Responsive Design for Household Surveys

Steven G. Heeringa and Robert M. Groves, University of Michigan and Joint Program in Survey Methodology
Steven G. Heeringa, Institute for Social Research, 426 Thompson, Ann Arbor, MI 48104.

Keywords: survey nonresponse, responsive design, propensity models, paradata, multi-phase sampling

1. Introduction

Established methods for survey design generally adhere to the following three-part recipe: pre-specification 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 less subject to uncertainty. Today, however, real-time access to information about the survey process and to the accumulating survey data enables survey researchers to analyze survey costs and errors, and to make mid-course decisions and design alterations (Hapuarachchi, March, and Wronski, 1997; Couper, 1998; Scheuren, 2001). In this paper, this is termed “responsive survey design.” This paper defines responsive design and uses examples to illustrate the responsive use of real-time information to guide mid-survey decisions affecting the nonresponse, measurement, and sampling variance properties of resulting statistics.

Over the past decade surveys have expanded to new populations, have incorporated measurement of new and more complex substantive domains, 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 (de Leeuw and de Heer, 2002; Groves and Couper, 1998). 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 tools to capture a variety of process data (“paradata”; Couper, 1998) that can be used to inform cost/quality tradeoff decisions in real time. The ability to continually monitor the streams of process data and survey data creates the opportunity to alter the design during the course of data collection in order to improve survey cost efficiency and achieve more precise, less biased estimates.

We make no pretense that our concept of responsive design represents a theoretical breakthrough. Nor do we wish to claim that we have invented new tools or even substantially refined methods for sample selection, questionnaire design, or survey data collection procedures. Techniques for replicated, two-phase and adaptive sampling (Cochran, 1977; Thompson and Seber, 1996) have been described by others and are used in sampling 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. The theme of this paper is that the entire survey process, from design through data collection should be responsive to both anticipated uncertainties that exist before the survey data collection begins and to real time information obtained throughout the survey data collection .

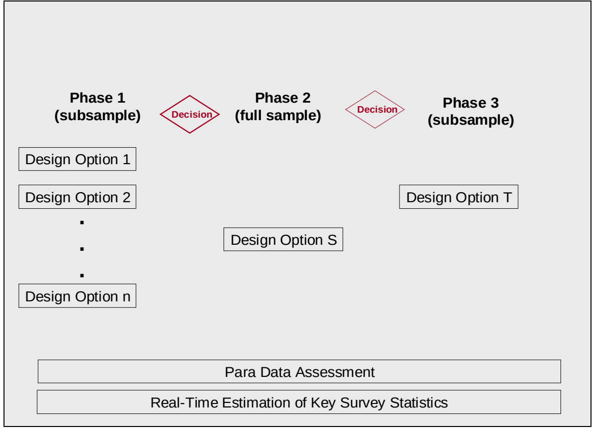
By way of definition, responsive survey designs:

a. pre-identify a set of design features potentially affecting costs and errors of survey statistics;
b. identify a set of indicators of the cost and error properties of those features;
c. monitor those indicators in initial phases of data collection;
d. alter the active features of the survey in subsequent phases based on cost/error tradeoff decision rules; and
e. combine data from the separate design phases into a single estimator.

Figure 1 illustrates the key components of a three phase responsive design, in which the first phase is mounted with N design options applied simultaneously (possibly on different replicate subsamples). Examples of these design options might include whether an incentive is offered, the maximum number of follow-up calls allowed to nonrespondent households, the use of a short or long version of a questionnaire, or alternatives for the number of sample persons to select per household. During Phase 1 (as displayed at the bottom of Figure 1) paradata are collected to inform the researcher of the interviewer hours spent calling on sample households, driving to sample areas, conversing with household members, and interviewing sample persons. The paradata may also include observations about the characteristics of housing units (e.g., whether they have some access impediments) or interactions with contacted sample persons predictive of later actions. Supplementing the paradata are key statistics from the survey analyzed as functions of interviewer effort, and computed on intermediate data sets as interviews are completed.

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 Figure 1). This decision will be guided by the paradata information on costs and sensitivity of values and standard errors of key statistics. Phase 3 is often a phase introduced to control the costs of the final stages of data collection while attaining desirable nonresponse error features for key statistics. This might involve a second phase sampling of remaining nonrespondents, the use of different modes of data collection, the use of larger incentives, etc. After the third phase is complete, the survey data collected in all three phases are combined to produce the final survey estimates.

Figure 1. An Illustration of a Three Phase Responsive Design



This paper reviews how responsive designs, informed by enriched process data, can reduce the uncertainties facing surveys. The next section of the paper introduces a nomenclature for responsive designs. Sections 3 to 5 of the paper provide examples of various responsive design methods that have been employed in surveys and evaluate their use based on paradata, cost, and error properties of the survey statistics. The paper concludes with a summary and an outline of theoretical challenges and next steps in the further development of the responsive survey design method.

2. A Nomenclature for Responsive Design

Responsive designs are organized about "design phases." A design phase is a period of data collection during which the same set of sampling frame, mode of data collection, sample design, recruitment protocols, and measurement conditions are extant. When different design phases are conducted on independent samples (e.g., distinct treatments assigned to sample replicates with known probabilities of selection for each replication/ treatment combination), they offer measurable contrasts of phases using traditional statistical estimators. Sometimes phases are simultaneously conducted; for example, when there is a randomized set of question modules assigned to sample replicates. Sometimes phases are sequentially conducted in a survey design (we provide an empirical example below) but apply to subsets of the sample respondents that are neither independent nor random samples (e.g. special incentives and procedures for final nonresponse follow-up). In such cases, the phase inclusion probabilities for sample elements must be modeled in a fashion similar to the response propensity models that are commonly used in addressing survey nonresponse (Little and Rubin, 2002). Note that this use of "phase" includes more design features than merely the sample design as implied in the term "multi-phase sampling" (Neyman, 1938).

Key to the operation of responsive designs is the notion that each set of design features (e.g., sample design, mode of data collection, recruitment protocol) brings with it a maximum level of quality for a given cost. "Phase capacity" is the minimum error condition for a statistic in a specific design phase; that is, the best outcome for a statistic that a particular set of design features can produce. In practice, phase capacity might be judged by evidence of stability of estimates of a statistic as the phase matures, a plateau of bias-relevant and variance properties of the statistic. When the phase achieves stability of an estimate, it can be said to be "fully matured" or have reached its capacity. One example of phase capacity is the stability of a statistic as a function of the number of callbacks made to sample cases to acquire an interview.

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 statistics for the replicate sample approach their expected 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 phase stability is cost-attractive; detecting phase capacity as soon as possible preserves more resources for later phases. However, it is important to note that not all error properties are functions of effort.

A valuable tool in implementing responsive designs is a set of "leading indicators" of error sensitivity. A "leading indicator" of error sensitivity is a statistic whose estimate is maximally sensitive to phase maturation. Leading indicators are ideally causes of a component of error that a phase's design capacity can remove as it matures. For example, Groves, Wissoker, Greene, McNeeley, and Montemarano (2001) suggest using a statistic to measure the maximum level of noncontact error among all statistics in a survey. They examine the percentage of households occupied by one person, who is employed outside the home, and lives in a unit subject to some sort of access impediment (answering machine, locked entrance). They demonstrate empirically that the design feature of maximum number of calls in a callback rule is a better predictor of the expected value of this statistic than any other statistic in the survey. Thus, this statistic would be a candidate for a leading indicator of noncontact error for the survey.

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 nonresponse; for example, telephone contact may be used for households with restricted physical access versus face-to-face contact for all others. They may be measurement features or mode choices that best fit different statistics; for example, using self-administered modes for sensitive items but face-to-face modes for items requiring burdensome retrospective recall.

The following sections describe examples of responsive design features and actions drawn from recent survey experience at the University of Michigan Survey Research Center (SRC).

3. Example 1: A Sample of Nonrespondents in a Responsive Design

Arguably, the most traditional responsive design option is the use of two-phase sampling for nonresponse (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 are completed, a probability subsample of the nonrespondent cases is selected. A more expensive and (theoretically) a 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).

Example. The Chicago Mind and Body Survey (CMB) 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 3,100 persons. A total of 1145 CMB interviews were collected with a special "focal area" sample for which physical measurement and neighborhood observation data were also obtained. The first phase of the CMB focal area survey design used a single set of design features and collected about 854 interviews using a promised $60 incentive, and callbacks guided by interviewer discretion. In early active monitoring of field costs and production rates, forecasts of the final response rate and number of completed interviews fell below the desired targets. 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 cooperate under the Phase 2 data collection protocol. A sampling fraction of 0.5 was used in a stratum 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 callbacks. Finally, a third phase protocol was implemented, on all remaining nonrespondent cases, raising the incentive to $150 and limiting effort to one additional contact.

Table 1. CMB Responsive Design. Phase 2, 3 Focal Area Nonrespondent Subsampling Outcomes.

Phase	Strat	Intvw Rate	Cum RR2	Hours/ Intvw	Miles/ Intvw
1	Tot	.922	. 509	~19.2	86
2	Tot	-	.616	~15.4	87
	Low	.385			
	High	.737			
3	Tot	-	.712	~19.9	88
	Low	.295			
	High	.479			
1-3	Tot	-	.712	~18.7	88

Evaluation: Table 1 provides a summary of results that help to answer several questions. Were the interviewers successful in predicting the response propensities for cases in the second phase? Yes. The stratum 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 typical in the end stages of a survey? Yes. 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 nonrespondent 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 an average of 87 miles driven for each interview completion. 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 a $100 to a $150 incentive, required an average 19.9 hours per interview, with 88 miles driven per case—slightly greater cost per case compared to Phase 1. 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 nonrespondents can increase response rates and control cost/case.

4. Example 2: Altering Within-Household Sample Design in a Responsive Design

The dominant model of optimal sample cluster size (see Cochran, 1977) minimizes the standard error of a sample statistic conditional on a cost model that includes cluster-specific costs and element costs within a cluster. When such models are applied to within household cluster samples of persons, it is attractive to include the effects of nonresponse and measurement error in the optimization concerns. Specifically, there are concerns about whether selecting two or more persons to respond to a person-level questionnaire generates higher likelihood of the second person refusing or providing answers contaminated by discussions with the first interviewed person.

Example: The National Comorbidity Survey-Replication (NCS-R; Kessler et al., 2004), was a US national area probability sample survey designed to measure the prevalence and severity of mental health disorders in the U.S. household population. Household screening and the majority of interviews were conducted face-to-face (FtF), although interviewers were permitted to conduct telephone interviews once contact with the designated respondent was established. Because the length of the NCS-R interview and therefore its cost was a function of the unknown prevalence and comorbidity of the mental health disorders, the first phase of the survey prepared for second phase design contingencies (as in Figure 1 above). Specifically, the CAPI 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 and paradata 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 available for real time 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 CAPI survey responses were processed through the mental health diagnostic coding algorithms within several days of interview completion to enable statistical evaluation of the prevalence of disorders in the full NCS-R sample and its subclasses, including second respondents.

Evaluation: There are four evaluative questions we can address for this example. First, did selecting a second adult in a subsample of NCS-R households with 2+ eligible adults reduce survey costs? Unfortunately in complex field operations it is difficult to precisely attribute costs to co-mingled survey activities such as household contact and screening, interviewing a first respondent in the household and interviewing a second eligible adult. However, the NCS-R paradata did provide reliable information on the number of call attempts required to contact and interview each respondent case. The number of call attempts and the mode of the attempt (FtF, 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.

Table 2 summarizes the call experience with the NCS-R primary and second adult respondents. The call distributions summarized in Table 1 illustrate significantly less interviewer time and labor for the secondary respondent than screening and interviewing an additional primary respondent in another sample household. The average number of calls to complete the second adult interview in a household was 4.7 compared to 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 second adult respondents were made by telephone, avoiding additional travel costs.

Table 2. Percentage Distribution of Number of Calls Required to Complete Interview for Primary and Secondary Respondents in the NCS-R

No. of Calls	% of Interviews by Respondent Type					
	Primary R			Second Adult		
	Tot	FtF	Tel	Tot	FtF	Tel
0	0%	0%	37%	0%	25%	31%
1	9	18	19	18	33	17
2	13	23	12	26	26	15
3	14	17	7	15	7	11
4	11	11	5	10	3	6
5+	53	31	19	31	6	20
Total	100	100	100	100	100	100
Mean	7.2	4.2	3.0	4.7	1.6	3.2
Med	5.0	3.0	1.0	3.0	1.0	2.0

A second question concerns the added potential for nonresponse 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 persons is 72.1% -- 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. To test the possibility of this nonresponse/response bias, Table 3 compares 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.
In this comparison and additional analyses not shown here, the only significant difference is in the estimated rate of lifetime experience with major depression— 13.4% for the primary respondents and 16.1% for the secondary respondents in the NCS-R sample of households with 2 respondents. Comparisons based on other DSM-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.

Table 3. Estimates of Prevalence for Mental Health Diagnoses for Primary and Secondary Respondents in NCS-R Sample Households With 2 Designated Respondents.

DSM-IV Lifetime Diagnosis of:	Prevalence Estimate	
	Primary	Second
Alcohol Dependence	4.9 %	4.4%
Drug Dependence	3.1 %	2.3%
Generalized Anxiety	6.1 %	6.9%
Major Depression	13.4 %	16.1%
Panic Disorder	3.6 %	3.4%
Social Phobia	11.7%	10.9%

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 vs. two respondents per household design. Selecting a second respondent introduces intra-household 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 intra-household correlation, the decision to select a second respondent reduces the variation in the selection weights, reducing variances for weighted estimates.

Table 4 presents an empirical evaluation of the variance impact 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 3,105 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 3,180 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 subsampling of this second group is performed to standardize the comparison on equal size samples of adult respondents and distribution across the strata and clusters of the NCS-R complex sample design.

The results in Table 4, although based on a small number of sample statistics, suggest that the NCS-R Phase 1 option to select a second adult respondent ineligible 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 contributed by households with 2 or more eligible adults.

Table 4. Sample Design Effects of the Phase 1 Design Choice to Select 2 Respondents in NCS-R Sample Households with 2 or More Adults

	Design Effect of Prevalence Esimate	
Item	One R Option	Two R Option
DSM-IV Lifetime Diagnosis		
Alcohol Dependence	1.118	1.066
Drug Dependence	0.983	1.290
Generalized Anxiety Disorder	1.131	1.409
Major Depression w/Hier	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 65+	1.240	1.292
High School Education	1.819	1.449
Low Income	1.226	2.585
Married	-	1.901
U.S. Born	2.799	2.449
African American	1.875	2.247
Average for demographic	*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 cost savings 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 procedure was informed by real field data.

5. Example 3: Assessing Phase Capacity Regarding Callback Rules

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, Singer, and Corning, 1999; Baumgartner and Rathbun, 1998) posit that different sets of influences act on sample persons 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) show, the number of questions and comments by both respondents and interviewers decline over the course of repeated contacts with a sample unit. Hence, we deduce that as the number of calls and contacts increase over the course of a data collection period, that the amount of change in nonresponse bias itself declines. This must be true because of the declining percentage of interviews obtained with each additional call. However, under the theory the phenomenon occurs also because the amount of change in the causes of the participation decision declines over the course of the study. Most all the reasons for refusing and accepting, most all the situational factors have been experienced by interviewers and respondents.

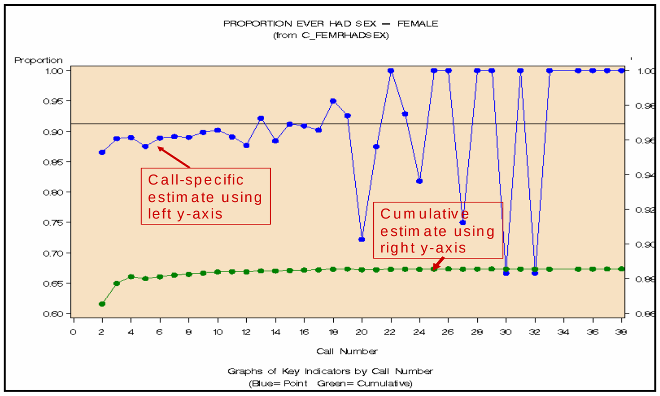
Example: The US National Survey of Family Growth (NSFG) Cycle 6 is an area probability sample of males and females age 15-44. Oversamples of teenagers, African-Americans, and Hispanics were introduced in order that separate estimates of key fertility statistics could be computed on those groups. Indeed, there were 18 age x gender x race/ethnicity groups that had targeted interview counts. Screening interviews with sample households collected household roster data in order to identify whether any persons 15-44 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 minutes; male interviews, 60 minutes. The targeted response rate for females was 80%; for males, 75%.

The first phase used a one-quarter sample of primary areas, a reduced interviewer corps, and unlimited callback rules. During the first phase, estimates of several key NSFG statistics were computed routinely. Using charts like those presented in Figure 2, the staff examined the impact of interviewer effort on key statistics (as indicated by number of contact attempts). Figure 2 has two y-axes and two associated plots; one, the cumulative estimate of the statistic, using all interviews collected on or before that call number. This cumulative graph uses the right y-axis and is very unchanging in its height. The second plot—corresponding to the left y axis-- is a much more variable plot. It is the value of the statistic based on the interviews taken only on a particular call number. As that plot moves to the right, the statistic is based on fewer and fewer cases; for that reason, the estimates become very erratic.

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 stability. The call-specific statistic plot was examined to look 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 important substantively. Simultaneously, multivariate models 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 statistics over the course of the data collection period was that 10-14 calls produced stable cumulative estimates on the vast majority of the key statistics. ("Stability" here was defined as values that would yield the same substantive conclusion.) This analysis during the first phase 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.

Figure 2. Estimated Proportion Females Who Ever Had Sex by Call Number of Interview, Cumulated and Call-specific.



Evaluation: Based on the Phase 1 experience, it is estimated that up to 9% of the screener call attempts were eliminated in Phase 2 and 3 screening. Separate paradata models suggested that marginal time required for each screener call was 4.2 minutes. At the volume of interviewer activities forecasted for this survey, this represented a saving of approximately 800-1,000 interviewer hours for the entire survey.

6. Summary

Responsive designs use paradata to guide changes in features of a data collection in order to maximize the quality of estimates 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 three examples of responsive design features, each of which affects some aspect of survey costs and the error properties of resulting estimates. All of the examples utilized real time cost-related data and (proxy) indicators of sampling, nonresponse, or measurement error properties of key survey statistics.

Responsive designs can reduce the cost inflation common in the later stages of survey data collection. When wise combinations of design options are chosen across sequential phases, responsive designs can offer evidence of reduced nonresponse errors.

7. 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 budget overruns and high nonresponse rates. 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, nonresponse, and measurement protocols, the analyst requires assessment of the impact of nonresponse, and measurement error differences across phases. Assessing the set of alternative design options to be mounted in the first phase of the study, in order to inform choices of later phases, requires 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 they are proxy indicators of cost or error properties of the key survey estimates. The fact that they are "proxy" indicators inherently means there is a compromise between the rigor 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 multi-phase 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. The variance computations currently used condition on the realized cost and error properties of the initial phases. The use of independent or quasi-independent replicates to be 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 increased use of responsive designs. We hope that a simultaneous research agenda will answer the questions above.

8. References

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

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," paper presented at the Annual Conference of the American Association of Public Opinion Research, St. Louis, Missouri.

Cochran, W.G. (1977), Sampling Techniques, 3rd Edition. John Wiley and Sons, New York.

Couper, M.P. (1998),,"Measuring Survey Quality in a CASIC Environment." *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1998, pp. 41-49.

De Leeuw, E., and de Heer, W. (2002), "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison," Chapter 3 in Groves, R., Dillman, D., Eltinge, J., and Little, R.J.A., *Survey Nonresponse*, pp. 41-54, New York: Wiley.

Deming, W.E (1953), "On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse," *Journal of the American Statistical Association*, 48, pp. 743-772.

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

Groves, R. M., Singer, E. and Corning, A. D. (1999), "Leverage-Saliency Theory of Survey Participation: Description and an Illustration", *Public Opinion Quarterly*, **64**, pp. 299-308.

Hansen, M. H. and Hurvitz, W. N. (1946), "The Problem of Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, **41**, pp. 517-529.

Hapuarachchi, Piyasena; March, Mary; and Wronski, Adam (1997), "Using Statistical Methods Applicable to Autocorrelated Processes to Analyze Survey Process Quality Data," Chapter 26 in Lyberg, L.; Biemer, P.; Collins, M.; de Leeuw, E.; Dippo, C.; Schwarz, N., and Trewin, D. (eds.) *Survey Measurement and Process Quality*, New York: Wiley, pp. 589-600.

Kersten, H.M.P. and Bethlehem, J.G. (1984), "Exploring and Reducing the Nonresponse Bias by Asking the Basic Question," *The Statistical Journal of the United Nations Commission for Europe*, pp. 258-263.

Kessler, R., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B., Walters, E., Zaslavsky, A., Zheng, H. (2004), "The U.S. National Comorbidity Survey Replication (NCS-R): An Overview of Design and Field Procedures". Manuscript submitted for review by the *International Journal of Methods in Psychiatric Research*, February 2004.

Kish, L. (1965), Survey Sampling. John Wiley and Sons, New York.

Little, R.J.A. and Rubin, D.B. (2002), Statistical Analysis with Missing Data, Second Edition. John Wiley and Sons, New York.

Neyman, J. (1938), "Contribution to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, Vol. 33, pp. 101-116.

Thompson, S.K. and Seber, G.A.F. (1996), Adaptive Sampling. John Wiley and Sons, New York.

Scheuren, Fritz (2001), "Macro and Micro Paradata for Survey Assessment," paper presented at the United Nations Work Session on Statistical Metadata, Washington, DC, 16pp.