J. R. Statist. Soc. A (2006)
169, *Part* 3, *pp.* 439–457

Responsive design for household surveys: tools for actively controlling survey errors and costs

Robert M. Groves and Steven G. Heeringa

University of Michigan, Ann Arbor, and Joint Program in Survey Methodology, College Park, USA

[Received July 2004. Revised February 2006]

Summary. Over the past few years surveys have expanded to new populations, have incorporated measurement of new and more complex substantive issues and have adopted new data collection tools. At the same time there has been a growing reluctance among many household populations to participate in surveys. These factors have combined to present survey designers and survey researchers with increased uncertainty about the performance of any given survey design at any particular point in time. This uncertainty has, in turn, challenged the survey practitioner's ability to control the cost of data collection and quality of resulting statistics. The development of computer-assisted methods for data collection has provided survey researchers with tools to capture a variety of process data ('paradata') that can be used to inform cost–quality trade-off decisions in realtime. The ability to monitor continually the streams of process data and survey data creates the opportunity to alter the design during the course of data collection to improve survey cost efficiency and to achieve more precise, less biased estimates. We label such surveys as 'responsive designs'. The paper defines responsive design and uses examples to illustrate the responsive use of paradata to guide mid-survey decisions affecting the non-response, measurement and sampling variance properties of resulting statistics.

Keywords: Multiphase sampling; Paradata; Propensity models; Responsive design; Survey non-response

1. Introduction

In recent years two phenomena have combined to prompt consideration of new design options for sample surveys of household populations in wealthier countries of the world:

(a) the growing reluctance of the household population to survey requests has increased the effort that is required to obtain interviews and, thereby, the costs of data collection (de Leeuw and de Heer, 2002; Groves and Couper, 1998) and
(b) the growth of computer-assisted data collection efforts provides tools to collect new process data about data collection efforts (Hapuarachchi *et al.*, 1997; Couper, 1998; Scheuren, 2001).

The first phenomenon has threatened survey field budgets with increased risk of cost overruns, missed response rate targets and shortfalls in meeting goals for numbers of interviews. The second phenomenon provides survey researchers with unprecedented amounts of information about the data collection processes. This paper describes a set of design approaches to use the second to ameliorate the damages of the first.

Address for correspondence: Robert M. Groves, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106, USA.
E-mail: bgroves@isr.umich.edu

 0964–1998/06/169439

Established methods for survey design generally adhere to the following three-part recipe: prespecification and standardization of all aspects of design, implementation of those specifications and analysis conditional on the design protocols. These time-tested methods were developed to control sampling and measurement errors in the survey process, and they remain effective in survey applications where survey costs and errors are subject to small uncertainty. Today, however, uncertainty abounds.

The traditional survey design options include various tools to reduce survey errors, but many have cost implications. For example, the presence or absence of sample clustering often affects standard errors of estimates and also costs. Repeated call-backs to reduce non-contact of sample units increases interviewer costs but reduces non-response rates (which are often linked to non-response error). The use of alternative modes of data collection (e.g. mail, telephone or face-to-face interviews) can affect measurement errors, but they also affect costs. The use of incentives can affect rates of co-operation but may increase total costs. Although sampling variance is a common concern in most of these issues, the major focus is bias, increasingly bias due to non-response. How much these alternative features improve the quality of survey statistics and how much they affect costs often have no certain answers before data collection. Hence, prespecified optimal designs are almost never achieved in practice.

In this paper, we describe an approach that is termed 'responsive survey design'. By way of definition, responsive survey designs

(a) preidentify a set of design features potentially affecting costs and errors of survey estimates,
(b) identify a set of indicators of the cost and error properties of those features and monitor those indicators in initial phases of data collection,
(c) alter the features of the survey in subsequent phases based on cost–error trade-off decision rules and
(d) combine data from the separate design phases into a single estimator.

Although there are similarities to sequential analysis in its use of accumulating data, we make no pretence that our concept of responsive design represents a theoretical breakthrough as significant as sequential analysis (Wald, 1947) was for experimental testing. Nor do we wish to claim that we have invented new tools or even substantially refined methods for surveys. Techniques for replicated, two-phase and adaptive sampling (Cochran, 1977; Thompson and Seber, 1996) have been described by others and are used in practice. Likewise, adaptive flexible procedures for questionnaire design, respondent selection and incentives, refusal conversion and other aspects of the survey process are the subject of a substantial literature and have been employed in survey practice.

This paper is a discussion of key components of responsive survey design, defining the terms, and presenting practical examples of the components: 'design phases', in Section 2; the role of randomized experiments, in Section 3; then the notion of 'phase capacity', in Section 4; how the use of process data (or 'paradata') aids responsive design, in Section 5; how early phases can include indicators that are sensitive to error properties being monitored, in Section 6; finally, in Section 7 we describe the notion of complementary design features.

2. The notion of a design phase

Responsive designs are organized about design phases. A design phase is a time period of a data collection during which the same set of sampling frame, mode of data collection, sample design, recruitment protocols and measurement conditions are extant. For example, a survey may start

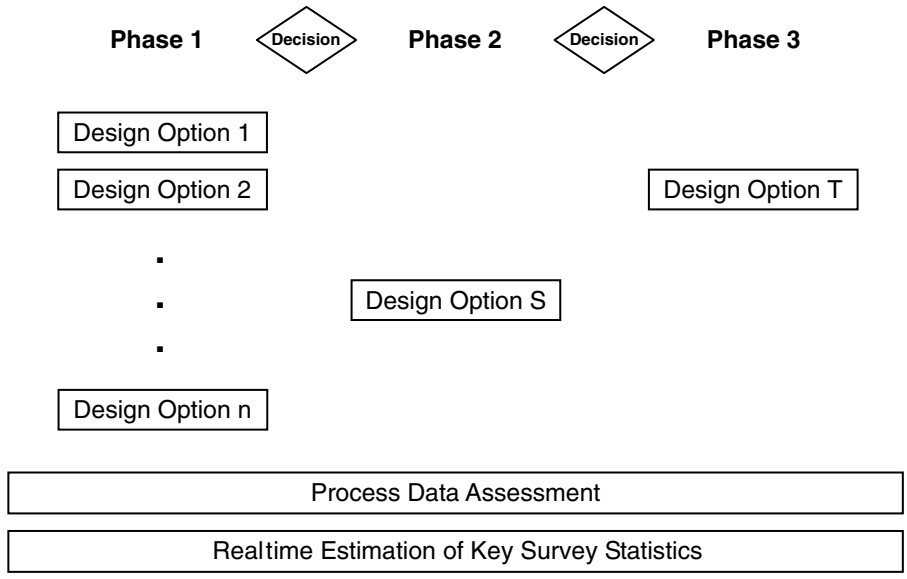

Fig. 1. Illustration of a three-phase responsive design

with a mail questionnaire attempt in the first phase, follow it with a telephone interview phase on non-respondents to the first phase and then have a final third phase of face-to-face interviewing. Sometimes phases are simultaneously conducted, e.g. when there is a randomized set of question modules assigned to sample replicates. Sometimes phases are sequentially conducted in a survey design (we use such an empirical example below) and apply to subsets of the sample respondents that are neither independent nor random samples (e.g. special incentives and procedures for final non-response follow-up, like those which are used in the American Community Survey). Note that this use of 'phase' includes more design features than merely the sample design, which are common to the term 'multiphase sampling' (Neyman, 1938).

Fig. 1 illustrates the key components of a three-phase responsive design, in which the first phase is mounted with n design options that are applied simultaneously (possibly on different replicate subsamples). Examples of these design options might include the level of incentive that is offered, the number of follow-up calls to non-respondent households, the use of a short or long version of a questionnaire or alternatives for the number of sample people to select per household. During phase 1 (as displayed at the bottom of Fig. 1) process data, called paradata (Couper, 1998), are collected to inform the researcher of the interviewer hours that are spent calling on sample households, driving to sample areas, conversing with household members and interviewing sample people.

At the end of phase 1, the researcher makes a decision about the phase 2 design options that appear to be prudent (the middle portion of Fig. 1). This decision will be guided by the information on costs and sensitivity of values and standard errors of key statistics. Phase 2 usually includes, whenever possible, a switch of recruitment protocol for the remaining unresolved phase 1 cases to the phase 2 protocol. Phase 3 is often a phase that is introduced to control the costs of the final stages of data collection while attaining desirable non-response error features for key statistics. This might involve a subsampling of remaining non-respondents, the use of different modes of data collection, the use of larger incentives, etc. After the third phase has been completed, the data from all three phases are combined to produce the final survey estimates.

2.1. Example of design phases

Arguably, the most traditional responsive design option is the use of two-phase sampling for non-response (Hansen and Hurwitz, 1946; Deming, 1953). In the first phase, all possible measurements using an initial mode and recruitment protocol are executed. When all cases that can possibly be measured under the first phase design have been completed, a probability subsample of the non-respondent cases is selected. A more expensive and (theoretically) totally successful method of data collection is applied to this subsample. The resulting sample statistics weight the subsampled cases by the inverse of their second phase selection probability (multiplied by any other selection weights).

It can be shown that such statistics, when linear in form, are unbiased estimates of population parameters, *if the second-phase sample of non-respondents is completely measured.* In practice, complete measurement is never achieved and second-phase response rates are functions of the nature of the first-phase non-response, the mode, incentives and efforts that are used in the second phase. It is possible that the inclusion of the second-phase respondents can increase the bias of the combined estimates. Ideally the researcher can evaluate the effects of the two-phase sampling by using some external criterion.

The Chicago Mind and Body Survey was an epidemiological survey of the Chicago, Illinois, household population and was based on a two-stage area probability sample. Face-to-face interviews, of about 2 hours in length, were administered to about 3100 people. The first phase used a single set of design features and collected about 2000 interviews by using a promised incentive of $60 and call-backs guided by the discretion of the interviewer. In early active monitoring of field costs and production rates, forecasts of the final response rate and number of completed interviews fell below the targets desired.

Phase 2 introduced two new design features. First, a second-phase subsample was drawn, using stratification based on interviewers' subjective assessments of the likelihood that a non-final case would co-operate under the phase 2 data collection protocol. A sampling fraction of 0.5 was used in a stratum that was judged to have low propensities to respond; a sampling fraction of 1.0 was used for the high propensity stratum. The second-phase recruitment protocol increased the incentive from $60 to $100, with a fixed number of call-backs. Finally, a third-phase protocol was implemented, on all remaining non-respondent cases, raising the incentive to $150 and limiting effort to one contact.

2.2. Evaluation of design phase alternatives

There are two evaluative opportunities with this example. Were the interviewers successful in predicting the response propensities for cases in the second phase? Yes, they were. Sample cases that interviewers expected to have low propensities achieved a second-phase response rate of 38.5%; the high propensity stratum, 73.7%. Were the second and third phases effective in avoiding the large inflation of costs per interview that are typical in the end stages of a survey? In the first phase of the survey the average number of interviewer hours per interview was 19.2, and interviewers drove on average 86 miles per interview. We would expect that additional efforts on the remaining non-respondent cases would require more calls per completed interview. Despite this expectation the hours per interview in the second-phase sample were reduced to 15.4, with 87 miles driven. The full cost of an interviewer is approximately $25 per hour (including all indirect costs) and they are reimbursed at $0.375 per driven mile. Taking into account the travel costs also, the second phase protocol produced interviews costing on average $45 less than those of the first phase. The third phase, which changed from an incentive of $100 to $150, required an average 19.9 hours per interview, with 88 miles driven per case. In short, interviewers can provide

useful information for stratification of second-phase samples, and second-phase designs that alter the benefit structure to be more appealing to the remaining non-respondents can achieve attractive cost features.

3. The value of embedded randomized experiments in initial design phases

The initial phases of a responsive design assemble information about the cost and error trade-offs of alternative data collection strategies. In essence the ideal initial phase of the survey provides an evaluation about the performance of all key design alternatives that are considered by the researcher. Thus, after the first phase has been completed, the researcher has sufficient information to identify the optimal design for the second phase of the survey. For example, prepaid incentives are often used to increase response propensities of sample units. Since incentives that have little effect on response propensities unnecessarily increase total costs of the survey, it is useful to know the effects of incentives. In the first phase of the data collection, comparisons between cases offered incentives and those who were not offered incentives would improve the researcher's evaluation of whether later phases should use incentives.

The comparisons of alternative data collection protocols that are mounted in the first phases are best made when the different protocols were applied to units that have the same characteristics on the survey variables, the same base response probabilities and the same measurement error tendencies. Such a state is best achieved by randomized assignment to probability subsamples.

3.1. Example of randomized experiments in initial design phases

The National Comorbidity Survey—Replication (NCS—R) (Kessler *et al*., 2004) was a US national area probability sample survey that was designed to measure the prevalence and severity of mental health disorders in the US household population. Household screening and the majority of interviews were conducted in person, although interviewers were permitted to conduct telephone interviews, once contact with the designated respondent had been established. Because the cost of the NCS—R interview was a function of the unknown prevalences and comorbidities of the mental health disorders, the first phase of the survey prepared for second-phase design contingencies (as in Fig. 1 above). Specifically, the computer-assisted personal interview code for the household screening interview was designed to select more than one sample person in a random subsample of approximately 25% of all households containing two or more eligible adults. In all other households, a single respondent was randomly designated for interview. Phase 1 of the study was therefore structured to evaluate two design options, one *and* two respondents per household. Survey interview data and paradata that were gathered in the experience with the initial sample replicates were used to inform the investigators about the potential costs and errors of selecting up to two respondents in a single household. The decision rule for the preferred within-household sample design was a function of costs, response rates and clustering effects on sampling precision.

The paradata that were available for realtime monitoring of the cost and error properties of the phase 1 design included sample control information such as total calls (in person and by telephone) and the intermediate or final disposition for each call to the sample case. The computer-assisted personal interview survey responses were processed through the mental health diagnostic coding algorithms within several days of completion of the interviews to enable statistical evaluation of the prevalence of disorders in the full NCS—R sample and its subclasses, including second respondents.

3.2. Evaluation of randomized experiments in initial phases

There are three evaluative questions that we can answer for this example. First, did selecting a second adult in a subsample of NCS—R households with two or more eligible adults reduce survey costs? The number of call attempts (additional to those for the first respondent) and the mode of the attempt (in person or by telephone) are indicators of the relative costs of selecting a second adult relative to that required to identify and interview a single designated respondent in a sample household. The average number of calls to complete the second adult interview in a household was 4.7 compared with an average 7.2 calls to complete an interview with a single primary respondent. One contributing factor to this efficiency is the fact that over 18% of all secondary respondents completed the NCS–R on the same visit as the primary respondent. In addition, a higher proportion of all calls to the second adult respondents were made by telephone, avoiding additional travel costs.

A second question concerns the added potential for non-response bias from the decision to select a secondary respondent. The combined screening and interview response rate for the NCS—R primary respondent sample was 70.9%. Conditional on a successful household screen, the response rate for the NCS—R second-adult sample was 80.4%. If we incorporate the 89.7% screening response rate for the total NCS—R household sample, the estimated overall response rate for secondary sample adults is 72.1%—which is slightly better than for the primary respondent sample.

A third major concern in selecting a second adult respondent in the household is that the experience of the primary adult respondent may affect the second adult's willingness to cooperate or bias their responses. We compared primary and second-adult estimates of several key NCS—R mental health diagnostic measures. These estimates are restricted to only sample adults in the subsample of households where two respondents were selected. The only significant difference is in the estimated rate of lifetime experience with major depression—which was 13.4% for the primary respondents and 16.1% for the secondary respondents in the NCS—R sample of households with two respondents. Comparisons based on other *Diagnostic and Statistical Manual* IV mental health diagnoses and a broad set of demographic and socio-economic characteristics found no further significant differences between the primary and second respondents where two sample adults were selected.

The fourth and final tool for evaluating the NCS—R phase 1 design is an empirical comparison of the relative sampling variance for the one *versus* two respondents per household design. Selecting a second respondent introduces intrahousehold correlations in the data that typically will lead to increases in variances of sample estimates based on a given sample size. Offsetting the effects of the intrahousehold correlation, the decision to select a second respondent reduces the variation in the selection weights, reducing variances for weighted estimates. Table 1 presents an empirical evaluation of the effect on the variance of selecting a second respondent from eligible sample households. Estimated design effects (Kish, 1965) are compared for two subsamples of the NCS—R data set. The first subsample includes the 3105 primary and secondary respondents from the phase 1 subsample of households in which two adults were selected. The second subsample is a random selection of 3180 single respondents from the balance of the NCS—R sample households in which a selection of a second respondent was possible (but was not made).

The results in Table 1, although based on a small number of sample estimates, suggest that the NCS—R phase 1 option to select a second adult respondent in eligible households may have resulted in an average increase of 10–15% (prevalence estimates) to as much as 33% (demographic characteristics) in the variance of sample estimates that were contributed by households with two or more eligible adults.

Table 1. Sample design effects of the phase 1 design choice to select two respondents in NCS—R sample households with two or more adults

Sample statistic	*Design effect for prevalence estimates in multiple-adult households*	
	1 respondent per household design	*2 respondents per household design*
Diagnostic and Statistical Manual IV lifetime diagnosis		
Alcohol dependence	1.118	1.066
Drug dependence	0.983	1.290
Generalized anxiety disorder	1.131	1.409
Major depression hierarchy	1.015	1.231
Panic disorder	0.898	1.016
Social phobia	1.224	1.023
Average for diagnoses	1.061	1.172
Demographic characteristics		
Age $\geqslant$65 years	1.240	1.292
High school education	1.819	1.449
Low income	1.226	2.585
Married	—	1.901
US born	2.799	2.449
African American	1.875	2.247
Average for demographic characteristics	1.493	1.987

At the conclusion of phase 1 of the NCS—R, the field budget appeared to be in line with its target. The continuing review of the expected savings in costs and the expected increase in design effects for sample estimates led to a decision not to expand the use of a second adult respondent in phase 2 of the data collection. The value of mounting the first phase with multiple-sampling options was that the decision for the second-phase within-household sample was informed by real field data.

4. The notion of phase capacity

Statistical inference from sample surveys has traditionally conditioned on sample design features (Little, 2004). As is increasingly obvious with probability sample surveys, the estimates can also be affected by the level of effort that is expended to obtain participation of sample units (Lynn *et al.*, 2002). Hence, key to the operation of responsive designs is the notion that each set of design features (e.g. the sample design, mode of data collection and recruitment protocol) brings with it a maximum level of quality for a given cost. Phase capacity is the stable condition of an estimate in a specific design phase, i.e. a limiting value of an estimate that a particular set of design features produces. (Although a full mean-square error treatment of the estimate might be considered, in practice, survey management is focused on biasing potentials of terminating a phase.) When the phase achieves stability of an estimate, it can be said to be 'fully matured' or to have reached its phase capacity. One example of phase capacity is the stability of an estimate as a function of the number of call-backs that are made to sample cases to acquire an interview. Since the productivity of a phase usually declines over call-backs (in terms of adding new complete-data records to the respondent file), the stability of estimates is a function of both the magnitude of late cases to the file and their values on the survey variables.

The existence of a phase capacity for a given statistic requires that, as a sample replicate becomes more fully measured using the features of a given phase, key estimates approach their limiting value under the phase. Thus, when stability is reached for values of key survey estimates, the phase capacity has been reached. Usually the earliest point of stability is cost attractive, i.e. detecting phase capacity as soon as possible permits more resources for later phases. However, it is important to note that not all error properties are functions of effort (e.g. mode switches as a recruitment protocol feature).

4.1. Example of phase capacity analysis

A common outcome is that the early days of the data collection are quite productive of contacts and interviews, but that the last days of the data collection period are quite inefficient. The current theories about survey participation (Groves *et al.*, 1999; Baumgartner *et al.*, 1998) posit that different sets of influences act on sample people to determine their likelihood of participation. For some, the topic of the survey is of great interest; for others, the use of an incentive is important; for others, the sponsor or data collection organization evokes interest. As Groves and Couper (1998) showed, the number of questions and comments by both respondents and interviewers decline over the course of repeated contacts with a sample unit. Hence, because of the declining percentage of interviews that are obtained with each additional call, we speculate that the amount of change in non-response bias itself declines over call-backs. As the phase matures, all of the influences on participation have manifested themselves and on-going efforts generate 'more of the same'—most of all the reasons for refusing or accepting and most of all the situational factors have been experienced by interviewers and respondents.

The US National Survey of Family Growth (NSFG) cycle 6 (National Center for Health Statistics, 2005) is an area probability sample of males and females age 15–44 years. Oversamples of teenagers, African-Americans and Hispanics were introduced so that separate estimates of key fertility statistics could be computed on those groups. Indeed, there were 18 age × gender × race or ethnicity groups that had targeted interview counts. Screening interviews with sample households collected household roster data to identify whether any people who were aged 15–44 years lived in the household. In age-eligible households, one and only one respondent was selected for a 'main' interview. Female main interviews required about 85 min; male interviews, 60 min. The targeted response rate for females was 80%; for males, 75%.

The complete NSFG cycle 6 field period included three responsive design phases and spanned 12 months beginning in March 2002 and ending March 2003. The approximate amount of time that was spent in each of the three phases was, for phase 1, 9 weeks, for phase 2, 37 weeks, and, for phase 3, 6 weeks. The first phase used a quarter sample of primary areas, a reduced interviewer corps and unlimited call-back rules. During the first phase, estimates of several key NSFG statistics were computed routinely.

Using charts like those which are presented in Fig. 2, the staff examined the effect of interviewer effort (as indicated by the number of contact attempts) on key statistics. Fig. 2 has two y-axes and two associated plots: one, the cumulative estimate of the statistic, using all interviews that had been collected on or before that call number. This cumulative plot uses the scale that is provided on the right-hand y-axis. The second plot—corresponding to the left-hand y-axis—is the value of the estimate based on the interviews taken only on a particular call number. As that plot moves to the right, the estimate is based on increasingly fewer cases; for that reason, the estimates become highly variable.

During the course of the data collection period, these plots were examined multiple times. The cumulative plot was examined to see at what call number the estimate began to show some

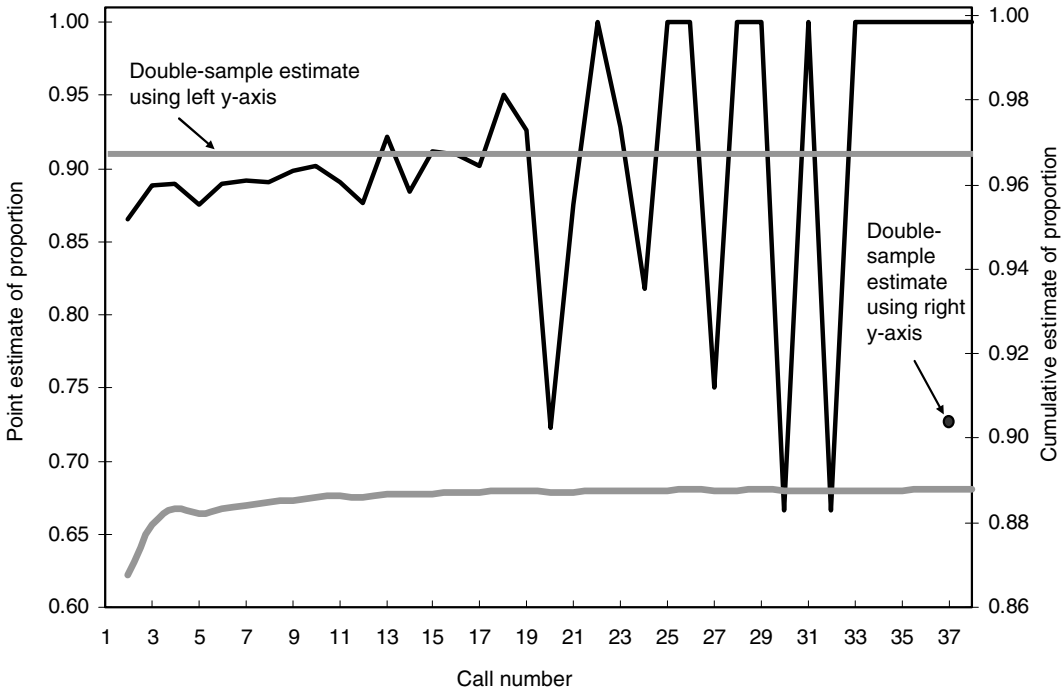

Fig. 2. Estimated proportion of females who ever had sexual intercourse by call number of interview, cumulated (———) and call specific (———)

stability. The call-specific estimate plot was examined for the direction of change in the early calls (i.e. 1–10). When there appeared to be a systematic pattern in the movement of the call-specific estimates, then closer attention was paid to the movement in the cumulative plot to see whether the changes were substantively important. Simultaneously, multivariate models that were estimated on call records and time reports from interviewers tracked the average costs of a call on a sample case.

The conclusion after examining these plots of key estimates over the course of the data collection period was that 10–14 calls produced stable cumulative estimates on the vast majority of the key estimates. ('Stability' here was defined as values that would yield the same substantive conclusion.) This analysis during phase 1 led to the choice of the design option for the later phases that a maximum of 10–14 calls would be made on sample cases. It is useful to note that estimates for some key statistics might display a 'local' maximum or minimum, such that stability of estimates occurs and then disappears at later numbers of calls. Choosing to have the maximum phase 1 effort (at least on some replicates) to be quite large is useful to protect against this outcome.

4.2. Evaluation of phase capacity analysis

On the basis of the phase 1 experience, it is estimated that up to 9% of the screener call attempts were eliminated in phase 2 and 3 screening. A cost analysis suggested that the marginal time that was required for each screener call was 4.2 min. At the volume of interviewer activities that was forecasted for this survey, this represented a saving of approximately 800–1000 interviewer hours for the entire survey.

5. Paradata inputs to design phase evaluation

Responsive designs require enriched process data, which were termed paradata by Couper (1998), to guide decisions that were made during data collection. Paradata can be proxy indicators of costs or errors. They can include data collection administrative data, such as records of contact attempts on sample cases, travel distance and time to reach sample cases and hours spent by interviewers on different tasks during interviewing. They can also include sampling frame data that might be useful in allocating data collection resources, such as the level of urbanicity of a sample unit (which is correlated with the effort that is required to produce an interview), the existence of multiunit structures (which typically require more visits to make contact) or demographic information that might be recorded on the sampling frame. They can also include *de novo* observations taken on sample units that might be proxy indicators of how difficult the sample case might be to gain survey participation, likely characteristics of the unit on key survey variables or whether the measurement errors of completed cases might be distinctive.

Paradata are valuable in responsive designs to the extent that they are good predictors of the propensity of a sample case to participate in the survey, the costs of gaining that participation, the non-response error impacts of failing to gain data on the case or the measurement error properties of data that are obtained from the case. For example, Lynn (2003) demonstrated how observations on multiunit structures and entrance door intercoms predict the amount of effort that is required to contact sample households.

5.1. Example of paradata inputs to design phases

Before the use of computer-assisted interviewing systems, interviewers (and perhaps their supervisors) were given the responsibility of making judgments about the likelihood that individual active sample cases would become respondents. As computer-assisted systems have matured, researchers can use paradata that are collected via the systems to build predictive response propensity models, i.e. it is possible at the end of each working day of the survey to estimate the probability that the next call on a sample case will produce an interview, given a set of values on paradata. To create propensity models that have predictive value, paradata on sample cases must be expanded over those which are normally collected.

In the NSFG, the process information that was collected for the design decisions began at the listing stage of the sample and ended at the last call on the last case. A brief review of the data lies below:

(a) *observations on each interviewer*—
  (i) experience in interviewing tasks and
  (ii) performance on training;
(b) *observations on segments made by people listing addresses*—
  (i) evidence of abandoned or unoccupied structures,
  (ii) extent of commercial, church, school and other activity,
  (iii) impediments to physical access (e.g. walled subdivisions),
  (iv) evidence of non-English speakers in the neighbourhood and
  (v) evidence of safety concerns for the interviewer;
(c) *observations about each listed housing unit made by people listing addresses*—
  (i) impediments to access to the unit (e.g. locked gates),
  (ii) number of housing units in the structure,
  (iii) evidence of children (e.g. toys visible) and

(iv) evidence of an adult at home during the day;

(d) *observations on each call to the unit—*
   (i) the time of day,
   (ii) the day of the week and
   (iii) the outcome of the call;

(e) *observations on each call yielding a contact with a household member—*
   (i) whether the householder asked a question,
   (ii) whether the householder noted that it was a bad time to talk and
   (iii) whether the householder delivered a negative statement about the survey request;

(f) *observations for each interviewer day—*
   (i) the number of hours spent travelling to the segment,
   (ii) the number of hours spent on administrative activities,
   (iii) the number of hours spent attempting screening interviews and
   (iv) the number of hours spent attempting main interviews.

Each of these items of data was included on the basis of prior studies indicating their relationship to the difficulty of contacting sample households or the difficulty of persuading them to co-operate with an NSFG interview.

Such a paradata design yields a nested data structure. Interviewers are the main production units under study; their productivity is affected by the sampling area characteristics that they have been assigned, the characteristics of the sample segment and housing unit, and the nature of calls and contacts on the case. Needless to say, the complexity of the paradata rivals that of the interview data themselves.

We built two models to estimate the propensity of a case to be interviewed on the next call. These models were discrete time hazard models (Singer and Willett, 1993) using call record data as time-varying covariates. They were logistic models predicting the odds of obtaining an interview on the next call, given a set of prior experiences with the sample case. In essence, each call was a separate data record; the data record contained all the observations on the segment and

Table 2. Discrete hazard model coefficients predicting the probability of a screening interview on the next call attempt

Predictor	*Coefficient*
Intercept	−9.37†
Urban = 1	−0.21‡
Access problems in segment = 1	−0.015
Residential neighbourhood = 1	0.11‡
Evidence of non-English speakers = 1	0.031
Evidence of safety concerns = 1	0.088‡
Evidence of unit level access impediments = 1	−0.42‡
Large multiunit structure = 1	−0.10‡
Number of prior calls	−0.12‡
Some prior contact with unit = 1	8.74‡
Number of prior contacts	−0.065‡
Some negative statements by householder = 1	−1.42‡
Last contact statements from householder = 1	0.65
Some questions asked in earlier contact = 1	−0.085‡
Questions asked in last contact = 1	0.056

†$p < 0.05$, assuming a simple random sample of calls.
‡$p < 0.01$, assuming a simple random sample of calls.

individual sample units made by the person listing the addresses and the recorded behaviours of the household members that occurred in prior calls. One model attempted to predict the likelihood that a case that had not yet provided a screener interview (a short household roster determining age eligibility for people in the household) would do so on the next call. It had the predictors that are displayed in Table 2, with their associated coefficients. The model was built in two steps; first a stepwise procedure using the variables at the segment, unit and call level was employed. Then the model was respecified using past literature and theories that were applicable to the response propensity. This second step removed some nonsensical features of the stepwise specification.

Table 2 shows very large positive effects on screener interview propensity of prior contact with the case (coefficient 8.74) and negative effects of expressions of some resistance from the householder (coefficient −1.42). The latter indicator was derived from results of prior studies that showed that what householders say to interviewers is predictive of their later decisions (Groves and Couper, 1998). There are relatively large negative effects of the number of prior calls (coefficient −0.12 for each additional call). Impediments to access reduce the propensity of an interview in general. Cases in large urban areas display lower propensities. We found that these models comported with the expert knowledge of field supervisors but offered a compact statistical summary to sort active cases into different groups based on the likelihood of successful measurement.

Table 3 shows a similar display of coefficients for a model predicting the propensity of a main interview (a long fertility questionnaire) on the next call. This model showed that there

Table 3. Discrete hazard model coefficients predicting the probability of a main interview on the next call attempt

Predictor	*Coefficient*
Intercept	−3.95†
Urban = 1	−0.12†
Uninhabited structures in segment = 1	−0.037
Public housing project = 1	0.071
Residential neighbourhood = 1	0.044
Evidence of non-English speakers = 1	0.023
Evidence of Spanish speakers = 1	0.039
Evidence of safety concerns = 1	0.0081
Some prior contact with unit = 1	4.58†
Resistance displayed on earlier contact = 1	−2.28†
Large multiunit structure = 1	0.13†
Evidence of unit level access impediments = 1	−0.088‡
Evidence of security measures in unit = 1	−0.016
Sample person is a teenager = 1	0.25†
Sample person is male = 1	−0.11†
Sample person is African-American = 1	−0.017
Sample person speaks Spanish = 1	−0.30†
Household has only one member = 1	0.30†
Previous call was a contact = 1	1.50†
Some statements by householder = 1	−1.43†
Last contact statements from householder = 1	0.71
Some questions asked in earlier contact = 1	0.060
Questions asked in last contact = 1	0.24†
Number of prior calls	−0.066†
Number of prior contacts	0.12†

†$p < 0.01$, assuming a simple random sample of calls.
‡$p < 0.05$, assuming a simple random sample of calls.

were strong positive effects of having prior contact with the unit (coefficient 4.58), the prior call being a contact (coefficient 1.50) and the number of prior contacts (coefficient 0.12 for each additional contact). Negative effects of notable magnitude were associated with cases where some resistance was displayed in earlier contact (coefficient −2.28), whether the householder made some statement during an earlier contact (coefficient −1.43) (these tend to be negative statements) and the number of prior calls on the case (coefficient −0.066 for each additional call). The main model could examine the effects of person level characteristics of the selected respondent on the propensity and found the expected positive effects of the respondent being a teenager (coefficient 0.25), negative effects of being male (coefficient −0.11) and negative effects of a Spanish-speaking respondent (coefficient −0.30).

These models were used several times during the data collection period to predict a screening and interview propensity for each active case. The expected propensities for screening and interview were summed over all cases within a sample segment (weighting the screener model expected values by the expected eligibility of the households). Two uses were made of the segment totals of expected values:

(a) during phase 2, segments were grouped into categories with low, medium and high total propensities, for use by supervisors to direct the work of interviewers to the most promising areas, and
(b) at the end of phase 2, segments were grouped into quartiles that formed strata for the phase 3 sample.

We discuss the results of that phase 3 sample design.

5.1.1. Phase 3 sample design

On the basis of the propensity models above, the 783 sample segments of phase 1–2 were stratified on two major dimensions: the number of cases in the segment that were not finalized and the total expected propensities for active cases in the segment.

Phase 3 used a stratified sample of segments, with all non-respondent cases in a selected segment included in the phase 3 sample. This was chosen on the basis of cost model estimates that were computed during phases 1 and 2 that showed that about 38% of an interviewer's working day was spent in travel. If the phase 3 sample maximized the clustering of the subsample, costs could be reduced. Note that this design option placed large emphasis on the cost efficiency of the phase 3 design to produce interviews, not on the minimization of standard errors of the resulting data set. The range of relative sampling fractions over the 16 strata was of the order of 1–4. Simulations before the selection suggested an increase in variance due to the additional selection weighting of approximately 20%. The highest selection probabilities were assigned to those segments with large total expected propensities to be interviewed or large numbers of active cases. The smallest selection probabilities for the phase 3 sample were assigned to segments with few active cases that had low propensities of being interviewed, given the models above.

5.2. Evaluation of use of paradata for design phase decisions

We assess whether the costs of the third-phase efforts varied by the strata that were used in the third-phase sample. This can be addressed with several evaluative indicators. First, did response rates in phase 3 follow patterns that were expected from the stratification by expected propensities? For strata that were defined by expected propensities, the phase 3 response rate was 35% for eligible respondents in the lowest propensity stratum and 54% for cases in the highest propensity stratum. Thus, the stratification by propensity estimates was predictive of the phase 3

response rates. Second, did the costs of data collection vary systematically by propensity strata? The number of interviews that were obtained per call attempt varied only between 0.13 and 0.17 interviews per call. However, the higher propensity strata did have higher cost efficiencies. The strata that were defined by numbers of remaining active cases in the sample segment performed as expected, ranging from 0.13 to 0.20 interviews per contact attempt. As expected the segments with more active cases had higher cost efficiencies.

6. The notion of error-sensitive indicators

A valuable tool in implementing responsive designs is a set of 'leading indicators' of error sensitivity. A leading indicator of error sensitivity is a parameter whose estimate is maximally sensitive to phase capacity. The examples of such leading indicators are easiest for non-response-related phenomena. As Bethlehem (2002) has noted, the bias of the unadjusted respondent mean of variable y can be expressed as

$$\sigma_{yp}/\bar{p}$$

where the numerator is the covariance between the survey variable y and the response propensity p, and the denominator is the mean response propensity for the full sample. If the researcher can theoretically or empirically determine a y that is likely to have maximum covariance with the response propensity, then it can be monitored during the data collection. When response rate increases do not affect the point estimate, then it is likely that the phase in question has reached phase capacity for the given source of non-response being monitored.

6.1. Example of error-sensitive indicators

Statistics that are treated as leading indicators are ideally based on variables that are causes of a component of error that declines as a phase matures. For example, Groves *et al.* (2001) suggested the use of a statistic to measure the maximum level of non-contact error among all statistics in a survey. Candidates for y-variables would be any that are unusually sensitive to the likelihood of contact with the sample unit, i.e. p is the propensity to be contacted, and y is some variable that is highly correlated with that propensity. They examined the percentage of households that are occupied by one person, who is employed outside the home and lives in a unit that is subject to some sort of impediment to access (an answering machine or locked entrance). This is proffered as a candidate that has maximum sensitivity to non-contact error because a proximate cause of non-contact rates is the absence from the home of the sample person. The expectation is that, if the percentage of households with impediments to access and which are occupied by one employed person stabilizes over repeated application of the recruitment protocol (e.g. repeated call-backs), then the phase has reached capacity and some other recruitment protocol would be required to reduce non-contact error.

6.2. Evaluation of error-sensitive indicators

There are bench-mark possibilities for the chosen indicator. In the USA, the monthly labour force survey covers the household population and achieves response rates that are in the mid-90% range. Hence, the estimated percentage of households with impediments to access and which are occupied by one employed person can be obtained from that source. Using that estimate as a target quantity, it was found in a series of random-digit dialling studies that the estimate starts at 2.2% after one call and increases monotonically to 3.2% by 15 call attempts (Groves *et al.*, 2001). Between 10 and 15 call attempts the estimate moves from 3.12 to 3.15. They show

that none of the key estimates of the survey are subject to such great sensitivity to the number of call-backs.

If we assert that this estimate has a greater non-contact bias than any other estimate in the survey, then it achieves 3.12/3.15 or 99% of its phase capacity with a 10-call-back rule. Since each added call-back costs money, such changes in estimates could be used to assess the cost per unit reduction in non-contact error. Following this, decision rules for defining the end of the phase can be constructed.

7. The notion of complementary design features

Another concept in responsive design is that of complementary design features. Complementary design features are those that, when combined, offer minimum error properties among a set of features. They may be recruitment features that are attractive to different parts of the target population, when considering non-response, e.g. telephone contact for those in households with restricted physical access *versus* face-to-face contact. They may be measurement features or mode choices that best fit different statistics, e.g. using self-administered modes for sensitive items but face-to-face modes for items requiring burdensome retrospective recall.

7.1. Example of complementary design features

The NSFG provides an example of an attempt to employ complementary design features that were successful in measuring sample cases that were not measured in the initial phases. The third phase of the NSFG was designed to use a double sample of non-respondents. To assess what recruitment protocol might be successful on those sample people who failed to be measured in the initial protocols, the paradata of the initial phases needed to be studied.

In the NSFG, a set of key estimates were being monitored for sensitivity to extended call-backs and refusal conversions. In addition, the demographic patterns of response rates among screened eligible households were tracked. During this analysis, there were three notable outcomes: sample people aged 15–19 years had higher response rates than those in other age groups; the screener response rate was lower than that of prior similar studies; interviewer variability in co-operation rates were high, despite efforts to reduce that variability.

We sought to invent a phase 3 recruitment protocol that was distinctive from that used in phase 2. Such a distinction is necessary (but not necessarily sufficient) to attract sample people for whom the ingredients of the earlier phase protocol were not effective. The phase 3 recruitment protocol involved the following ingredients:

(a) use of the most productive interviewers on staff (this was an attempt to eliminate interviewers with demonstrated lower effectiveness at increasing base response propensities);
(b) increased use of proxy informants for the screening interview (this was an attempt to lower the burden for obtaining screener information);
(c) a prepaid incentive of $5 (*versus* no incentive) for cases that had not yet completed the screening interview (this was an attempt to increase the benefits of providing screener information);
(d) a prepaid incentive of $40 for the adult main interview (*versus* an incentive of $40 provided after the informed consent was signed) (this was an attempt to set incentives at levels that might be more attractive to sample people who were over 19 years of age);
(e) a promised additional incentive of $40 for a completed main interview.

We limited phase 3 to a 1-month period, following an 11-month combined period for phases 1 and 2.

7.2. Evaluation of complementary design features

The overall response rate at the end of phases 1 and 2 of the NSFG was approximately 64%, using the American Association for Public Opinion Research definition that uses an estimated rate of eligibility among the non-respondent screener cases (American Association for Public Opinion Research, 2004). The third-phase response rate was approximately 40%, which yielded a combined response rate of between 78% and 79%, using the approved American Association for Public Opinion Research double-sample computation. In that sense, the third-phase sample was a very successful increase in 1 month of the overall response rate.

Second, we assess the evidence regarding the non-response bias characteristics of the three-phase design. As always with non-response error, the evidence is indirect only. The NSFG attempts to control achieved interview counts on 18 different subpopulations defined by age, gender and race or ethnicity groups. Examining the phase 1–2 and overall response rates of the 18 subpopulations is one way to examine the imbalance of the phase 1 performance across these 18 subpopulations. The coefficient of variation of the response rates in phases 1 and 2 is 7.6% of the mean response rate; the same measure of variation for the overall rates is 4.4%, which is a large decrease in the variation in response rates. We take this as an indirect indication of reduced non-response error associated with age, gender, race and ethnicity variation. Fig. 3 shows that the variation is quite systematic by age of the sample people. At the end of the first two phases teenagers had a response rate 6 percentage points higher than for those in the oldest age group, with those who were 20–24 years old in between the two. At the end of the third phase the difference in response rates between the three groups was just 3 percentage points.

Are there indications of the effect on estimates of this reduced variation in response rates? We would expect that the effect on estimates of the phase 3 response rates would be to impact all the fertility experience variables that are a function of age. As expected, the proportion of females who ever had sexual intercourse, the lifetime number of sexual partners and a variety of other measures of sexual experience are higher among phase 3 respondents than among the first- and second-phase respondents.

All these findings are consistent with the findings of lower response rate variation by age of the sample people. The first two phases ended with a deficit of respondents who were 20 years old or older. (We believe that this is a function of an enhanced value of the incentive of $40 to teenagers in the phase 1–2 design relative to older sample people.) The third phase was more successful in attracting older respondents. The effect of this on key statistics of the survey is the prevalence of attributes that reflect longer sexually active lives.

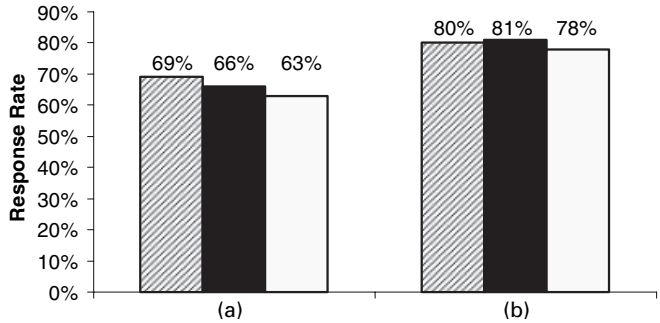

Fig. 3. Response rate by age group by phase: (▨, 15–19 years group; ■, 20–24 years group; □, 25–44 years group): (a) phases 1–2; (b) phases 1–3

8. Summarizing remarks

Responsive designs use paradata to guide changes in features of a data collection to achieve higher quality statistics per unit cost. Responsive designs require the creation and active use of paradata to determine when a phase of the survey has reached its phase capacity and what additional features might be complementary to those of the current phase.

This paper has provided five examples of responsive design features, each of which affects some aspect of survey costs and many of which appear to affect the error properties of resulting statistics. All the examples utilized realtime cost-related data and (proxy) indicators of sampling, non-response or measurement error properties of key survey statistics.

Responsive designs can reduce the cost inflation that is common in the later stages of survey data collection. They can provide a data set with a larger estimated proportion of respondents, when the combined probabilities of selection are reflected. When wise combinations of design options are chosen across sequential phases, responsive designs can offer evidence of reduced non-response errors.

9. Needed next steps in the development of responsive designs

It is appropriate to note that most of the invention of responsive designs has been driven not by formal theory and specified optimal design models but by the practical need to reduce risks of overruns of budget and the increasing non-response rate problem. As with all such developments, practice sometimes outpaces theory. Hence, we note some unanswered questions in responsive designs.

First, it is clear that, since responsive designs combine data from different recruitment, non-response and measurement protocols, the analyst requires an assessment of the effect of non-response and measurement error differences across phases. Assessing the set of alternative design options to be mounted in the first phase of the study, to inform choices of later phases, requires an intelligent assessment of likely cost efficient alternatives to the preferred design. Further, some survey resources are used in mounting the multiple-design options in the early phases. Studies on how best to do this are sorely needed.

Second, paradata are like all other survey data—they need conceptual development, measurement development and pretesting. Paradata are useful to the extent that they are proxy indicators of cost or error properties of the key survey statistics. The fact that they are 'proxy' indicators inherently means that there is a compromise between the rigour of the measurement and the utility of the measurement. The field is just beginning to exploit computer-assisted data collection systems to provide question timing data, digital audio recording of speech, interviewer observations using programmed function keys and complicated question contingencies.

Third, the field needs to study how the survey statistician should best model paradata from early phases. In a real sense, responsive designs are model-assisted designs, not just on sample design issues, but on all the aspects of the data collection. These models, as all models, are imperfect characterizations of the world. They need development, sensitivity analyses for alternative specifications, diagnostic scrutiny, studies of the meaning of outliers, etc.

Finally, variance estimation for survey statistics from multiphase designs with mixed protocols is complicated. Since early phases are used to collect information on cost and error properties for later phase decisions, all aspects of their realizations can contribute to variation in the final estimators combining data from several phases. Currently, variance computations condition on the realized cost and error properties of the initial phases. The use of independent or quasi-independent replicates coextensive with the design phases permits design-based

contrasts across replicates of estimates. The properties of traditional variance estimators for statistics based on combined design phases needs much work.

We expect that the continued pressures on sample surveys to control costs will lead to an increased use of responsive designs. We hope that a concurrent research programme will answer the questions above.

Acknowledgements

The authors express appreciation to John Van Hoewyk, Paul Schulz, Emilia Peytcheva and James Wagner for assistance in preparing the paper. The research was partially supported by a US National Science Foundation grant (SES-0207435; Robert Groves, principal investigator), a contract from the US National Center for Health Statistics (200-2000-07001) and National Institutes for Health grants (3P50HD38986; James House, principal investigator; and U01MH60220, Ronald Kessler, principal investigator).

References

American Association for Public Opinion Research (2004) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Lexana: American Association for Public Opinion Research.

Baumgartner, R., Rathbun, P., Boyle, K., Welsh, M. and Laughland, D. (1998) The effect of prepaid monetary incentives on mail survey response rates and response quality. *A. Conf. American Association of Public Opinion Research, St Louis*.

Bethlehem, J. (2002) Weighting nonresponse adjustments based on auxiliary information. In *Survey Nonresponse* (eds R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), ch. 18, pp. 275–288. New York: Wiley.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.

Couper, M. P. (1998) Measuring survey quality in a CASIC environment. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 41–49.

Deming, W. E. (1953) On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse. *J. Am. Statist. Ass.*, **48**, 743–772.

Groves, R. and Couper, M. (1998) *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., Singer, E. and Corning, A. D. (1999) Leverage-saliency theory of survey participation: description and an illustration. *Publ. Opin. Q.*, **64**, 299–308.

Groves, R., Wissoker, D., Greene, L., McNeeley, M. and Montemarano, D. (2001) Common influences on noncontact nonresponse across household surveys: theory and data. *Meet. American Association for Public Opinion Research*.

Hansen, M. H. and Hurwitz, W. N. (1946) The problem of nonresponse in sample surveys. *J. Am. Statist. Ass.*, **41**, 517–529.

Hapuarachchi, P., March, M. and Wronski, A. (1997) Using statistical methods applicable to autocorrelated processes to analyze survey process quality data. In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), ch. 26, pp. 589–600. New York: Wiley.

Kessler, R., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B., Walters, E., Zaslavsky, A. and Zheng, H. (2004) The U.S. national comorbidity survey replication (NCS-R): an overview of design and field procedures. *Int. J. Meth. Psychiatr. Res.*, **13**, 69–92.

Kish, L. (1965) *Survey Sampling*. New York: Wiley.

de Leeuw, E. and de Heer, W. (2002) Trends in household survey nonresponse: a longitudinal and international comparison. In *Survey Nonresponse* (eds R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), ch. 3, pp. 41–54. New York: Wiley.

Little, R. J. A. (2004) To model or not to model?: competing modes of inference for finite population sampling. *J. Am. Statist. Ass.*, **99**, 546–556.

Lynn, P. (2003) PEDAKSI: methodology for collecting data about survey nonrespondents. *Qual. Quant.*, **37**, 239–261.

Lynn, P., Clarke, P., Martin, J. and Sturgis, P. (2002) The effects of extended interviewer efforts on non-response bias. In *Survey Nonresponse* (eds R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), ch. 9, pp. 135–148. New York: Wiley.

National Center for Health Statistics, Groves, R. M., Benson, G., Mosher, W. D., Rosenbaum, J., Granda, P., Axinn, W., Lepkowski, J. and Chandra, A. (2005) Plan and operation of Cycle 6 of the National Survey of Family Growth. In *Vital and Health Statistics*, vol. 1. Washington DC: Department of Health and Human Services.

Neyman, J. (1938) Contribution to the theory of sampling human populations. *J. Am. Statist. Ass.*, **33**, 101–116.
Scheuren, F. (2001) Macro and micro paradata for survey assessment. *United Nations Wrk Sess. Statistical Metadata, Washington DC*.
Singer, J. D. and Willett, J. B. (1993) It's about time: using discrete-time survival analysis to study duration and timing of events. *J. Educ. Statist.*, **18**, 155–195.
Thompson, S. K. and Seber, G. A. F. (1996) *Adaptive Sampling*. New York: Wiley.
Wald, A. (1947) *Sequential Analysis*. New York: Dover Publications.