Using Paradata and Responsive Design to Manage Survey Nonresponse

Couper, Mick P.
University of Michigan Institute for Social Research
P.O. Box 1248
Ann Arbor, MI 48106, U.S.A.
E-mail: mcouper@umich.edu

Wagner, James
University of Michigan Institute for Social Research
P.O. Box 1248
Ann Arbor, MI 48106, U.S.A.
E-mail: jameswag@isr.umich.edu

1. Introduction

Groves and Heeringa introduced the notion of responsive survey design in a 2006 paper. Since that time, a number of applications of responsive design have been developed (e.g., Mohl and Laflamme, 2007; Peytchev et al., 2010). At about the same time, there has been work on the development of adaptive survey designs (Wagner, 2008; Schouten and Calinescu, 2010). In this paper we review these developments and attempt to elucidate their differences and similarities. We then present several case studies of our work in responsive design to illustrate the new approach to survey design. We end with some observations on this burgeoning field of research in survey methods.

1.1 Background: Responsive or Adaptive Designs

An explicit element of Groves and Heeringa's (2006) definition of responsive design is the notion of two (or more) phases of data collection. They define a design phase as "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." This notion is consistent with the early work on two-phase samples (Deming, 1953; Hansen and Hurwitz, 1946).

We view the notion of responsiveness as occurring only in separate phases as too restrictive. In fact, in Groves and colleagues' implementation of responsive design in the National Survey of Family Growth (NSFG, see Groves et al., 2009; Lepkowski et al. 2010, 2011), changes in protocol during the first phase of data collection are common, as are the more explicit subsampling and changed protocols that were characterized in the Groves and Heeringa paper.

Wagner (2008, 2011) developed the notion of adaptive or dynamic treatment regimes from the world of clinical trials (see, e.g., Murphy, 2003, 2005) where a sequence of treatments may be attempted until "success" (response in the survey world) is obtained. An optimal adaptive treatment regime is the sequence of treatments with the highest probability of success. In contrast, a strategy that maximizes response for each step in isolation may achieve a lower overall probability of response. In adaptive design, the treatments (e.g., timing of call, incentive, mode switch, etc.) are tailored to the individual, including to the history of previous treatments applied to each individual.

Schouten and Calinescu (2010) and colleagues at Statistics Netherlands use the term "adaptive" to describe different treatments or protocols assigned to different groups at the start of data collection, based on frame data. But they also note that such treatments may depend on data observed during data collection. They call the former static adaptive designs, and the latter dynamic adaptive designs.

This recent literature has led to some confusion over whether adaptive designs are different from responsive designs, or how the two are (or ought to be) related. We argue that the two terms can be used interchangeably, and we use (for convenience) the term "responsive" here. Further, there are some who argue that responsive or adaptive designs are nothing new and survey organizations have been responding to changing circumstances for as long as surveys have been around. There are a variety of different ways in which a survey design can be responsive.

In adaptive design, the treatment is usually tailored to the individual and the tailoring variable is time-varying, i.e., it changes during the field period (Wagner, 2008). If we think of treatments targeted at groups of individual cases sharing a common set of features, and if we relax the notion of explicit phases, then we can think of adaptive and responsive designs as sharing the same continuum. At one extreme (where there is no responsive design) the entire sample gets the same protocol for the entire survey period. Many mail surveys are examples of this type, where the protocol is set at the design stage, and all sample cases are treated equally. Further along on this continuum are the adaptive designs described by Schouten and Calinescu (2010) where the sample is classified into different groups at the outset, and different (fixed) protocols are applied to the different groups. Further along still are the responsive design described by Groves and Heeringa, where the survey is divided into separate phases, and different protocol are applied to subgroups in later phases based on information gathered in early phases. Finally, at the other end of the spectrum as the individually-tailored, time-varying adaptive treatment regimes described by Wagner (2008).

Key elements of responsive design that distinguish the approach from earlier efforts to change field protocols in mid-course include the following (consistent with Groves and Heeringa):

1) Monitoring of key indicators (i.e., the use of statistical process control methods to track activities)
2) Statistically-based decisions to alter the design based on the evidence and targeted at one or more measurable outcomes;
3) Interventions targeted at particular subsets of the sample (rather than more broadly changing a strategy for the entire sample);
4) Documentation of the decision process; and
5) Evaluation of the success of the intervention.

With regard to the third feature, a subset of cases could be as small as an individual case (as in the extremes of adaptive treatment regimes), or it could be a subsample of cases selected for a phase 2 treatment (as in the Groves and Heeringa responsive design perspective). But changes in protocol targeted at the entire (remaining) sample, such as an increase in incentives for everyone as a last-ditch effort, do not qualify as responsive, in our view.

1.2. Responsive Design Notation

We will use the following notation. At the beginning of each time period or phase t in the interval $\{1,2,3,\ldots,K\}$, sampled unit $i=1,\ldots,n$ will have its current status encoded in the vector $\boldsymbol{S_{it}}$. This vector may include fixed covariates that are known at the time the unit is sampled. For a sample of housing units, an example of this type of covariates is the characteristics of the neighborhood available from a census. They may also include interviewer observations about the sampled housing unit itself, as opposed to contact with persons living in the unit. The vector $\boldsymbol{S_{it}}$ also includes time-varying covariates. These covariates are generally drawn from paradata and reflect the history of previous attempts (for example, call records or interviewer observations about contacts).

Pre-specified rules determine when a sampled unit will iterate to the next time period or phase. We can imagine at least two types of rules. One type of rule is specified for groups of cases. For instance, one rule might specify that when a particular response rate has been achieved, a subset of cases will be selected and a new protocol will be applied to these cases. A second type of rule is specified at the case level. For example, when a case achieves a specified number of telephone calls, it is switched to a face-to-face mode.

At each time period or phase, an action is taken. The action taken at time period t on unit i is $\boldsymbol{a_{it}}$, which may be encoded as a vector. If there are several actions to be taken, then we will call this set of actions a protocol. An example vector with two components would be to call the case again ($a_{1it=1}$) and offer an incentive ($a_{2it=1}$). One possible action might be to stop attempting to interview the unit. The actions at time periods 1 to $t-1$ will become

part of the status encoded in the vector $\boldsymbol{S_{it}}$. In this way, prior treatments $\boldsymbol{a_{i1}}$ to $\boldsymbol{a_{it-1}}$ can be used as predictors when determining what the next treatment at time *t* should be.

Finally, we define the outcome variable Y_i for the i^{th} unit. The variable is the final outcome at the end of the time series. This outcome can vary across surveys. It might be a response indicator, where $Y_i=1$ indicates that the i^{th} unit responded while $Y_i=0$ indicates that the i^{th} unit did not respond. Other outcomes are possible. The measure can be continuous or categorical. Other outcomes might relate to measurement error or sample balance.

With this setup, we can imagine analyses aimed at many different goals. We can define objective functions that will need to be minimized or maximized to meet these goals. For instance, we might want to minimize total effort to reach a target response rate. If $Y_i=1$ indicates response, $\bar{Y}^*$ is the target response rate, and $a_{1it}=1$ indicates that we will place an additional call at time *t* (and implicitly assume that each call has the same cost), then our goal is the following:

$$\text{Minimize} \quad \sum_{t=1}^{K}\sum_{i=1}^{n} a_{1it}$$

$$\text{subject to} \quad \frac{\sum_{i=1}^{n} Y_i}{n} \geq \bar{Y}^* .$$

We might also want to produce the most balanced dataset (with balance defined using R-Indicators, see Schouten, Cobben, and Bethlehem [2009]) possible for a fixed budget. In this case, we have the R-Indicator, denoted $\hat{R}(\rho)$, as the function that we hope to maximize. The R-Indicator is a function of the variability of the estimated response propensities. We can reduce this variance by matching protocols to cases subject to budgetary constraints. For example, we might offer higher incentives to lower propensity cases. The series of actions are denoted $\boldsymbol{a_{i1}, a_{i2}, \ldots, a_{ik}}$. Each of these vectors has *p* components, indicating that there are *p* possible actions. In this case, each component of these vectors is a binary variable indicating whether that action is undertaken for that unit at the specified time. Each of these actions has an associated cost (assumed to not vary over time), which results in the $1\times p$ dimensional vector $\boldsymbol{c}$. For example, if the first element is $\boldsymbol{a_{i1}}$ is an indicator as to whether a call is undertaken at time 1, then the first element of $\boldsymbol{c}$ is the cost of a call. The total budget is denoted *C*. Our goal is the following

$$\text{Maximize} \qquad \hat{R}(\rho)$$

$$\text{Subject to} \qquad \sum_{t=1}^{K}\sum_{i=1}^{n} \mathbf{c}'\mathbf{a}_{\mathbf{it}} \leq C ,$$

$$\hat{R}(\rho_i) = f(\mathbf{S}_{\mathbf{iK}}) .$$

The latter constraint simply indicates that the response probabilities are a function of the status of a case – a logistic regression model being one such function. The status includes both fixed covariates and the history of previous actions. It might include interactions of the two types of covariates. In other words, we can impact the response propensities by our choice of the sequence of actions $\boldsymbol{a_{i1}, a_{i2}, \ldots, a_{ik}}$. To the extent that there are interactions between the actions and the fixed characteristics of each case, the protocol can be tailored to the case. For example, an interaction between the indicator for offering an incentive and a fixed characteristic like urbanicity might suggest tailoring the incentive to urbanicity. The product of these analyses will be a set of rules that govern the actions that are taken at each time *t* for each unit.

2. Case Studies: Examples of Responsive Designs

In this section we describe a few examples of responsive design to illustrate the different approaches that could fall under the same broad rubric.

2.1 NSFG

As mentioned earlier, the National Survey of Family Growth (NSFG) has employed several different types of responsive design interventions. We briefly describe three here to illustrate the range of different responsive designs that are possible (for further details, see Lepkowski et al. 2011).

The NSFG is an ongoing cross-section, multistage area probability sample of households. In each sample household interviewers make a roster of members and select at random, one person aged 15-44. The interviewer then seeks a 60-80 minute interview from that person. The questionnaire contains questions on sexual and fertility experiences of the respondent. More sensitive items are administered using ACASI. Each year of the NSFG consists of four replicate samples, introduced at the beginning of each quarter. The full data collection period lasts 48 weeks. At any one point the sample consists of 25 nonself-representing areas and 8 self-representing areas, with about 38 interviewers working.

Interviewers use CAPI and a computerized case management system. Production is monitored on a daily basis, and interventions are targeted at increasing effort on sample cases to achieve desired goals. The first type of intervention targets interviewer effort placed on screener calls versus main interview calls at a specific point in each quarter. There is a tradeoff between these two activities – the more effort interviewers spend on screening to determine eligible households, the less time they will have to complete the main interviews. On the other hand, if interviewers focused on completing each main interview immediately after screening in an eligible respondent, many screener cases may remain incomplete. Managers monitor the ratio of main calls to screener calls, and when it is judged that this ratio would yield an insufficiently lower screener response rate, screener cases are targeted. This intervention has bee successful employed across 16 quarters of NSFG data collection to optimize the effort targeted at each activity. While effort was successfully increased on screener cases during these targeted times, there was considerable variation across quarters in the yield of screener interviewers resulting from these interventions.

The second type of intervention involves classifying a randomly selected subset of active cases meeting certain criteria as ‘high priority’ cases, and comparing effort and production for the high priority cases to a control group of active cases meeting the same criteria. Across four years of NFSG data collection, 16 different interventions were targeted at a variety of key subgroups of interest. The goal of these interventions were to increase response rates for these targeted subgroups relate to other groups to help ensure data set balance. While effort was consistently higher in the targeted treatment groups than the control, these yielded significantly higher response rates in only 2 of the 16 cases, with non-significantly higher rates for the treatment groups in an additional 10 cases. Sample sizes were generally small (as low as 20 for some treatment groups), limiting power to detect differences. In general, this suggests some success across the 16 interventions, but there is room for improvement in learning what kinds o interventions work better than others.

The third type of intervention involves late quarter targeting of important cases for improving overall response rates. This employs the two-phase design approach described in Groves and Heeringa (2006). Using response propensity models developed during phase 1, active cases that have large selection weights and high probabilities of an interview are subsampled for additional effort in phase 2. This strategy has been found to decrease the coefficient of variation of key estimates across major subgroups of interest, and decrease estimates of nonresponse bias for key indicators.

2.2 SCA calling protocols

In this section, we focus on the use of an adaptive calling protocol that is aimed at improving the efficiency of contacting sampled households. In this application, a new protocol (timing of the next call) may be invoked for a case after each call. Therefore, in this example, the phase is determined at the case level.

The data come from an RDD telephone survey that is conducted on a monthly basis – the Survey of Consumer Attitudes (SCA). The survey collects approximately 300 RDD interviews per month. The sample is prepared by a vendor that attaches contextual data to the sample file. The ZIP code of each telephone number is estimated using listed numbers from the same 100-bank. Census data for the associated ZIP Code Tabulation Area (ZCTA) are then attached to each telephone number.

The calling protocol described here attempts to learn, over time (measured by calls), which calling times are most likely to lead to contact. As we accrue more data on any particular household, we are learning about

their patterns for being at-home and willingness to answer the telephone. Successfully contacting a household at one time of day or day of the week (i.e. in a particular calling "window") increases our estimate of the chance of success in that window for that household. Unsuccessful attempts to contact a household in another window decrease our estimate of the probability of achieving contact in that window. A similar approach was recommended by Bollapragada and Nair (2010).

Multi-level logistic regression models are fit daily in order to estimate these probabilities of contact using all prior information. These models could be fit more frequently if sampled units were to be called more frequently than daily. The models provide household-specific estimates of the probability of contact for each of four "call windows." The predictor variables in this model are context variables available on the sampling frame.

The household-level estimates of the probability of contact are used to decide which cases have their highest probability of contact (not necessarily the highest of all cases) in the current window. Those cases those cases that have their highest probability of contact in the current window are prioritized in the current window. The sample is re-prioritized at the beginning of each call window.

For operational reasons related to the sample management software, refusal conversion and Spanish language calls were eliminated from the analysis.

At the beginning of the field period, there are no call histories for the current sample. Therefore, we use data from prior months. Specifically we used the call records from the same month in the prior year (in order to capture any seasonal effects in the data) and the month prior to the current. Data from the current month are analyzed daily. The models are re-estimated daily, and the results are updated daily with call records from the first day through the prior day included.

This protocol was implemented experimentally on a random half-sample of cases for each month. The control protocol was the protocol currently employed by field management staff.

The experimental protocol appears to improve contact rates. Table 1 presents the overall contact rates for the experimental and control groups for the "eligible" calls by month and combined. The χ^2 test reported here uses the Rao-Scott approach to account for the clustering of the observations within households. In addition to these differences, we found that approximately 30% of all calls in the experimental group resulted in a "Ring-no-answer" (RNA) while 39% of the calls in the control group produced this result.

Table 1. Calls, Contacts, and Contact Rates by Experimental Group and Month

	Control			Experiment			
Month	**Calls**	**Contacts**	**Contact Rate**	**Calls**	**Contacts**	**Contact Rate**	**Pr> χ^2**
August	4,467	470	0.105	4,238	517	0.122	0.179
September	5,418	507	0.094	5,025	596	0.119	0.016
Combined	9,885	977	0.099	9,263	1,113	0.120	0.008

Although contact rates and efficiency were improved by the experimental method, this method was not applied to refusal conversions or Spanish language cases. The efficiency gains for the calls governed by the experimental method were lost later in the process. This was because refusal conversions in the experimental group required more calls than those cases in the control group. The total calls are nearly equal for the experimental and control group, despite the early efficiency of the experimental protocol.

This result seems to imply that there is some interaction between what happens in the early attempts at contact and the later phase of refusal conversion. It may be that the experimental method leads to contacts at times that are inconvenient for persons in the household. As a result, they are more reluctant to be interviewed at later calls made during a refusal conversion phase. This interaction could be interpreted as evidence that early treatments do impact the effects of later treatments in surveys. As a result, considering the sequence of treatments, as an adaptive treatment regime approach would, should have the potential to improve results.

3. The Use of Paradata in Responsive Design

An essential element of the responsive design approach is the collection of paradata or process data to inform the design decisions and monitor the outcome of the interventions. The development of computer-

assisted surveys, and especially of computer-based case management and call scheduling systems was a key catalyst in the development of responsive designs. Much of the paradata used in the designs we discuss here are from the case management systems, and include call record data, information on interviewer time and travel, and so on. Increasingly such data are being supplemented with interviewer observations at various levels (e.g., neighborhood, sampled household, contact description, respondent, etc.). As such operational data are increasingly being used to manage survey process and from the basis of response propensity models used for both responsive design and for nonresponse adjustment, questions are being raised about the measurement error of such observations (see West, 2010). We are working on ways to evaluate and improve the observational measures used in responsive design, as well as on tools to facilitate the presentation of the paradata generated by projects to facilitate management decisions.

4. Discussion

We have argued in this paper that there is a range of different ways in which responsive design principles can be applied to survey protocols. They share some key elements, however. Regardless of the particular approach used, responsive designs are only successful if the appropriate paradata can be collected, if appropriate survey design protocols can be developed based on the paradata, and if the management of a survey is flexible enough to follow these design protocols. There may thus be circumstances under which the methods described here may not work well, but we believe there are an increasing number of survey settings where these approaches could be adopted.

Our work has demonstrated the utility of these methods, but we still have a long way to goal to fully exploit these ideas in practice. Additional challenge are the development of better analytic tools to analyze complex sequences of events (see, e.g., Kreuter and Kohler, 2009) and better ideas of what sequences are optimal according to a variety of possible criteria (e.g., maximizing response rates, minimizing nonresponse bias, minimizing costs).

A key driving force behind the development of responsive design approaches is the growing recognition that one-size-fits-all protocols, while meeting a requirement of standardization, replicability, and equivalence of stimulus, do not meet the needs of the modern, diverse and rapidly-changing societies we are surveying. Over the past several decades, research has shown the different sample persons react differently to different approaches, particularly with regard to minimizing nonresponse (e.g., Groves and Couper, 1998), and that treating everyone in exactly the same manner may be counter-productive. Thus, the need for the development of responsive design methods is stronger than ever.

References

Bollapragada, S., and Nair, S.K. (2010), "Improving Right Party Contact Rates at Outbound Call Centers." *Production and Operations Management*, 19 (6): 769-779.

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: 743–772.

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

Groves, R.M., and Heeringa, S.G. (2006), "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Nonresponse and Costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169 (3): 439-457.

Groves, R.M., Mosher, W.D., Lepkowski, J., and Kirgis, N.G. (2009), "Planning and Development of the Continuous National Survey of Family Growth." *Vital and Health Statistics*, Series 1, No. 48.

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

Kreuter, F., and Kohler, U. (2009), "Analyzing Contact Sequences in Call Record Data. Potential and Limitations of Sequence Indicators for Nonresponse Adjustments in the European Social Survey." *Journal of Official Statistics*, 25 (2): 203-226.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010), "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173 (2): 389-407.

Laflamme, F., and Karaganis, M. (2010), "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." Paper presented at the Symposium on Recent Advances in the Use of Paradata (Process Data) in Social Survey Research, London, December.

Lepkowski, J.M., Axinn, W., Kirgis, N.G., West, B.T., Mosher, W.D., and Groves, R.M. (2011), "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection." Survey Research Center, University of Michigan, unpublished paper.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M., and Van Hoewyk, J. (2010). "The 2006-2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey." *Vital and Health Statistics*, Series 2, No. 150.

Mohl, C., and Laflamme, F. (2007), "Research and Responsive Design Options for Survey Data Collection at Statistics Canada." *Proceedings of the Joint Statistical Meetings,* Salt Lake City, UT, July 29-August 2.

Murphy, S.A. (2003), "Optimal Dynamic Treatment Regimes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65 (2): 331-355.

Murphy, S.A. (2005), "An Experimental Design for the Development of Adaptive Treatment Strategies." *Statistics in Medicine*, 24 (10): 1455-1481.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010), "Reduction of Nonresponse Bias in Surveys through Case Prioritization." *Survey Research Methods*, 4 (1): 21-29.

Schouten, B., Cobben, F., and Bethlehem, J.G. (2009), "Indicators for the Representativeness of Survey Response." *Survey Methodology,* 35(1): 101-113.

Schouten, B., and Calinescu, M. (2010), "Optimizing Quality of Response Through Adaptive Survey Designs." Den Haag, Netherlands: Statistics Netherlands, Division of Methodology and Quality, unpublished paper.

Sinibaldi, J., Casas-Cordero, C., Eckman, S., McCulloch, S.K., Kreuter, F., and West, B.T. (2010), "Assessment of the Quality of Interviewer Observations." Paper presented at the Symposium on Recent Advances in the Use of Paradata (Process Data) in Social Survey Research, London, December.

Wagner, J. (2008), *Adaptive Survey Design to Reduce Nonresponse Bias*. Ann Arbor, MI: University of Michigan, Ph.D. thesis.

Wagner, J. (2011), "Adaptive Treatment Regimes and Survey Design." Survey Research Center, University of Michigan, unpublished paper.

West, B.T. (2010), "A Practical Technique for Improving the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth." Ann Arbor, MI: Institute for Social Research, NSFG Survey Methodology Working Paper No. 10-013.