thus, a very pressing question is: What guidance on individual case management can we give to interviewers and their managers to help them reach the statistical goals of a survey? To respond, we need both a framework for classifying cases in terms that reflect the statistical objectives of the survey and a mechanism for transmitting sufficiently precise information to and from the field.

Previous SCF efforts late in the field period have typically been targeted to even out large differences in response rates across PSUs for AP cases, and to achieve specific numbers of completed interviews within the sample strata for LS cases. Detailed investigation of nonresponse issues in the SCF led to the use of various post-strata at the weighting adjustment stage to address a set of potential biases (Kennickell and McManus, 1993), and that practice has been refined over time (Kennickell and Woodburn, 1999). However, there has never been any previous effort in the SCF to develop a more detailed case management plan to address potential bias and efficiency issues *during* the field period. In essence, interviewers and managers were allowed to pursue their own "convenience" in case management.

The information available to make management decisions about a sample is often patchy at best. When a sample is first selected, some characteristics of the elements may be known from the sample frame; some additional information might be matched from other sources. As interviewers visit the cases, call records are generated and some other information about the cases is revealed; some such information may be captured systematically. However, much "local" information may be so idiosyncratic as to be difficult to use systematically or insufficiently salient to be noticed in all relevant instances by all interviewers. Because local information may sometimes bear on the evaluation of interviewers' performance, it may also be important to consider ways to manage interviewers' incentives so that they would be willing to reveal such information. As field operations progress, more detailed information becomes available on the set of sample elements that actually complete an interview.

Ideally, in order to reduce bias or estimation variance, cases would be classified dynamically through the field period on the basis of all available information into ones

that should be disproportionately targeted and those that should not. Loosely speaking, cases believed to be "like" existing cases or to show low variability in terms of *a priori* unobservable characteristics within important *a priori* observable groups would be subsampled, and cases believed to be "unlike" existing cases or relatively variable in *a priori* unobservable characteristics within important *a priori* observable groups would be targeted with relatively more effort. Informational and cost limitations inevitably force a compromise.

There are very many possible formal strategies. Each strategy (including the one of allowing interviewers to persist in traditional minimally guided behavior) entails some sort of "model" of what is known and controllable in a sample. In the classical sampling perspective of Hansen and Hurwitz [1946], at some point in a field process, uncompleted cases are randomly subsampled. By forcing effort more intensively onto a smaller number of cases, the idea is that more could be learned about the nonrespondent population (reduced bias) at the cost of some direct variance inflation, but with lower mean squared error if the level of subsampling can be calibrated sufficiently. Depending on ultimate response goals and differences in quantity and reliability of the available information, one might extend this model to differential subsampling rates for different subpopulations. Sudman [1966] offers another perspective. As in the Hansen and Hurwitz model, there is an initial probability sample that experiences nonresponse. Here subsampling is performed ideally using the probability of nonresponse; those with lower probabilities are oversampled and those with higher probabilities are undersampled. Usually the operational implication is taken to be the generation of "quotas" for field staff of certain classes of cases. Although bias reduction would lead to a direct reduction in mean squared error, the direct variance implications of the subsampling are not straightforward, but depend on the interpretation of the operation. A classical interpretation implies variance inflation through increased variability of weights, while a strong model-based interpretation assuming a credible mapping from groups of participants to nonrespondents would not necessarily imply any such variance inflation.

There are many other arguable approaches to subsampling. For example, as suggested earlier, one could use traditional stratification arguments to sample differentially observable groups discovered during the field period to have strongly differing variances for key variables. If much is known about the nonresponse mechanism when a sample is first selected, differential sampling at that point (or the creation of reserve replicates to allow more control of differential sampling later) could lead to efficiency improvements and bias reduction. Clearly, there are many other possibilities blending many of these arguments and others. All subsampling plans should also be examined in light of post-survey adjustment, such as post-stratification, that might otherwise be made or be made to larger effect in the absence of subsampling.

One factor which may conflict with straightforward sample management plans is the drive to make a credible level of effort to inform every selected sample element of the nature of the survey and the respondent's role in the process. Informed refusal (at least taken to the limits of something less than a "hard" refusal) seems as large an ethical concern as informed consent. The effect of a lower standard of work on the behavior of interviewers and their managers could also undermine the key sense of legitimacy field staff require to persuade respondents. Perhaps more seriously, by signaling to interviewers a lower importance of interviewing cases in general, it seems almost certain that new selectivity effects would be induced on survey participation.

A structured initial case management plan entailing significant work on all cases could serve reasonably as the first part of a two-phase sample management plan. In the first phase, all sample observations in the original sample would be subjected to a specified level of effort which would play out through a series of alternative branches depending on the difficulties in contacting or persuading respondents to participate. There are two important informational benefits of enforcing this phase of work uniformly. First, because the endogeneity between the application of effort and expectations of success would be broken, it would be possible to make more meaningful estimates of respondents' propensity to cooperate. Second, more uniform and reliable case-specific information would be available. Together, this information could be used to target resources in a second phase to achieve a bias reduction or an improvement in statistical efficiency. With sufficient information and resources, such targeting could proceed dynamically through the remaining field period. A very important side benefit of the phased approach is that effort should become more predictable and controllable, and thus, costs should also become more predictable and controllable.

One very important issue in moving from a model in which variations in effort are largely ignored (though probably not statistically ingorable) to one in which effort is systematically controlled is that the model of control becomes observable and must be defended directly. The Hansen and Hurwitz model of subsampling, which has the advantage of not requiring any assumptions about the distribution of nonrespondents, can nest fully within the original probability structure of a sample; pursued with sufficient vigor, this approach may reduce some nonresponse biases. But often at least something is known about the sources of nonresponse, and if that information is sufficiently reliable, one should be able to gain by incorporating it into the sample management. One way of incorporating such information is to start with the framework of the Hansen and Hurwitz model and

subsample disproportionately as required to offset nonresponse along dimensions believed to be important for nonresponse. If at the end of the field period the evidence is believed to be strong that the subsampled population differed from the earlier respondents in key ways, then a classical subsampling-adjusted weight could be taken to apply. At the other extreme, if the populations within the subsampled groups were believed to be identical, then no such adjustment would be required. In practice, something intermediate seems more likely to reflect reality, but a formal framework would be need to be developed to support the choice of an optimal intermediate adjustment.

V. Conclusion and future research

Typically, we care about nonresponse because of its implications for bias and inefficiency in the estimation of survey statistics. Nonresponse is a joint product of the degree to which respondents can be persuaded to participate in an interview and the amount of effort expended in the effort to gain cooperation. One root of the problem is in the respondent, and thus cannot be controlled directly. Usually, persuasion and information come to respondents from an interviewer. But unless supplemental information is available, variations in the effort spent in persuasion would be indistinguishable *ex post* from variations in respondents' behaviors.

This paper uses data from the administration of the 2001 SCF to look at the distribution of effort in that survey, and it attempts to draw some conclusions for future practice. Several things emerge clearly. First, there was very substantial variability in the efforts devoted to cases; this variability appears to exceed any reasonable bounds of simple measurement error in the administrative records. Second, the application of effort is correlated with some potentially important characteristics of respondents, even when there are controls for the level of difficulty. Finally, although there is insufficient information to disentangle fully the application of effort and respondents' reactions to effort, the data do indicate that variations in effort have consequences for the distribution of outcomes. The results of the investigation suggest that other surveys might also benefit from a systematic evaluation of variations in effort and its implications for nonresponse.

Many factors may be important in characterizing nonresponse in a given survey. Generally, some such factors are very difficult or impossible to observe directly, and the structure that makes sense of all the factors is not known. In the absence of such information and structure, a very large number of strategies for the application of effort might be equally appropriate. One might simply push for the highest response rate possible, in hopes that this approach, applied over time, would yield at least time series comparability, if not reduced bias in any given period. However, any approach that falls short of specific instructions to interviewers and their managers on how to work the sample cases risks introducing selectivity effects in the set of participants; such a realized sample would inevitably have aspects of a convenience sample. One might apply effort in proportion to the degree of respondents' resistance, though such an approach would very likely imply declining to interview some very "easy" cases and pursuing strong refusals to the point of harassment. Alternatively, if one could develop proxies for some important dimensions of nonresponse, then those proxies might be used systematically to target effort in a staged fashion over the field period.

This paper addresses some problems of targeted and phased effort and proposes a general approach. The first phase would lead every sample case released to the field through a process designed to inform respondents to a degree that would allow them to make an informed decision to participate or an informed initial decision to decline participation. This initial phase might be further controlled through the use of sample replicates that would be released as needed to meet the statistical goals. The second phase would operate more indirectly through control of the sample. A variety of ways of subsampling potentially mixed with differential initial sampling are discussed. Each approach turns at least implicitly on a model of the process that generates nonresponse and what might be done to alter the composition of the nonrespondent population. The implications of subsampling for bias reduction and inflation of estimation variance depend on the interpretation of the model.

Targeted and phased effort of the sort described here has two particularly large potential benefits. First, if the targets are meaningfully related to important nonresponse factors, this approach should tend to reduce bias and perhaps increase some aspects of estimation efficiency. Second, by providing a more structured approach to interviewing practice, it would make field activities more predictable—and most likely, more controllable—as well as ensuring that every case receives a credible minimum exposure to effort. As a subsidiary benefit, forcing a minimum level of effort on every case makes it possible to estimate meaningful models of nonresponse uncontaminated by differential effort and these models could be used to guide further work.

It is hoped that field work for the 2004 SCF will be able to proceed in a two-phased fashion: including a phase of specified effort on all cases and a phase of sample management to reduce nonresponse biases. The clearer administrative information required to implement such a strategy will also be useful in a post-survey evaluation of the 2004 procedures and the design of more refined procedures for the 2007 SCF sample.

Endnotes

1. The opinions stated in this paper are those of the author and do not necessarily reflect the views of the Federal Reserve Board. The author is grateful for discussions with Leslie Athey, Steven Pedlow, and Fritz Scheuren.

This version of the paper is an abridgement of a paper of the same name, which may be found at http://www.federalreserve.gov/pubs/oss/oss2/method.html.

2. See Kennickell (2000b) for a review of SCF methodology and Aizcorbe *et al.* (2003) for a summary of data from the survey.

3. Because the analysis reported in this paper is largely concerned with field procedures, the LS postcard refusals and the deleted cases are excluded from the analysis.

4. This discussion abstracts from technical complications that might cause the solution not to exist.

5. Omission of these administrative variables causes very little qualitative change in the other estimates.

Bibliography

Aizcorbe, A.M., A.B. Kennickell, and K.B. Moore [2003] "Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances," *Federal Reserve Bulletin*, pp. 1-32.

Groves, R.M., J. Van Hoewyk, F. Benson, P. Schultz, P. Maher, L.r Hoelter, W. Mosher, J. Abma, and A. Chandra [2003] "Using Process Data from Computer-Assisted Face to Face Surveys to Help Make Survey Management Decisions," Paper presented at AAPOR.

Hansen, M.H. and W.N. Hurwitz [1946] "The Problem of Non-Response in Surveys," *JASA*, pp. 517-529.

Kennickell, A.B. [2000a] "Asymmetric Information, Interviewer Behavior, and Unit Nonresponse," working paper, http://www.federalreserve.gov/pubs/oss/oss2/method.html.

Kennickell, A.B. [2000b] "Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research," paper presented at AAPOR, http://www.federalreserve.gov/pubs /oss/oss2/method.html.

Kennickell, A.B. [1999a] "Analysis of Nonresponse Effects in the 1995 Survey of Consumer Finances,"*Journal of Official Statistics*, pp. 283-304.

Kennickell, A.B. [1999b] "What Do the 'Late' Cases Tell Us? Evidence from the 1998 Survey of Consumer Finances," paper presented at the International Conference on Survey Nonresponse, Portland, http://www.federalreserve.gov/pubs/oss/oss2/method.html.

Kennickell, A.B. and D.A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section on Survey Research Methods*, 1993 Annual Meetings of the ASA.

Kennickell, A.B. and R.L. Woodburn [1999] "Consistent Weight Design for the 1989, 1992, and 1995 SCFs, and the Distribution of Wealth," *Review of Income and Wealth*, pp. 193-215.

Little, R.A.J. [1993] "Post-Stratification, a modeler's perspective," *JASA*, pp. 1001-1012.

Sudman, S. [1966] "Probability Sampling with Quotas," *JASA*, pp. 749-771.

Table 1: Probit models of propensity to follow an observation; AP and LS cases.

		AP		LS				AP		LS	
							-1	-0.629	#	0.227	
								0.179		*0.288*	
Intercept		5.664	#	4.276	#		0	.		-0.241	
		0.575		*0.230*				.		*0.149*	
LSSTRAT	7	.		-0.174	+	SPACING	3	-0.032		0.088	
		.		*0.093*				*0.115*		*0.080*	
	6	.		-0.3047			2	0.168	*	-0.027	
	#							*0.084*		*0.074*	
		.		*0.055*			0	.		-0.121	
	5	.		-0.799	#			.		*0.140*	
		.		*0.054*		OBSTACL	0	.		-0.134	
	4	.		0.603	#			.		*0.130*	
		.		*0.089*			1	0.281		-0.241	*
	3	.		0.611	#			*0.312*		*0.112*	
		.		*0.098*			3	-0.768	+	0.157	
	2	.		0.844	#			*0.419*		*0.162*	
		.		*0.128*			4	-0.037		-0.058	
PSUTYP	1	-0.228	*	-0.413	*			*0.214*		*0.101*	
		0.108		*0.183*		P_LE17		-0.015		-0.005	
	2	-0.104		0.140				*0.015*		*0.006*	
		0.080		*0.098*		P_GE65		-0.024	*	0.008	*
REGION	4	0.217		-0.322	#			*0.011*		*0.004*	
		0.155		*0.060*		PERCMED		-0.005	#	0.000	
	3	-0.097		0.118	*			*0.002*		*0.000*	
		0.106		*0.056*		PHHWPAI		-0.007		0.023	
	2	0.389	#	0.075				*0.029*		*0.016*	
		0.119		*0.064*		POWNOCC		0.006		-0.005	*
NYC		0.136		-0.242	#			*0.005*		*0.002*	
		0.155		*0.056*		MAGEHU		0.009		-0.000	
LA		-0.394	*	0.162	#	.		*0.005*		*0.002*	
		0.169		*0.061*		PMINOR		-0.002		-0.006	#
NEIBLDG	0	.		-0.384	+			*0.004*		*0.002*	
		.		*0.213*		PNOENG		-0.039	+	-0.014	
	1	-0.189		0.071				*0.020*		*0.008*	
		0.121		*0.0935*		PHISP		0.030	#	0.015	#
								0.008		*0.003*	
	2	-0.243	+	0.182		DAYS		-0.018	#	-0.016	#
		0.130		*0.119*				*0.002*		*0.001*	
BLDGCON	0	.		0.686	#	ATTEMPT		0.036	#	-0.007	
		.		*0.216*				*0.010*		*0.005*	
	1	0.255		-0.237	+	CONTACT		-0.057	#	-0.034	#
.		*0.164*		*0.139*				*0.014*		*0.008*	
	2	0.014		-0.067		EVCONT		0.160		0.042	
		0.120		*0.095*				*0.120*		*0.038*	
TYPBLDG	-4	0.319	+	0.022.		EVREF		-0.846	#	-0.621	#
		0.172		*0.141*				*0.095*		*0.032*	
	-3	0.362	#	0.011		N		32,573		40,975	
		0.132		*0.157*		Likelihood ratio		3,185		10,190	

P-values: #: ≤1%, *: ≤5%, +: ≤10% SE in italics below each parameter estimate.
LSSTRAT: LS stratum (1 is omitted category).
PSUTYP: Overall urbanization of PSU: 1=non-MSA, 2=non-self-representing MSAs (self-representing MSAs is the omitted category).
REGION: Region of the country: 2=north central, 3=south, 4=west (northeast is the omitted category).
NYC: Observation located in the New York City PSU.
LA: Observation located in the Los Angeles PSU.
NEIBLDG: Types of buildings in the neighborhood of the sample address: 0=interviewer did not see the sample address, 1=all residential, 2=mostly residential (omitted category is half or more nonresidential).
BLDGCON: Condition of unit at sample address relative to others in the neighrorhood: 0=interviewer did not observe sample address, 1=others better, 2=about the same (omitted category is others not as good).
TYPBLDG: Type of building at sample address: 0=interviewer did not see sample address, 1=mobile home, 3=building has 2 to 9 units, 4=building has 10 or more units (omitted category is single-family building).
SPACING: Spacing of units in neighborhood: 0=interviewer did not observe the sample address, 2=21 to 100 feet apart, 3=greater than 100 feet apart (omitted category is 20 feet apart of less).
OBSTACL: Obstacle to reaching the sample address: 0=interviewer did not observe the sample address, 1=doorman or guardhouse, 3=other "gatekeeper" at the sample address, 4=locked lobby or locked gate (omitted category is no such obstacle).
P_LE17: % of census tract population age 17 or less.
P_GE65: % of census tract population age 65 and older.
PERCMED: Median income of the census tract as a percent of the area median income.
PHHWPAI: % of households in the census tract receiving public assistance.
POWNOCC: % of housing units in the census tract that are owner occupied.
GE5UNITS: % of housing units in the census tract in buildings with 5 or more units.
MAGEHU: Median age of housing units in the census tract.
PMINOR: Racial and ethnic minorities as a percent of the population of the census tract.
PNOENG: % of the census tract population who speak English "poorly" or "not at all."
PHISP: % of the census tract population reporting Hispanic origin.
DAYS: Number of days from the first attempt on a case to the current attempt.
ATTEMPT: # of attempts made from the first attempt on a case to the current attempt/.
CONTACT: # of contacts made from the first attempt on a case to the current attempt.
EVCONT: Indicator for whether case has ever been contacted as of current attempt.
EVREF: Iindicator for whether R has ever been uncooperative as of current attempt.