Journal of Official Statistics, Vol. 28, No. 4, 2012, pp. 477–499

Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection

James Wagner[1], *Brady T. West*[1], *Nicole Kirgis*[1], *James M. Lepkowski*[1], *William G. Axinn*[1], *and Shonda Kruger Ndiaye*[1]

In many surveys there is a great deal of uncertainty about assumptions regarding key design parameters. This leads to uncertainty about the cost and error structures of the surveys. Responsive survey designs use indicators of potential survey error to determine when design changes should be made on an ongoing basis during data collection. These changes are meant to minimize total survey error. They are made during the field period as updated estimates of proxy indicators for the various sources of error become available. In this article we illustrate responsive design in a large continuous data collection: the 2006–2010 U.S. National Survey of Family Growth. We describe three paradata-guided interventions designed to improve survey quality: case prioritization, "screener week," and sample balance. Our analyses demonstrate that these interventions systematically alter interviewer behavior, creating beneficial effects on both efficiency and proxy measures of the risk of nonresponse bias, such as variation in subgroup response rates.

Key words: Nonresponse; paradata; responsive design; interviewing.

1. Introduction

Survey data collection is filled with uncertainty. This is particularly true for large, face-to-face surveys that rely on interviewers to make most of the decisions about how to achieve contact with (and cooperation from) sampled units. For these surveys, many aspects of the process can only be quantified with probability statements. Commonly used sampling frames (e.g., address lists) may contain many ineligible units. Often, our ability to predict eligibility is weak. Interviewers vary in their ability to find the best times to call on households to maximize contact rates and in their ability to obtain cooperation once contact has been made. Overall, our ability to predict the likelihood of either contact or cooperation is also often weak. Unfortunately, each of these uncertainties interferes with our ability to control the cost, timeliness, and error properties of survey data. This article illustrates the application of a new generation of methodological tools for addressing these uncertainties.

Pre-specified survey designs are not well suited to highly uncertain settings. Any departure from the expectations of the design may lead to a failure to meet some or all of

[1] University of Michigan, USA, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48105, U.S.A. Corresponding author: James Wagner, telephone: 734-647-5600; facsimile: 734-764-8263. Email: jameswag@isr.umich.edu, Brady T. West, Email: bwest@umich.edu, Nicole Kirgis, Email: nkirgis@isr.umich.edu, James M. Lepkowski, Email: jimlep@isr.umich.edu, William G. Axinn, Email: baxinn@isr.umich.edu and Shonda Kruger-Ndiaye, Email: shondak@isr.umich.edu.

the targeted outcomes. These failures frequently include both cost and error failures (Groves 1989), leading to costs that run higher than budgets or errors that are larger than expected. For example, if more effort to complete interviews is required than initially expected, then fewer interviews may be completed and the sampling error of estimates will increase.

Responsive survey designs attempt to address these issues by gathering information about the survey data collection process and using these data to compute indicators that decrease this uncertainty (Groves and Heeringa 2006). These data are used to make decisions about altering design features *during* the survey field work. Groves and Heeringa define five steps for these responsive designs:

1. Pre-identify a set of design features potentially affecting costs and errors of survey statistics;
2. Identify a set of indicators of the cost and error properties of those features;
3. Monitor those indicators in initial phases of data collection;
4. Alter the active features of the survey in subsequent phases based on cost/error tradeoff decision rules; and
5. Combine data from the separate design phases into a single estimator.

These responsive designs rely upon indicators that are built from the available data. Frequently, sampling frames include auxiliary variables that are only weakly predictive of important outcomes of the survey process, including indicators of response and measures on key survey variables collected from respondents. For this reason, researchers have turned to *paradata*, or survey process data (Couper 1998; Couper and Lyberg 2005), as an additional source of auxiliary data. These data may include records of call attempts; interviewer observations about the neighborhood, sampled unit, or sampled person; and timing data from computerized instruments. Responsive designs incorporating paradata to guide design decisions during the field work have the potential to reduce the costs and errors associated with survey data collection. Survey methodology has made advances in the use of paradata (Kreuter et al. 2010; Durrant et al. 2011), but there is very little published research evaluating responsive design tools.

To advance this area of science, this article reviews several responsive design features of a large, face-to-face demographic survey – the 2006–2010 U.S. National Survey of Family Growth (NSFG). The NSFG is sponsored by the National Center for Health Statistics and was conducted by the University of Michigan's Survey Research Center. The responsive design tools described in this article are built upon paradata that have been tailored to the demographic data collected in the NSFG. They are meant to increase our control over the costs, timeliness, and quality of the collected data. Conceptually, responsive designs can be understood from a total survey error perspective, and include monitoring and control of other error sources. We focus on the use of responsive design principles to control the risk of nonresponse bias as a crucial dimension of total survey error.

In most situations, researchers do not have direct information about nonresponse bias. Surveys that do have "gold standard" data or true values available on selected variables for an entire sample are usually performed for methodological – as opposed to substantive – research purposes. Therefore, in order to control the risk of nonresponse bias in a

production environment, proxy indicators of nonresponse bias are needed. For example, the NSFG sample includes multiple subgroups defined by the cross-classification of age, race, ethnicity and gender. A recent review of specialized studies of nonresponse found that the variation in response rates across groups defined by these sorts of demographic variables was not predictive of nonresponse biases (Peytcheva and Groves 2009). In the case of the NSFG, however, these demographic factors are predictive of key survey variables (Martinez et al. 2012). To the extent that these characteristics are predictive of the key statistics measured by the NSFG, large variance in the response rates across these groups is an indicator for potential nonresponse biases (Groves and Heeringa 2006). In another NSFG-specific example, the NSFG asks interviewers to make observations about the sampled persons. These observations are highly correlated with several of the key statistics produced by the survey (Groves et al. 2009; West 2013). These proxy indicators may also be used as indicators for the risk of nonresponse bias. The assumption here is that once we have equalized response rates across subgroups defined by these proxy indicators, the nonresponders and responders within each subgroup will not differ with respect to the survey variables being collected. In other words, we assume that the nonrespondents are "Missing at Random" (Little and Rubin 2002), conditional upon the characteristics used to balance the sample. In this article, we discuss attempts to use such proxy indicators in a responsive design framework to control the risk of nonresponse bias in the NSFG.

In order to make effective use of these proxy indicators, the NSFG design called for centralized direction of data collection effort. We believe that this is a unique feature of the NSFG design. In contrast, most large-scale face-to-face surveys leave the prioritization of effort to the interviewer. The interviewers determine which cases to call and when. Many surveys provide general guidelines to interviewers in this regard. For example, the European Social Survey (ESS) guidelines suggest that a minimum of four calls be placed to each household and that these calls should be spread over different times of day and days of the week, with at least one call in the evening and one on the weekend (Stoop et al. 2010). Others have used prioritization schemes developed prior to fielding the survey in order to increase these sorts of proxy indicators for nonresponse bias (Peytchev et al. 2010). The NSFG is unique in that interviewer behaviors are at times guided by centralized decisions of the managers based on the analysis of paradata. These altered behaviors lead to greater balance on the proxy indicators for nonresponse bias, and we illustrate this result in this article. The special centralized design of the 2006–2010 NSFG gives us a distinctive opportunity to investigate responsive design tools that are intended to alter interviewer behavior in response to incoming paradata during field data collection.

After describing relevant aspects of the NSFG design, we investigate three types of paradata-driven responsive design interventions. The first (Section 3.1) is a set of interventions that was designed to determine our ability to alter interviewer behavior and had specific objectives with relation to the cost and error properties of the data. The second set of interventions (Section 3.2) was aimed at identifying eligible persons earlier in the field period than might have otherwise occurred, thereby procuring data that are informative about the risk of nonresponse bias as quickly as possible in order to enable better control over this error source. The third type of intervention (Section 3.3) uses the variation in subgroup response rates as a proxy indicator for the risk of nonresponse bias. Investigations of these three types of responsive design interventions provide a crucial

advance in the tool set for implementing responsive designs and for using such designs to reduce the uncertainty in survey data collection.

2. NSFG Management Framework

The interventions reported here were developed in the context of a survey management framework that used paradata to guide decision making about survey design features. These responsive design interventions were implemented in the NSFG, which collects data from an ongoing, national, cross-sectional, multistage area probability sample of households in the United States (Groves et al. 2009). In each sampled household, interviewers completed a screening interview by collecting a roster of household members. One person aged 15–44 was selected at random from the age-eligible persons within the household. The interviewer then sought a 60–80 min interview from the selected person. The interview involved the administration of a computer-assisted personal interview (CAPI) questionnaire that contained questions on the respondent's sexual and fertility experiences. More sensitive items (e.g., risk behaviors for HIV) were administered using an audio computer-assisted self-interview (ACASI) application on the interviewer's laptop. A token of appreciation ($40) was paid to respondents upon completion of the main interview.

Each year of data collection for the NSFG consisted of four replicate samples yielding 5,500 completed interviews per year on average. Replicate samples were introduced at the beginning of each quarter. The full data collection period for a year lasted 48 weeks (four 12-week quarters), with four weeks for end-of-year holidays and mid-year training of new interviewers. New interviewers were introduced as part of a rotation of the primary sampling units (PSUs) each year. During any given year, the sample consisted of 33 PSUs and about 38 interviewers across them. The American Community Survey (ACS) uses a similar continuous measurement design to produce timely, frequent, and high-quality data in place of the previous United States Census long form (National Research Council 2007).

Unlike many surveys, the NSFG used a two-phase or double sample process to address the problem of nonresponse. Each 12-week quarter was divided into a 10-week period (Phase 1) and a 2-week period (Phase 2). During Phase 1, interviewers were assigned an average of about 120 sample addresses to screen and interview. At the end of ten weeks, some addresses remained outstanding, that is, they had not yet been finalized as an interview, a refusal, a non-sample case, or some other final disposition. A sample of about one-third of the outstanding addresses was selected and sent back to the interviewers. This sample was selected as a stratified random sample of cases. The strata were defined by eligibility status (eligible or unknown) and tertiles of the estimated probability of response. The sampling rate was chosen based on management experience from Cycle 6 of the NSFG. The sampling rate effectively triples the effort on the selected cases (since the interviewers work a constant 30 per week). This sampling rate allowed us to meet targeted response rates while controlling costs. More information on the second phase sample design is available in the NSFG Series 2 report (Lepkowski et al. 2010). The interviewers then had two weeks at the same weekly effort level to complete interviews with as many of the double sample addresses as possible. The NSFG was also able to provide a higher

token of appreciation ($80 for adult respondents) during Phase 2. Later, during data processing, the Phase 1 and 2 samples were combined in the final survey data set. Weighted response rates were computed to account for the additional interviews obtained from the Phase 2 respondents. Additional details about the design and operations of the Continuous NSFG, including detailed descriptions of paradata collected, can be found in Groves et al. (2009).

The NSFG used a management decision model to guide interventions in a responsive design framework. The model has three input elements that management can manipulate (Effort, Materials, and Quality of Materials), and three broadly defined outcomes (Interviews, Cost, and Sample Characteristics). All inputs and outcomes are monitored through the processing and analysis of paradata. *Effort* refers to survey features such as number of calls, whether in total or within a time frame (e.g., per day); proportion of hours worked during "peak" calling times; number of interviewers working on the study; and hours worked by interviewers. *Materials* include active cases remaining to be screened in the field data collection; cases with identified eligible persons who have yet to be interviewed; or the number of cases not attempted as of a fixed date in the data collection. *Quality of Materials* includes such measures as the proportion of remaining cases that have ever resisted an interview attempt through refusal or other actions indicating a reluctance to participate; the proportion of active cases in a locked building; or the mean of the estimated response propensities for each active case.

Three primary outcomes were of interest to NSFG managers. *Interviews* were measured by such outcomes as the number of interviews completed by day or response rates by day, week, or other time period. *Cost* was measured by hours required to complete an interview or expenditure to travel to a sample location. *Sample characteristics* included measures of how well the set of respondents matched the characteristics of the sample (for example age, sex, race, ethnicity, or interviewer observations about relevant household characteristics), and whether estimates from the observed data converged after a specified number of calls. The overall production model asserts that the number and cost of interviews as well as the characteristics of the sample are a function of the field effort applied and the current state of the active sample (materials and the quality of the materials). This model was applied to the dynamic process of daily data collection.

The elements in the production model were monitored through a "dashboard" consisting of graphs portraying the status of various measures for each of these elements (see Groves et al. 2009). The graphs were updated daily throughout the data collection period. The dashboard served as a central feature in the management process, allowing for monitoring of all elements in the model and guiding management decisions about how and when to intervene.

3. Three Paradata-Driven Interventions

The 2006–2010 NSFG implemented three different types of management interventions in the responsive design framework: *case prioritization*, *screener week*, and *sample balance*. Each of the three types of interventions had different objectives. The *case prioritization* intervention was aimed at checking whether the central office could influence field outcomes by requesting that particular cases be prioritized by the interviewers. If this

prioritization proved to be successful, then the second objective was to determine what impact these case prioritizations could have on the composition of the final set of respondents. *Screener week* sought to shift the emphasis of field work in such a way that eligible persons (and proxy indicators of nonresponse bias for those persons) would be identified as early as possible. Since the screening interview also generates valuable data about the demographic characteristics of sampled persons, screener week improved our ability to balance the sample. The "*sample balance"* intervention sought to minimize the risk of nonresponse bias by endeavoring to have the set of respondents match the characteristics of the original sample (including nonresponders) along key dimensions, such as race, ethnicity, sex, and age. We describe each of these interventions in detail and provide examples of their implementation in the following subsections.

3.1. Case Prioritization: Paradata-Guided Randomized Experiments

The idea of embedding randomized experiments in an ongoing survey data collection is not new. Possible reductions in survey errors from adaptively embedding randomized experiments in survey designs have been discussed previously by Fienberg and Tanur (1988, 1989). The first set of interventions that we describe here involved assigning a random subset of active cases with specific characteristics to receive higher priority from the interviewers. NSFG managers targeted these cases for intervention in response to trends in selected elements of the production model that indicated possibly increased risks of survey errors. This type of intervention was replicated 16 times on different quarterly samples.

The case prioritization interventions involved late-quarter targeting of specific types of sampled housing units or persons (if already screened) to increase the number of calls to these specific groups. The first objective of these experiments was to determine whether interviewers would respond to a request to prioritize particular cases. While one can assume that interviewers will do what is requested of them, we knew of no research examining the outcomes of such requests in a field data collection. It was hoped that if the calls were increased, then response rates for the targeted cases would rise, relative to those of other cases. In this section, we focus our analysis on determining whether these types of interventions can have an impact on effort and, subsequently, on response rates for the targeted subgroups. If these interventions are successful, they may be an important tool in reducing interviewer variance and controlling the composition of the set of respondents. In subsequent sections, we will consider how these interventions might be used to improve survey outcomes relative to the risk of nonresponse bias.

Each of the experiments also had a secondary objective related to reduction of survey errors. Table 1 lists all 16 of the randomized experiments and describes the secondary objectives for targeting each of the specified subgroups. In some cases, the objective was to improve overall response rates. In other cases, the objective was to evaluate the utility of data available on the sampling frame. In still other cases, the objective was to bring the distribution of the characteristics of the respondents closer to the distribution of the characteristics of the original sample.

All 16 of these interventions were randomized experiments in which one half of the target cases was assigned to the intervention and one half remained as a control group.

Table 1. 16 randomized interventions, 2006–2010 Continuous NSFG

Intervention Type[a]	Description	Objective	Length (Days)	Sample size	
				Intervention	Control
EXT1	Active screener addresses matched with Experian data indicating household eligibility (at least one person age 15–44 in household)	Evaluate the utility of commercially available data. Evaluate whether prioritizing likely eligible persons leads to better sample balance.	11	759	755
EXT2	Active screener addresses matched with Experian data indicating household not eligible (no person age 15–44 in household)		11	637	624
EXT3	Active screener addresses with no Experian match (indeterminate household eligibility)		11	430	434
INT1	Active screener addresses with high predicted probability of eligibility (based on NSFG paradata)	Determine whether prioritizing likely eligible persons leads to better sample balance	13	204	165
INT2	Active main addresses with high predicted probability of response (based on NSFG paradata), no children, and high predicted probability of eligibility (based on NSFG paradata)		14	115	109[b]
INT3	Active screener addresses with high predicted probability of response (based on NSFG paradata), no children, and high predicted probability of eligibility (based on NSFG paradata)		14	146	146
INT4	Active main addresses with high base weights and large or medium predicted probabilities of response (based on NSFG paradata)	Determine whether it is possible to improve response rates by prioritizing cases with relatively high weights.	8	100	88[b]
INT5	Active screener addresses with high base weights and larger or medium predicted probabilities of response (based on NSFG paradata)		8	133	133

Table 1. Continued

Intervention Type[a]	Description	Objective	Length (Days)	Sample size	
				Intervention	Control
DS1	Active main addresses in double sample with large or medium base weights	Determine whether it is possible to prioritize cases during the second phase.	11	46	46
DS2	Active screener addresses in double sample with large or medium base weights		11	26	25
DS3	Active main addresses in double sample with large base weights		10	28	28
DS4	Active screener addresses in double sample with large base weights		10	20	20
SB1	Active main addresses with no children under 15 years of age on household roster	Determine whether it is possible to improve sample balance through prioritization.	15	232	188[b]
SB2	Active main addresses with no children under 15 years of age by interviewer observation		8	167	315
SB3	Active main addresses with older (age 20–44) non-Black and non-Hispanic males		13	103	85
SB4	Active main addresses with older (age 20–44) Hispanic males		11	69	62

[a] EXT = subgroup defined by external data; INT = subgroup defined by internal paradata used to estimate predicted probabilities of response; DS = Phase 2 subgroup defined by stratification and weight paradata; SB = sample balance subgroup.

[b] Subset of control cases that were also part of simultaneous non-randomized sample balance intervention (see Section 3.3) deleted from comparison.

The prioritized cases were "flagged" in the interviewers' view of the sample management system. Figure 1 shows how these flags appeared to the interviewer. Interviewers were asked to prioritize the "flagged" cases and apply more effort to these cases. Instructions about the interventions were communicated to interviewers in a weekly telephone call (or "team meeting") and by email. Subsequent analyses examined effort on intervention and control addresses to determine if a null hypothesis of no difference in number of calls or response rates between the two groups of cases could be rejected.

Since these interventions occurred later in some quarters and also targeted different types of cases (given secondary objectives), sample sizes in intervention and control groups were sometimes small. There was limited power to detect even modest differences in response rates in many of these randomized experiments. Rather than focus on individual experiments that rejected the null hypothesis of no difference, we summarize findings across experiments. Across the 16 interventions, the null hypotheses for the number of calls or response rates might be expected to be rejected (using a 5% level of significance) in less than one intervention by chance alone.

For each randomized intervention, there are two questions posed:

1. Do interviewers do what we ask of them (that is, do they increase the number of calls to high priority cases)?
2. Does the intervention increase the response rate among the target high priority cases?

Table 1 summarizes the characteristics of the 16 randomized interventions. Intervention periods ranged from eight to 15 days, and sample sizes from 20 to 759 per intervention or control group. Subgroups subject to intervention varied on a number of characteristics; specifically, we distinguish between four types of case prioritization interventions. Three interventions were primarily based on external (EXT) commercial data purchased to determine whether household eligibility could be reliably predicted for addresses from the external source before the screening interview was completed. Five were based on internal

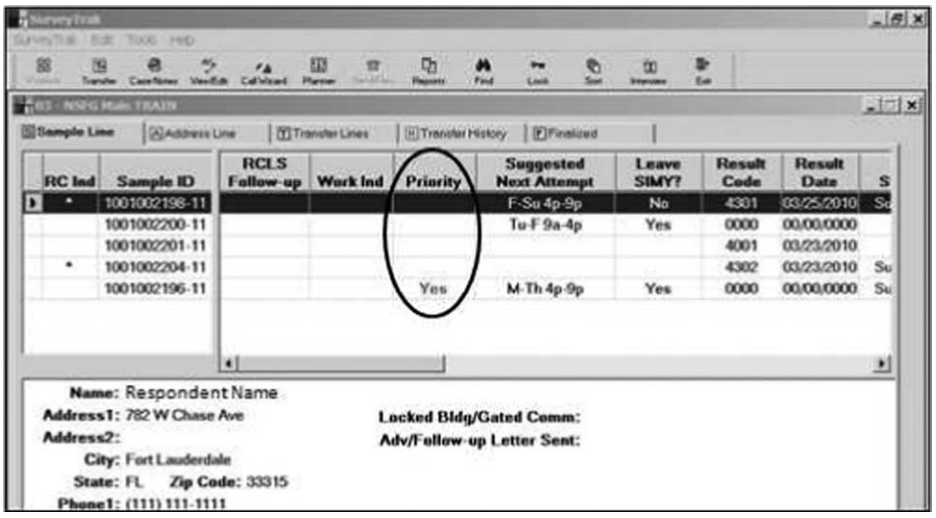

Fig. 1. Screen shot of an active sample line "flagged" as high priority in the sample management system, 2006–2010 Continuous National Survey of Family Growth

(INT) NSFG paradata used to predict either propensity to respond on a given day of the quarter or eligibility status. These predictions were based on logistic regression models fitted to addresses or households for which response status was known (responded or not) or household eligibility was known (eligible or not). Predictors included contact information recorded by interviewers at each household contact, interviewer observations about sample block or sample address characteristics, or interviewer judgments about individuals living in the household. Interventions of this type targeted addresses with high or medium predicted probabilities of response, addresses with high base weights in an effort to improve response rates, or high predicted probabilities that an address had one or more eligible persons residing there. These interventions helped to increase the yield of the sample, which was an important objective. Other interventions and design features were aimed at minimizing nonresponse bias. Four randomized interventions (INT2, INT3, INT4, and INT5) involved combinations of sample selection criteria. The subgroups for these four interventions were all based, though, on internal models driven by paradata, and are thus classified as the internal type.

Four interventions were conducted on the Phase 2 or double sample (DS) selected addresses. Cases with a high selection weight or a high probability of response were prioritized during the second phase.

Four additional interventions were randomized experiments to assess whether sample balance (SB) on key subgroups could be restored by intervention on high priority addresses. In addition to a sample balance intervention on Hispanic males ages 20–44 years, interventions were conducted on main addresses judged by interviewers to have no children under 15, with no children on the household roster from the screening interview, and non-Black and non-Hispanic males ages 20–44 years, groups that were observed to have lower response rates in particular quarters. The interviewer judgment about the presence of young children was one of several interviewer observations collected to provide NSFG managers with auxiliary information enabling comparisons of responding and nonresponding households. Groves et al. (2009) provide a more detailed description of these interviewer observations, and West (2013) examines the accuracy of the observations and shows that the observations are correlated with both response propensity and several key variables collected in the NSFG interview.

We consider first whether flagging high priority addresses changed interviewer behavior. Figure 2 presents bar charts of the mean cumulative calls per address for both the intervention and control groups of addresses at the conclusion of each of the 16 interventions. Significant differences in mean cumulative calls at the $P < 0.05$ level based on independent samples t-tests are highlighted. The means in Figure 2 consistently show the intervention addresses receiving more calls than the control addresses. Approximately half (seven) of the experiments resulted in statistically significant two-sample hypothesis test results.

The interventions clearly had a consistent impact on interviewer calling efforts. The next question was whether the increased effort led to corresponding increases in response rates for intervention relative to control addresses. Figure 3 presents comparisons of final response rates (according to the AAPOR RR1 definition) at the end of each intervention for the intervention and control groups. Significant differences in final response rates with $P < 0.05$ for a χ^2 test of independence (where distributions on a binary indicator of

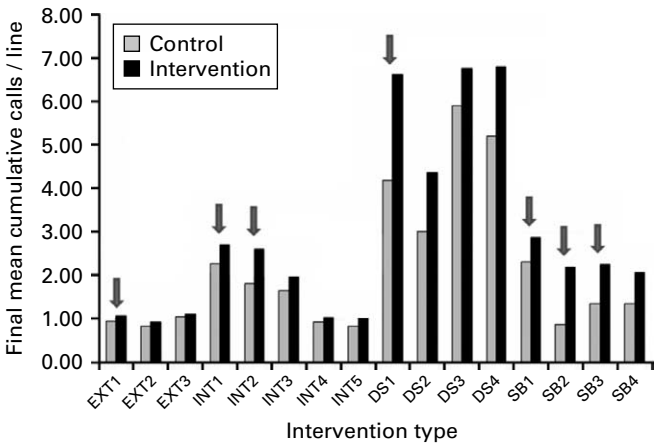

Fig. 2. Mean cumulative calls at the end of an intervention (arrows indicate significance at $\alpha = 0.05$ for independent samples t-tests) for intervention and control groups in 16 randomized trials, 2006–2010 Continuous National Survey of Family Growth

response were compared between the intervention and control groups of addresses) were found for only two of the 16 interventions (screener addresses predicted to have high eligibility and main addresses with no children under age 15 years in the household from the roster data). Response rates were generally found to be higher in the intervention groups, but there were four experiments with slightly higher response rates in the control group. Thus, there is some evidence that increased calling efforts tended to result in increases in response rates, although statistically significant increases occurred in only two of the interventions. Across all 16 interventions, there was a weak positive association between increased calling effort and increased response rates.

Two of the four interventions with higher response rates for control cases are interventions that were implemented during the double sample (DS) period. A third double sample intervention (DS3) resulted in equal final response rates in the two groups. During the second-phase period of the NSFG's double sampling operation, more attention was

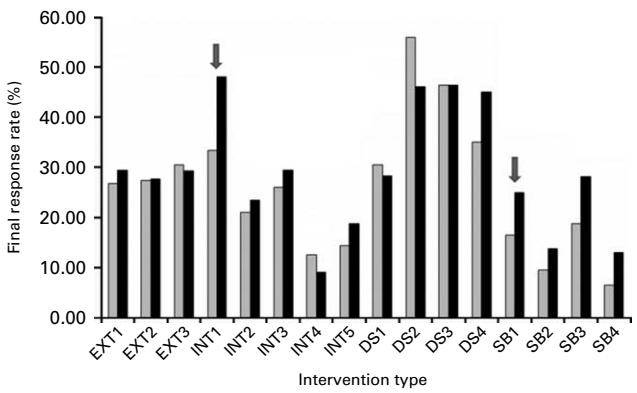

Fig. 3. Response rates (arrows indicate significant differences at $\alpha = 0.05$ in χ^2 tests of independence) for intervention and control groups in 16 randomized interventions, 2006–2010 Continuous National Survey of Family Growth

being paid to all active addresses. If these three double sample interventions (DS1, DS3, and DS4) are removed, there is much clearer evidence of a positive association between increase in effort and increase in response rates. These results for the double sample interventions indicate that intervening during an already intensive effort in a double sample period will not necessarily increase response rates. Because interviewers have a greatly reduced workload (approximately 1/3 of their assigned cases that have not been finalized are retained during the second phase), it may be that all cases are already being called more frequently than during Phase 1, and the additional calls on prioritized cases do not lead to additional contacts and interviews.

As a final evaluation of the randomized interventions, we present a more detailed analysis of the effectiveness of one of the 16 "internal" interventions, INT5. During each of the interventions, interviewers established appointments with active cases, and interviews were then completed after the end of the intervention period (which was chosen arbitrarily by NSFG managers). The question of interest is whether higher effort levels continued for intervention addresses after the end of the intervention period, and whether there is an increase in completed interviews relative to control cases. We suspected that higher calling rates would continue for intervention cases, because more calls should yield more contact with household members, more appointments, and interviewer visit patterns guided by more information about when household members are more likely to be at home.

We chose INT5 for this analysis for three reasons. First, this intervention had a balanced design, with a relatively large sample of 133 addresses in each arm of the experiment. Second, anecdotal reports from interviewers indicated that this intervention, although it did not lead to a significant difference in response rates, did lead to an increase in the number of appointments for the group receiving the intervention. We hypothesized that this appointment-setting work may have led to increased response rates after the intervention concluded. Third, because the experimental group did not receive higher numbers of calls or have higher response rates in this intervention, we wanted to see if this was an artifact of our arbitrarily ending the analysis of the treatment effect with the end of the prioritization.

A total of 266 active addresses without a screening interview, with larger base sampling weights (the largest tercile of the distribution of weights) and higher estimated response propensities predicted by the paradata (upper one-third of all active addresses), were selected for the INT5 intervention. One half (133) of these addresses were assigned to the intervention group, and the rest were assigned to the control group (where they received standard effort from the interviewers). There was a clear long-run benefit of the intervention on calling behavior. After the "end date" of this intervention (29 August 2007), at which time there was a slightly higher number of calls in the intervention group, the gap between the groups continued to increase, eventually leading to roughly 0.5 calls per address more on average than control cases. This result indicates that intervention addresses did receive higher calling effort during the intervention periods, and that the higher call effort continued with more calls being placed to intervention cases until the end of the quarter.

When we examined the cumulative response rate for each group in INT5, the largest gap in response rates between the two groups occurred when the intervention was originally

stopped on 29 August 2007. After this date, the gap between the two groups remained similar, with response rates increasing at the same rate in both groups, and the intervention group continuing to have a higher response rate until the end of the quarter. This constant gap may have been a function of the continued increase in calls to these cases after the intervention was stopped. In sum, these case prioritization experiments demonstrated that we have the ability to alter field data collection efforts from a central office. This capability should aid the reduction of interviewer variability while improving the balance of selected characteristics of the set of interviewed cases relative to the full sample. The latter may result from identifying eligible cases more quickly or by improving sample balance (see Section 3.3). Finally, we note the importance of continued experimentation with these techniques for discovering unintended consequences. For example, in the case of the interventions applied to the second phase samples, we found that the interventions were not effective, as interviewers were essentially prioritizing all of their remaining cases.

3.2. Screener Week: Shifting Effort to Incomplete Screener Addresses

From our experience with the implementation of NSFG Cycle 6 (Groves et al. 2005), the management team had observed that interviewers varied in how they scheduled work. Interviewers typically scheduled main interviews even when assignments included a large proportion of incomplete screener addresses. In the last weeks of Cycle 6, there remained data collection screener addresses with a limited number of calls and no completed screener interview. These indicators pointed to an interviewer preference for completing main interviews over screening households for eligible persons.

This apparent preference created two issues for the continuous design employed in the 2006–2010 NSFG. First, because main interviews could not be completed until screener interviews were completed, interviewers had limited time to complete main interviews with cases that were screened later in the process. This hampered our ability to attain high response rates in a study with a relatively short field period each quarter (12 weeks). Second, the screening interview generates important auxiliary data for further responsive design decisions. Information about the age, race, ethnicity, and sex of the selected person as well as an interviewer judgment about whether the selected person is in a sexually active relationship with a person of the opposite sex are used in subsequent interventions to improve the balance of the interviewed cases relative to the full sample along these dimensions (see Section 3.3 for a full description).

In the Continuous NSFG, project management sought to divert interviewer effort to screener addresses at an earlier point in the data collection. An intervention strategy was sought that would increase effort to call at any remaining previously not-contacted screener sample addresses, resolve access impediment issues that blocked contact attempts, and ultimately produce more screener interviews (regardless of whether age-eligible persons were present).

The field management strategy in week 5 of the first quarter was to instruct interviewers to keep all current firm main interview appointments made previously, to set main interview appointments at screener interview completion with selected eligible respondents during week 5 if a later time was not available, and to schedule main interview appointments with sample persons not present at the completion of the screener

interview *after* week 5. Field management then emphasized the importance of making calls on screener addresses during this week. The instructions were given in regularly scheduled telephone conference calls and in email correspondence.

Field management monitored screener calling and interview progress by using daily electronically-submitted interviewer call records before, during, and after week 5. There was an increase in screener calls and an increase in the ratio of screener to main calls during week 5 of year 1, quarter 1 (Y1Q1). Field management instituted screener week in week 5 (days 29 to 35) in Y1Q1, and in each subsequent quarter until the conclusion of data collection in 2010. There was one exception – in Y2Q2, screener week was implemented in week 4.

Graphs of daily and seven-day moving averages of completed screener and main interviews, such as those shown in Figure 4, were examined throughout each quarter. The upper black lines in Figure 4 track the daily and seven-day moving average number of screener interviews. In later quarters after the first, field management compared current quarter results to a previous quarter, to a previous quarter one year earlier, or a yearly average across quarters from a previous year. The lower grey lines similarly track the corresponding daily and seven-day moving average number of main interviews.

The number of screener interviews in the first three weeks of a quarter (Figure 4 presents data from Y4Q1) was between 80 and 100. The count gradually declined to less than 20 per day at the end of Phase 1 each quarter. There were relatively steady main interview counts per day of around 20 after the first three or four weeks of each quarter. The upper black lines in Figure 4 show (where vertical lines separate the weeks) a rise in the number of screener interviews in week 5, and little change in the number of main interviews.

The number of calls to active screeners and the screener to main call ratio increased in each quarter during screener week. While the size and consistency of the increases in each

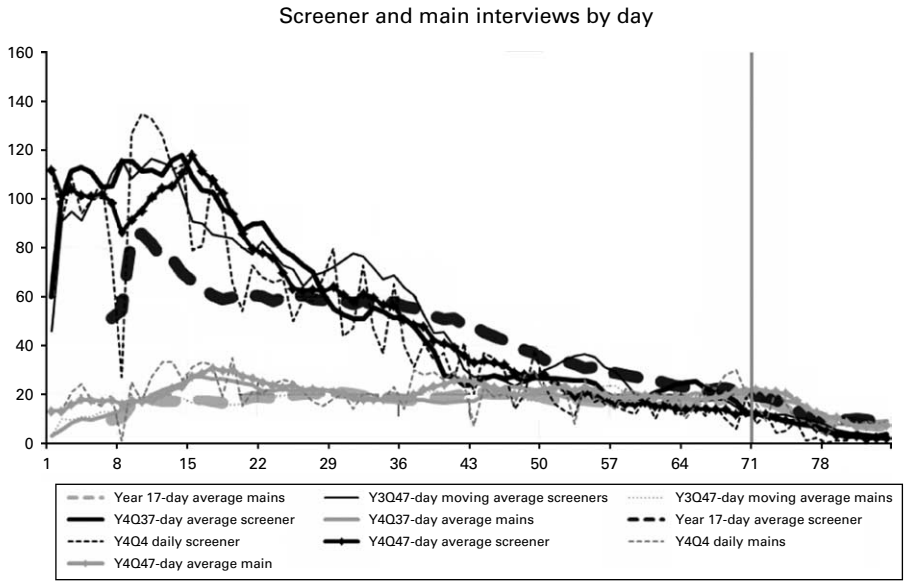

Fig. 4. Number of and seven-day moving average screener and main interviews by day for year 1 and data collection quarters Y3Q2, Y4Q1, and Y4Q2, 2006–2010 Continuous National Survey of Family Growth

screener week suggested a change in interviewer behavior, there was no experimental validation of this result. In an effort to evaluate further whether "screener week" had an impact on the volume of screener calls, two statistical models were fit to the paradata, and hypothesis tests about model parameters were conducted.

In the first model, the dependent variable was the daily number of screener calls for weeks 3 through 7 (days 15 to 49) in each of the 16 quarters – that is, the days before (days 15–28), during (days 29–35), and after (days 36–49) screener week. This included two weeks before screener week, the screener week, and two weeks after the screener week. There was one exception. InY2Q2, screener week was initiated in week 4, and for that quarter, weeks 2 through 6 are included in the analysis. There were 559 days across 16 quarters in the analysis (Y2Q2 only included 34 days, because the screener week intervention lasted only six days in that quarter).

The number of screener calls was regressed on the day number, an indicator of whether the day was in screener week, and the quarter number. Interactions among day, the screener week indicator, and the quarter were also included in the model. A three-way screener week by day by quarter interaction suggests a complex interviewer response in which screener week call levels were irregular during screener week and across quarters. Two-way screener week by quarter and day by quarter interactions would indicate whether screener call levels differ across quarters and across days within quarters. A two-way day by screener week interaction indicates whether there was a different number of screener calls across days in screener week. The screener week by quarter interaction was expected to be statistically significant, because there was observed variation in the number of screener calls during screener week across quarters. The day by screener week interaction was also expected to be significant, because in each screener week there was a rising number of screener calls from the beginning to the end of the week. Table 2 and Figure 5a and 5b summarize the model and the results of tests of null hypotheses about model parameters for the number of screener calls. Figure 5a presents predicted screener call levels by day for each quarter obtained from a reduced model that used only the statistically significant coefficients to compute the predicted values. That is, Figure 5a presents a "smoothed" image of the daily screener call levels as estimated from the reduced model.

Table 2. Analysis of factors affecting the number of calls per day made before, during, and after screener week, 2006–2010 Continuous National Survey of Family Growth

Factor	F-Statistic	Numerator DF	Denominator DF	*P*-value
Day of field period	159.20	1	525	<0.0001
Screener week	7.44	1	525	0.0066
Quarter	2.13	15	525	0.0079
Screener week × day	12.04	1	525	0.0006
Screener week × quarter	NS	–	–	–
Day × quarter	1.73	15	525	0.0425
Screener week × day × quarter	NS	–	–	–

NS = Not significant. Model $R^2 = 0.345$. The F-statistics test the hypothesis that the factor coefficients are different from zero in the presence of the other factors in the model.

There were statistically significant interactions between day and screener week and day and quarter, as expected. After removing the parameters associated with the other factors which could not be distinguished from zero, the remaining five factors explained 34.5% of the variance in daily screener calls. The significant interactions indicate that the number of

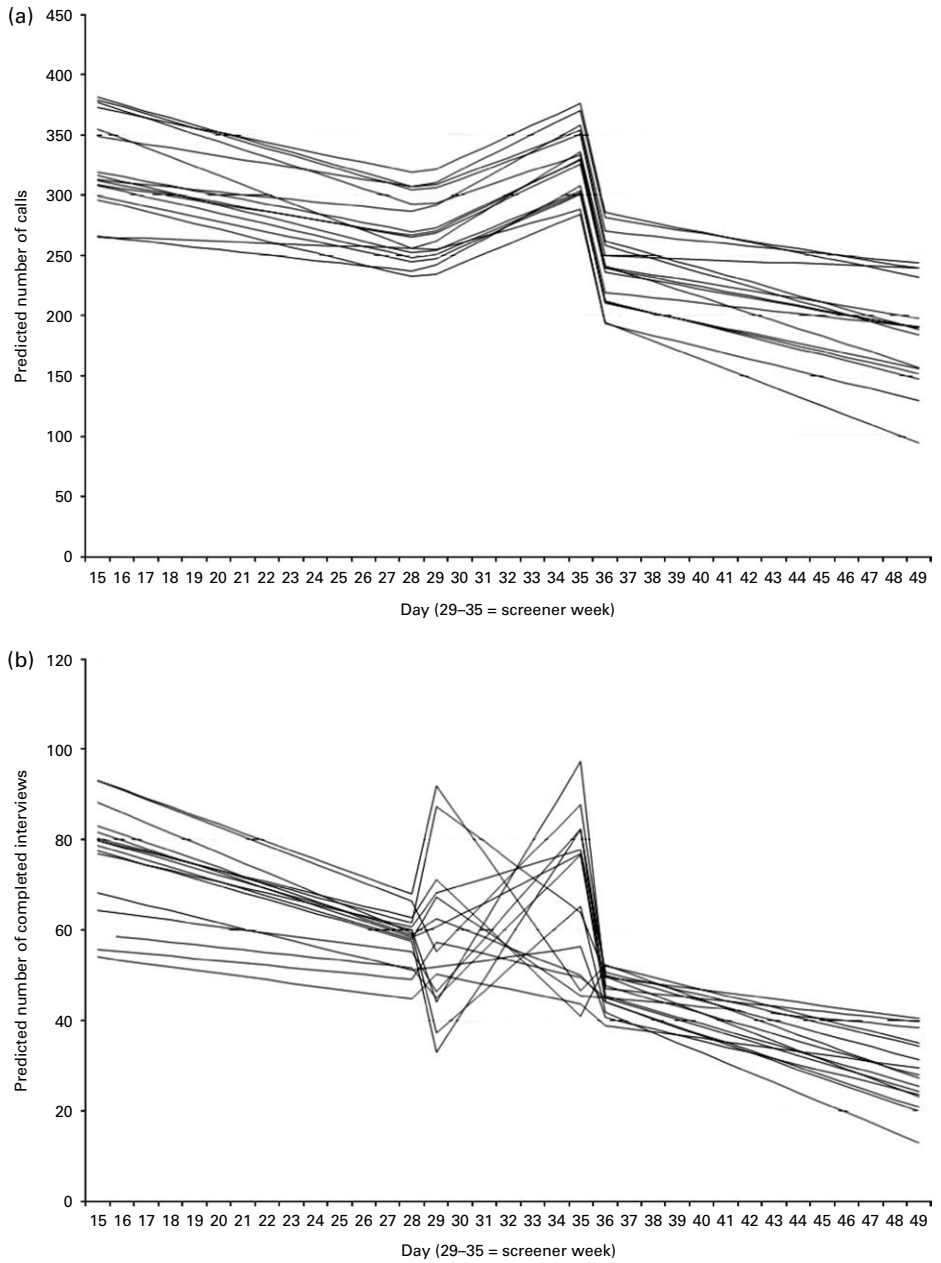

Fig. 5. Number of daily calls (Figure 5a) and number of daily completed screeners (Figure 5b) predicted under models with only statistically significant coefficients for weeks 3–7 (screener week days 29–35) for the 16 quarters of data collection, 2006–2010 Continuous National Survey of Family Growth

screener calls each day changes during screener week, and that the number of screener calls on average was different across quarters.

These findings are confirmed in Figure 5a. The figure shows that the predicted number of calls declines, except during screener week (days 29–35). There is an increasing predicted number of calls made per day across screener week. This increase occurred consistently across all quarters. The increasing number of screener calls during week 5 reverses a negative trend in screener calls per day before, and after, screener week. There is also evidence of variation in calling behavior across quarters, with some quarters having more calls than others over the five-week period. The effects of screener week and changes in the screener calling trends during screener week were, however, consistent across the 16 quarters.

The second model had identical predictors, but the dependent variable was changed to the daily number of completed screener interviews. Table 3 summarizes the test statistics for the second model. There is a significant three-way interaction between day, screener week indicator, and quarter, suggesting that changes in the number of interviews occurred across day within screener week, and that the day by screener week trend was not the same across quarters. The consistent increases in the number of screener calls across days in screener week were not repeated across quarters for the number of completed screener interviews.

Figure 5b shows "smoothed" predicted counts of screener interviews per day based on the fitted regression model including the three-way interaction. Figure 5b is a striking contrast to Figure 5a. The general trend of decreasing numbers of screener interviews before, and again after, screener week, is interrupted by a complex rise and fall of completed screeners during screener week in each quarter. The expected rate of completed screener interviews per day did not consistently increase during screener week across the 16 quarters. Consistent increases in screener calls across days of screener week did not produce consistently increasing numbers of completed screener interviews. There were initial decreases in screener interviews followed by increases during one half of the screener weeks, while in the other weeks there were sharp to modest increases early in screener week followed by decreases. Across all 16 quarters, during screener week there was an average effect of increased numbers of completed screeners, but the rates of

Table 3. Analysis of factors affecting the number of completed screener interviews per day before, during, and after screener week, 2006–2010 Continuous National Survey of Family Growth

Factor	F-Statistic	Numerator DF	Denominator DF	*P*-value
Day of field period	309.26	1	495	<0.0001
Screener week	4.82	1	495	0.0286
Quarter	3.57	15	495	<0.0001
Screener week × day	7.94	1	495	0.0050
Screener week × quarter	1.78	15	495	0.0351
Day × quarter	2.81	15	495	0.0003
Screener week × day × quarter	1.86	15	495	0.0245

Model $R^2 = 0.463$. The F-statistics test the hypothesis that the factor coefficents are different from zero in the presence of the other factors in the model.

completed screeners were not consistent across quarters. The reasons for this inconsistent effect may have to do with changes in the interviewing staff, variation between samples, or possible seasonal effects.

Although not experimentally implemented, we would argue that the emphasis on early screening helped to improve response rates. Logically, the screening interview needs to be completed before the main interview. The sooner that this task is completed, the more opportunity there is to complete the main interview. The screening week intervention was implemented during days 29 to 35 each quarter. Empirically, about 93% of the cases are interviewed within 49 days after being screened as eligible. Only 89.5% of cases are interviewed within 42 days after being screened as eligible. Identifying eligible persons as early as possible will therefore increase the likelihood of completing an interview. It was our experience from a prior Cycle of the NSFG (and other large-scale surveys using screening) that interviewers prefer to complete main interviews. They may delay screening, thus decreasing the time available to complete interviews with newly identified eligible persons. In addition, the rapid screening of households enabled us to use paradata from the household screening to create a proxy indicator for nonresponse bias that guided the types of interventions described in the next section.

3.3. *Sample Balance: Targeting Subgroups in Order to Reduce Variation in Subgroup Response Rates*

The third type of intervention, *sample balance*, was designed to reduce the risk of nonresponse bias. Since the survey variables for nonresponders are not known, a proxy indicator for nonresponse bias was needed. The proxy indicator chosen for this purpose was variation in subgroup response rates. NSFG management monitored the response rates of 12 individual subgroups and the coefficient of variation of these subgroup response rates on a daily basis. The variation in subgroup response rates reflects how closely the set of interviewed cases matches the sampled cases on the key characteristics used to define the subgroups – in this case, age, race, ethnicity and sex. In this sense, this indicator is very similar to the R-Indicator (Schouten et al. 2009). This type of intervention sought to bring the composition of the set of interviewed cases closer to the composition of the full sample by prioritizing cases from subgroups that were responding at lower rates. The key characteristics used for this purpose were age, race-ethnicity, sex, and presence of children under the age of 15 in the household (each collected during the screener interview), as well as presence of children under the age of 15 in the household (from interviewer observation). Of course, we cannot be certain that this approach actually reduces bias.

The distribution of the daily response rates by subgroups varied some over the years and quarters. This variation could be due to changes in the composition of the samples each quarter and changes in the interviewing staff each year. In many quarters, one subgroup showed lower numbers of interviews and lower response rates: Hispanic males ages 20–44. Figure 6 is an actual dashboard display monitored by NSFG management. It shows response rates for days 1 to 70 (the first 10 weeks) of Y4Q2. The denominator for each subgroup changes daily, as new cases are identified through the screening process. For instance, on the first day of Y4Q2, one Hispanic male 15–19 years of age was identified and interviewed. Therefore, the response rate for this subgroup is 100% on day 1, and goes

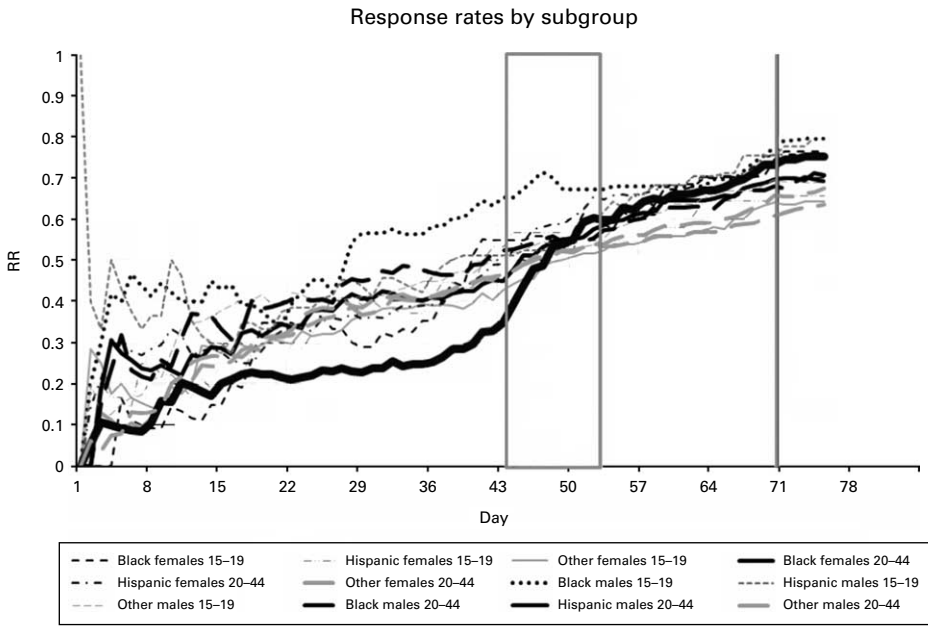

Fig. 6. Daily cumulative response rates for twelve subgroups defined by gender, race-ethnicity, and age, with the intervention period for Hispanic males 20–44 years of age in days 44–55 highlighted, 2006–2010 Continuous National Survey of Family Growth

down the next day as new cases are identified. As sample sizes increase, differences in response rate stabilize. In the case of Y4Q2, through week 6, Hispanic males 20–44 years of age had lower response rates.

In response to observed trends in a given quarter, field management developed an intervention to restore balance in the composition of the interviewed cases. At different points in each quarter, all outstanding addresses known to contain selected persons in a low response rate subgroup were identified. At the start of a sample balance field intervention, field management marked these addresses as high priority in the central sample management system. During nightly uploads of data, interviewers also downloaded the updated priority data from the sample management system.

In several quarters when sample balance interventions were conducted, the high priority designation was randomly assigned to one half the target subgroup addresses. The results of these randomized experiments are discussed in Section 3.1. Here only the non-randomized interventions are examined.

The high priority cases in randomized and non-randomized sample balance interventions were marked in laptop address lists with a high priority indicator (see Figure 1). Field management subsequently monitored daily response rates and numbers of interviews to observe if the priority assignment yielded the desired effect. Since some Hispanic males may require bilingual interviewers, field managers also assigned traveling interviewers with bilingual capabilities to segments containing addresses in this target subgroup.

Figure 6 also shows the effect of the "Hispanic male 20–44 years of age" intervention on response rates. The intervention began on day 44 of this quarter, with high priority case flags assigned in the sample management system to all addresses with a selected person from the targeted subgroup. There is a clear increase in response rates for this subgroup over the next week. This Y4Q2 intervention yielded, at the end of ten weeks of data collection, a response rate for Hispanic males 20–44 years of age that was similar to that for the other eleven subgroups. The intervention, therefore, had the beneficial effect of decreasing variation in response rates among these six subgroups. The variation in subgroup response rates is a process indicator monitored by NSFG managers to assess balance in the data set. This beneficial effect also translates to a reduction in the variation in nonresponse adjustment weights, and reduced sampling variance of weighted estimates.

The age-race-ethnicity-sex subgroups were not the only ones monitored and for which sample balance interventions, randomized and non-randomized, were attempted. For example, during Y4Q3, field managers noticed, while reviewing the dashboard, lagging response rates among sample households with children under the age of 15 (identified with screening data). Since the presence of young children is correlated with many of the key outcome statistics produced by the NSFG (West 2013), this indicator was also used as a proxy indicator for nonresponse bias. Establishing balance on this proxy indicator is meant to reduce the risk of nonresponse bias and mitigate the inflation of variance estimates due to the variability of nonresponse adjustment weights. Just as for Hispanic males ages 20–44 years, field management "flagged" high priority addresses for subgroups such as households without children less than 15 years of age (from the screening interview) to receive increased effort from the interviewers. Field management also sent email reminders advising interviewers that high priority addresses required extra effort on their part.

In sum, the interviewers followed centralized directions of how to prioritize their sample. This centralized prioritization can be used to improve the composition of the final set of respondents by increasing the response rates of groups that are "underrepresented" by the response process.

4. Summary and Conclusions

This article presents case studies of responsive design interventions generated from active monitoring of paradata. Three types of paradata-driven management interventions were examined: one applied to subgroups identified through a variety of internal and external paradata (*case prioritization*), one applied to all interviewers on a very broad level (*screener week*), and one applied to a selection of addresses with known key subgroup members (*sample balance*). Each illustrates important dimensions of the tools of responsive design, including the ability to use paradata to systematically alter interviewer behaviors during field work and the consequences of those behavioral changes for the nature of the survey data collected.

For case prioritization interventions, we found that interviewers will respond to centralized requests that set priorities on cases from key subgroups that are underresponding. The first analysis examined 16 different randomized interventions applied to addresses selected from groups defined by a variety of paradata. Interviewers followed intervention guidelines, making more calls on the experimental intervention

addresses than on the control addresses. The intervention addresses also tended to achieve larger increases in response rates than the control cases during the intervention period.

The second intervention successfully increased the effort on active screener cases. This led to earlier identification of eligible sample persons and the collection of key information used later to assess sample balance. To model the impact of the second intervention, the week of the screener intervention was contrasted to two prior weeks and two following weeks. Rates of interviewer calling significantly increased in a consistent manner across quarters during the screener week, indicating that the intervention did indeed influence interviewer behavior. The increased rates of calling, however, did not consistently lead to increased numbers of interviews during the same period. Once again, responsive design tools can be effective at altering interviewer behavior as desired, but tests across a broader range of interventions will be required to determine the most effective tools.

The third set of interventions was based on a proxy indicator for the risk of nonresponse bias – variation in subgroup response rates. Intervention on cases from "under-represented" subgroups not only affected interviewer behaviors, but in this important case also increased the subgroup response rate and reduced the variation of the response rates among key subgroups. This type of targeted intervention was successful at improving the balance of cases interviewed across subgroups defined by age, race-ethnicity, sex and other characteristics important in predicting survey outcome variables.

The overall conclusion that can be drawn from these findings is that interviewers were attentive to and accepted the centralized intervention strategies in the NSFG, despite not being told the reason for increased effort on certain addresses in most interventions. Interviewers were notified electronically and via conference calls that certain addresses would be high priority and to place emphasis on these lines as they planned their work.

A *sine qua non* of responsive design is, therefore, the ability of the central office staff to instruct the field interviewers to change their focus from one task to another. The three case studies in this article show that real-time interventions can lead to changes in key indicators of survey quality. All interventions were successful at altering interviewer behaviors, but not all interventions were successful at altering survey outcomes. Continuous examination of the practice of responsive design and investigation across a broader set of interventions is necessary to identify the types of interventions that further improve survey costs and reduce survey errors.

The techniques demonstrated in this article can be used by survey organizations to control progress toward key quality indicators (Kirgis and Lepkowski forthcoming). These techniques require the development of reporting mechanisms that allow managers to review progress on a frequent basis. Managers may decide to intervene based on the information in these reports. If, for example, important subgroups are responding at a lower rate, managers may wish to redirect interviewer effort toward cases in these low-responding subgroups. In order to re-prioritize field interviewer effort toward specific cases, managers must have the means to do so – for example, the use of "flags" in interviewer sample management systems. In this way, survey managers can control progress toward key indicators.

Given what we have learned in this investigation, the highest priority for new research in this area is to understand the circumstances under which centralized prioritization will lead to increased effort. We experienced variation in outcomes across the 16 interventions.

Understanding the sources of this variation may help researchers design more effective interventions. What factors mitigate the effectiveness of these experimental treatments? Is it a factor that varies across interviewers, or other factors that vary across samples? Or is it interactions between features of the design? For example, it appears that when interviewers have small workloads where all cases are receiving high priority (as in the NSFG second phase), centralized prioritization will be less effective. In addition, the consequences of using proxy indicators for nonresponse bias need to be evaluated. Understanding when this practice produces the desired results may require methodological "gold standard" studies designed specifically to investigate this question. There is certainly more work to be done in the development of these proxy indicators. In the case of the NSFG, demographic variables such as age, sex, race, and ethnicity are predictive of the key survey measures (Martinez et al. 2012). This may not be true for every study. A recent study by Peytcheva and Groves (2009) found that the types of demographic variables used to define some of our interventions were not predictive of nonresponse bias in the 23 specialized studies that they examined. More work is needed to develop "tailored" paradata suited for predicting the key survey variables of each particular study. Finally, although our focus was on the risk of nonresponse bias, other sources of error need to be included in the planning and execution of responsive designs. The tools outlined in this article are a valuable first step toward a "total survey error" perspective for responsive designs.

5. References

Couper, M.P. (1998). Measuring Survey Quality in a CASIC Environment. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas, TX.

Couper, M.P. and L. Lyberg (2005). The Use of Paradata in Survey Research. Proceedings of the International Statistical Institute Meetings.

Durrant, G.B., D'Arrigo, J., and Steele, F. (2011). Using Paradata to Predict Best Times of Contact, Conditioning on Household and Interviewer Influences. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174, 1029–1049.

Fienberg, S.E. and Tanur, J.M. (1988). From the Inside Out and the Outside In: Combining Experimental and Sampling Structures. The Canadian Journal of Statistics, 19, 135–151.

Fienberg, S.E. and Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. Science, 243, 1017–1022.

Groves, R.M. (1989). Survey Errors and Survey Costs. 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, 169, 439–457.

Groves, R.M., Benson, G., Mosher, W.D., Rosenbaum, J., Granda, P., Axinn, W., Lepkowski, J.M., and Chandra, A. (2005). Plan and Operation of Cycle 6 of the National Survey of Family Growth. Vital and Health Statistics, Series 1, No. 42. Hyattsville, MD: National Center for Health Statistics (Available from http://www.cdc.gov/nchs/data/series/sr_01/sr01_042.pdf, accessed October 11, 2011).

Groves, R.M., Mosher, W.D., Lepkowski, J., and Kirgis, N.G. (2009). Planning and Development of the Continuous National Survey of Family Growth. National Center for Health Statistics. Vital Health Statistics, Series 1, No. 48. Hyattsville, MD: National Center for Health Statistics (Available from http://www.cdc.gov/nchs/data/series/sr_01/sr01_048.pdf, accessed October 11, 2011).

Kirgis, N. and Lepkowski, J.M. (forthcoming). Design and Management Strategies for Paradata-Driven Responsive Design. Improving Surveys with Paradata: Analytic Use of Process Information, Frauke Kreuter (ed.).

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), 17, 389–407.

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, National Center for Health Statistics, 2(150).

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. Hoboken, N.J.: Wiley.

Martinez, G., Daniels, K., and Chandra, A. (2012). Fertility of men and women aged 15–44 years in the United States: National Survey of Family Growth, 2006–2010. National health statistics reports; no. 51. Hyattsville, MD: National Center for Health Statistics (Available from http://www.cdc.gov/nchs/data/nhsr/nhsr051.pdf, accessed August 12, 2012).

National Research Council (2007). Using the American Community Survey: Benefits and Challenges. Panel on the Functionality and Usability of Data from the American Community Survey. Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Constance F. Citro and Graham Kalton (eds). Washington, D.C. The National Academies Press.

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, 21–29.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness of Response. Survey Methodology, 35, 101–113.

Stoop, I.A.L., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving Survey Response: Lessons Learned from the European Social Survey. Chichester, West Sussex, U.K. Hoboken, N.J.: Wiley.

West, B.T. (2013). An Examination of the Quality and Utility of Interviewer Observations of Household Characteristics in the National Survey of Family Growth. Journal of the Royal Statistical Society, Series A (forthcoming).

Received September 2011
Revised March 2012