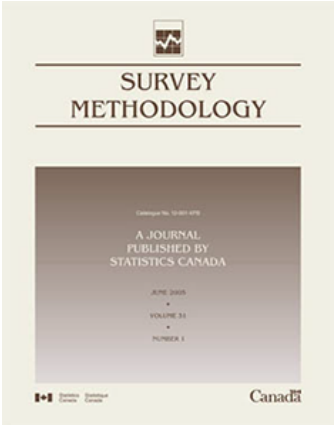
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Science and survey management



by Roger Tourangeau

Release date: June 24, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 3-28
Statistics Canada, Catalogue No. 12-001-X

Science and survey management

Roger Tourangeau[1]

Abstract

It is now possible to manage surveys using statistical models and other tools that can be applied in real time. This paper focuses on three developments that reflect the attempt to take a more scientific approach to the management of survey field work: 1) the use of responsive and adaptive designs to reduce nonresponse bias, other sources of error, or costs; 2) optimal routing of interviewer travel to reduce costs; and 3) rapid feedback to interviewers to reduce measurement error. The article begins by reviewing experiments and simulation studies examining the effectiveness of responsive and adaptive designs. These studies suggest that these designs can produce modest gains in the representativeness of survey samples or modest cost savings, but can also backfire. The next section of the paper examines efforts to provide interviewers with a recommended route for their next trip to the field. The aim is to bring interviewers' field work into closer alignment with research priorities while reducing travel time. However, a study testing this strategy found that interviewers often ignore such instructions. Then, the paper describes attempts to give rapid feedback to interviewers, based on automated recordings of their interviews. Interviewers often read questions in ways that affect respondents' answers; correcting these problems quickly yielded marked improvements in data quality. All of the methods are efforts to replace the judgment of interviewers, field supervisors, and survey managers with statistical models and scientific findings.

Key Words: Survey management; Responsive design; Adaptive design; Optimal routing.

1. Introduction

Surveys are in trouble these days, faced with the twin dilemmas of rising costs and falling response rates (e.g., Tourangeau, 2017; Williams and Brick, 2018). Both trends have been apparent in the United States since the 1970s (Atrostic, Bates, Burt and Silberstein, 2001; Steeh, Kirgis, Cannon and Dewitt, 2001), but seem to have accelerated in the last ten years or so. The same trends hold throughout the developed world (de Leeuw and de Heer, 2002). It seems fair to say that survey researchers do not really know what hit them (although see Brick and Williams (2013), for a thoughtful exploration of the possible causes behind these trends). But it is clear that fewer and fewer people want to do surveys these days; the downward trend in response rates mainly reflects increasing resistance to surveys among members of the general public.

Partly in response to this global industry-wide crisis, researchers have taken a closer look at the impact of falling response rates on the accuracy of survey estimates and have also proposed various measures to counter declining response rates. For example, more and more surveys have begun to offer incentives, make use of advance letters, and increase the number of contact attempts they make.

But another trend has been the use of a range of methods to improve the *management* of surveys to reduce the potential for error, data collection costs, or both. In Section 2, we review these efforts, generally known as *responsive* and *adaptive* designs. In Section 3, we look at another method for reducing cost and increasing efficiency in face-to-face surveys. This method – *optimal routing* – involves survey managers giving field interviewers detailed instructions about which cases to try to interview and what

1. Roger Tourangeau, 1601 Third Avenue, Apt. 19AW, New York, NY 10128. E-mail: RTourang@gmail.com.

route to follow in their next venture into the field. In Section 4, we look at another development with the potential to improve the performance of interviewers with the computer audio-recording of interviews, or *CARI* (Hicks, Edwards, Tourangeau, McBride, Harris-Kojetin and Moss, 2010). CARI allows central office staff the opportunity to hear how the interviewers are administering the questions in the field and make midcourse corrections in their performance. Research has shown that field interviewers depart from script more often than telephone interviewers do (Schaeffer, Dykema and Maynard, 2010; West and Blom, 2017), presumably because telephone interviewers can be monitored and given feedback in real time. In this fourth section, we describe two experiments in which central office staff provided *rapid feedback* to field interviewers – feedback provided within two or three days of the interview. What these techniques have in common is replacing the judgment of interviewers and field staff with the evidence- based prescriptions of survey managers – that is, they are attempts to replace management art with management science. Finally, Section 5 presents some conclusions.

2. Responsive and adaptive design

Responsive and adaptive designs refer to a family of methods for tailoring field work to reduce bias, variance, or cost (see Chun, Heeringa and Schouten (2018); Schouten, Peytchev and Wagner (2017); and Tourangeau, Brick, Lohr and Li (2017), for reviews). With responsive designs, researchers use multiple phases of data collection to reduce survey costs or errors. Adaptive designs use various forms of case prioritization, tailoring, and rules for stopping data collection to achieve similar goals.

Groves and Heeringa (2006) got this particular ball rolling with their description of responsive designs:

> Responsive designs are organized about design phases. A design phase is a time period of a data collection during which the same set of sampling frame, mode of data collection, sample design, recruitment protocols, and measurement conditions are extant. For example, a survey may start with a mail questionnaire attempt in the first phase, follow it with a telephone interview phase on non-respondents to the first phase and then have a final third phase of face-to-face interviewing. … Note that this use of "phase" includes more design features than merely the sample design, which are common to the term "multi-phase sampling". (Pages 440-441)

Of course, the American Community Survey had been using a three-phase design (mail followed by telephone follow-up followed by face-to-face follow-up with a subsample of the remaining cases) just like the one Groves and Heeringa described years before Groves and Heeringa dubbed these "responsive designs" (U.S. Census Bureau, 2014).

Groves and Heeringa cite several surveys that used responsive designs but focus mainly on Cycle 6 of the National Survey of Family Growth (NSFG). Most of the surveys they discuss, including Cycle 6 of the NSFG, applied two-phase sampling (that is, they selected a subsample of the nonrespondents remaining at a certain point in the field period and restricted further follow-up to this subsample) and offered larger incentives or made other changes to the data collection protocol for these final-phase cases. The real

innovation in the NSFG was not in its use of multiple phases of sampling (which had been around since Hansen and Hurwitz (1946)) or multiple modes of data collection (in fact, in the NSFG, all the cases were interviewed face-to-face) but in the application of paradata and real-time propensity modeling to guide the field work. The subsampling of nonrespondents in the Cycle 6 of the NSFG was based on propensity models that were updated frequently and that incorporated information gleaned from prior contacts with the sample case. In the final phase of Cycle 6 the NSFG, data collection was restricted to certain sample areas, with areas with larger numbers of active cases and those with cases with relatively high estimated propensities more likely to be retained for further follow-up field work.

Another difference between responsive designs and more traditional multi-phase designs, at least conceptually, is the notion of phase capacity. Groves and Heeringa argue that a given phase of data collection approaches a limit in its ability to change the survey estimates (and reduce any biases). Once it reaches this capacity limit, a change in protocol may be needed to improve the representativeness of the sample and reduce bias. Ideally, the later phases of data collection bring in different types of respondents from the earlier phases, reducing any remaining nonresponse biases. Different types of people may be inclined to respond by mail from those who respond to a face-to-face interview; larger incentives may help recruit those who are not interested in the topic (Groves, Singer and Corning, 2000). In the best case, the different phases of data collection are complementary and, together, create a more representative sample than each of the individual phases.

2.1 Case prioritization and related strategies

Cycle 6 of the NSFG is an early example of a strategy known as *case prioritization* – deliberately allocating more effort to some sample cases than to others. Of course, survey managers have always given priority to some cases over others. Interviewers are instructed to make sure they keep appointments, for example, or to set "soft" refusal cases aside for a while. What is different about the recent uses of case prioritization is that they are not based on a case's disposition but on models of the case's response propensity. In the Cycle 6 of the NSFG, a probability subsample of cases was kept for further work, with the second phase sampling probabilities partly based on the predicted propensities of the remaining cases. Later efforts have been explicit in their use of response propensities to guide the field work.

Depending on which cases are prioritized, case prioritization can serve a variety of goals. For example, focusing field work on cases with high response propensities may maximize the final sample size or reduce the costs per case. Beaumont, Bocci and Haziza (2014) distinguish three potential goals for such designs:

1) Minimizing variance;
2) Minimizing nonresponse bias or some proxy for it, such as sample imbalance (Särndal, 2011; see also Schouten, Cobben and Bethlehem, 2009); or
3) Maximizing response rates.

The first and third goals are related in that maximizing response rates tends to produce larger samples and, as a result, lower sample variances. Although some researchers have begun looking at the use of such designs to reduce measurement errors (Calinescu, Bhulai and Schouten, 2013), most efforts to date have been attempts to reduce nonresponse bias or costs.

With Cycle 6 of the NSFG, it is not completely clear what the statistical goal was. Oversampling areas with larger numbers of remaining cases and those with higher-propensity cases would tend to maximize the final sample size and reduce costs per case. Consistent with this, Groves, Benson, Mosher, Rosenbaum, Granda, Axinn, Lepkowski and Chandra (2005) noted that "this design option placed large emphasis on the cost efficiency of the … [final] phase design to produce interviews, not on minimizing standard errors of the resulting data set". However, Groves et al. (2005) also said that the final phase of data collection was intended to produce a "more representative" sample (page 38) by altering the data collection protocol to appeal to sample members who had failed to respond earlier. However, targeting areas with more cases with high estimated response propensities – that is, the cases predicted to be easiest to get – might actually exacerbate any problems with representativeness by bringing in additional respondents similar to those who had already responded.

Most later applications of case prioritization have taken the opposite tack, attempting to equalize the overall response propensities by focusing the field effort on the hardest cases. To see why this is a reasonable strategy, it is useful to take a closer look at the mathematics of nonresponse bias.

2.2 Factors affecting nonresponse bias

Under a stochastic perspective (e.g., Bethlehem, 1988), the bias of the unadjusted estimator of a mean or proportion $(\hat{\bar{y}})$ can be expressed as

$$\text{Bias}(\hat{\bar{y}}) \approx \frac{\sigma_\phi \sigma_y \rho_{\phi,y}}{\bar{\phi}}, \tag{2.1}$$

where $\bar{\phi}$ and σ_ϕ are the mean and standard deviation of the response propensities, σ_y is the standard deviation of a survey variable, and $\rho_{\phi,y}$ is the correlation between the response propensities and that survey variable. As (2.1) clearly demonstrates, both the overall response rate $(\bar{\phi})$ and the variation in the response rates (σ_ϕ) play a role in the bias, so that trying to maximize the response rates (e.g., by prioritizing the relatively easy cases) or to equalize the response propensities (by prioritizing the harder cases) are both reasonable things to do.

As a number of researchers have pointed out, nonresponse bias is a property of a survey estimate not of a survey, and, as (2.1) makes explicit, two variable-level properties also affect the bias – the correlation between the survey variable and the response propensities $(\rho_{\phi,y})$ and the variability of the survey variable (σ_y), both of which vary from one survey variable to the next. Given that two of the ingredients in the bias expression are study-level factors and two are variable-level, the question arises how much of the variation in nonresponse bias is between surveys and how much is within surveys.

Brick and I (Brick and Tourangeau, 2017) attempted to address this issue by reanalyzing data from a study done by Groves and Peytcheva (2008). They examined 959 nonresponse bias estimates from 59 studies. Eight hundred and four of these bias estimates involved proportions; almost all the others were means. (Four of the estimates seemed problematic to us, so we dropped them from our reanalysis.) Like Groves and Peytcheva, we examined the absolute relative bias statistic (absolute relbias), or the absolute

difference between the respondent estimate and the full sample estimate divided by the full sample estimate:

$$R_i = \frac{|\theta_{ri} - \theta_{ni}|}{\theta_{ni}}, \quad (2.2)$$

in which R_i is the absolute relbias for statistic i, θ_{ri} is the estimated value for that statistic based on the respondents, and θ_{ni} is the corresponding full sample estimate. The absolute relbias is useful in that it puts all the bias estimates on the same metric the percentage by which the estimate is off. Our reanalysis also examined the absolute differences (the numerator in (2.2)) for the estimated proportions.

Table 2.1 displays various statistics from the reanalysis. For example, we calculated the correlation between the individual bias estimates and the study-level response rates; these results are shown in the top panel of the table. The middle three panels of the table show what happens when the average bias from the study is used in place of the individual bias estimates. Some of the correlations based on study-level averages are considerably higher than those based on the individual estimates, particularly when the data are weighted by the number of estimates from each study (r's of 0.40 to 0.55). The bottom two panels of the table show that there is a substantial study-level component to the nonresponse bias. For example, the R^2 estimates from a one-way ANOVA indicate that the between-study component accounts for 21 to 40 percent of the overall variation in the nonresponse bias estimates. The results from multi-level models lead to similar conclusions. This between-study component of the bias presumably reflects two main variables – the mean response propensity (reflected in the overall response rate) and the variation across respondents in the response propensities.

Table 2.1
Relationship between response rates and bias measures at the estimate and study level

	All statistics	Proportions only
Estimate-level correlations		
Response rate and absolute relbias	-0.191(n = 955)	-0.256 (n = 802)
Response rate and absolute difference	-	-0.323 (n = 802)
Unweighted study-level correlations		
Response rate and mean absolute relbias	-0.255 (n = 57)	-0.315 (n = 43)
Response rate and mean absolute difference	-	-0.246 (n = 43)
Study-level correlations weighted by number of estimates		
Response rate and mean absolute relbias	-0.402(n = 57)	-0.552 (n = 43)
Response rate and mean absolute difference	-	-0.508 (n = 43)
Study-level correlations weighted by mean sample size		
Response rate and mean absolute relbias	-0.413 (n = 57)	-0.247 (n = 43)
Response rate and mean absolute difference	-	-0.208 (n = 43)
Estimate-level ICCs from multilevel model		
Absolute relbiases	0.164 (n = 955)	0.161 (n = 802)
Absolute differences	-	0.509 (n = 802)
Estimate-level R^2 from one-way ANOVA		
Absolute relbiases	0.221(n = 955)	0.211 (n = 802)
Absolute differences	-	0.395 (n = 802)

The results in Table 2.1 are important because responsive and adaptive designs work primarily at the study level. For example, case prioritization generally either increases the overall response propensities or reduces the variation in the propensities, and these are the two main study-level variables affecting the level on nonresponse bias. In addition, if a design succeeds in reducing the overall variation in the response propensities, this will tend to attenuate the correlations between the propensities and the survey variables across the board. At the extreme, if there is no variation in the response propensities, the correlation with all the survey variables will be zero and there won't be any nonresponse bias. The results in Table 2.1 seem to contradict the view that response rates don't matter. Nonresponse rates are clearly an imperfect proxy for nonresponse bias, but they are an important predictor of the average level of bias in the estimates from a survey.

2.3 Experimental evaluations of responsive and adaptive designs

How well do responsive and adaptive designs achieve their goals? At the outset, I should note that our expectations shouldn't be too high. As we noted in an earlier paper (Tourangeau et al., 2017, page 208), these designs "represent an attempt to do more with less or at least to do as much as possible with less" in an increasingly survey unfavorable environment. To date, studies have used four basic strategies to achieve one or more statistical goals – multi-phase designs (like the one described by Groves and Heeringa, 2006), other types of case prioritization (in which different cases are slated to receive different levels of effort), adaptive contact strategies (changing the timing of contact attempts based on propensity models to maximize the chances of making contact), and tailoring of the field work or mode of data collection based on what is known about the cases before they are fielded. I briefly review some of the major efforts to evaluate each of these approaches.

Multi-phase designs and case prioritization. Peytchev, Baxter and Carley-Baxter (2009) report another study that, like Cycle 6 of the NSFG, employed a multi-phase design. They conducted a telephone study with two phases. The second phase used a much shorter questionnaire and offered a larger incentive than the first. Cases received up to twenty calls during Phase 1, with some cases getting even more. Overall, this phase produced a response rate of 28.5 percent. In Phase 2, the researchers subsampled the remaining nonrespondents, shortened the questionnaire from 30 to 14 minutes, gave a prepaid incentive of $5, and offered a conditional incentive of $20. (Phase 1 had offered only conditional incentives.) Phase 2 produced a response rate of 9.8 percent (or 35.5 percent overall). The evaluation of the design was based two sets of comparisons: Peytchev and his colleagues compared early and late respondents from Phase 1 and they compared Phase 1 to Phase 2 respondents. They reasoned that the late respondents (interviewed after at least six call attempts) from Phase 1 were unlikely to differ on the key study variables – reported crime victimizations of various sorts – from the early respondents (interviewed in five or fewer attempts) because they were recruited via the same protocol. The results indicated that the addition of the late Phase 1 respondents did not significantly change the estimates. In contrast, the authors believed the Phase 2 respondents were likely to differ from the Phase 1 respondents, because the changes in protocol would attract different types of respondents. There was some support for this line of argument for males.

The Phase 1 male respondents were more likely to report victimizations than the Phase 2 male respondents, with significant differences on four of six victimization rates. However, there was less evidence that the change in protocol in Phase 2 affected the estimates for females. In addition, even within the Phase 1 sample, there were differences between male cases who never refused and those who were converted after refusing. Like the Phase 2 male respondents, the converted Phase 1 male refusals also showed significantly lower victimization rates on four of six key estimates. This suggests that the refusal conversion protocols changed the make-up of the Phase 1 sample and did not just bring in more of the same type of respondents.

Peytchev, Riley, Rosen, Murphy and Lindblad (2010) report a study that tailored the data collection protocol for different groups of cases from the outset. Their study involved a panel survey and the response propensities for each case was estimated using information from the prior round. Cases with low predicted response propensities were randomly assigned to an experimental or control treatment. For most of the data collection period, interviewers got a $10 bonus for each completed interview with one of the control cases, but $20 for each completed interview with one of the experimental cases. (During Phase 1, there was no bonus for control interviews and a $10 bonus for experimental interviews.) There was little difference in the final response rates for the two groups of cases (89.8 percent for the control cases versus 90.8 percent for the experimental cases) or in the average number of contact attempts per case (5.0 for the controls versus 4.9 for the experimental cases). Although the variance in the estimated response propensities was lower among the experimental cases, the estimated nonresponse biases (based on the correlations between the survey variables and the fitted response propensities) were higher.

Another set of experiments illustrates some of the practical difficulties with case prioritization. Wagner, West, Kirgis, Lepkowski, Axinn and Kruger Ndiaye (2012; see also Lepkowski, Mosher, Groves, West, Wagner and Gu (2013)) carried out 16 experiments over the course of Cycle 7 of the NSFG, which fielded 20 quarterly samples. The experiments examined the effectiveness of "assigning a random subset of active cases with specific characteristics to receive higher priority from the interviewers… The first objective of these experiments was to determine whether interviewers would respond to a request to prioritize particular cases" (Wagner et al., 2012, page 482). In only seven of the 16 experiments did the priority cases actually receive significantly more calls than the control cases, and only twice did this lead to a significant increase in response rates for the priority cases. Additional experiments attempted to shift the effort of NSFG interviewers from trying to complete main interviews to trying to complete screeners during one week of the field period. This intervention did lead to more screener calls than in prior or later weeks, but the impact on the number of *completed* screeners varied across quarters. In both cases, the efforts at case prioritization in Cycle 7 of the NSFG had some impact on what the interviewers did, but less impact on the intended survey outcomes, such as response rates.

Statistics Canada has also begun implementing responsive designs for its CATI surveys and carried out two experiments assessing these designs. Both experiments used three phases of data collection with case prioritization in one phase (Laflamme and Karaganis, 2010; Laflamme and St-Jean, 2011). In Phase 1, cases were categorized by response propensities; in Phase 2, cases were randomly assigned either to the

responsive collection condition (in which cases were assigned priorities and the high priority cases got more calls) or the control condition; and in Phase 3, all remaining cases got the same treatment. In Phase 2, the priority cases in the responsive collection group were apparently those with high predicted response propensities. The goal in Phase 3 was to equalize response propensities across key subgroups. Once again, the results indicated modest effects. The overall response rates were essentially unaffected by case prioritization. In one survey, the response rates were 74.0 percent for the control group versus 74.1 for the responsive collection group; in the other, the control group had a slightly higher response rate (73.0 versus 72.8 percent). This is a little surprising since the responsive collection targeted the easier cases in Phase 2. In addition, neither the new three-phase design nor the responsive collection protocol had a clear effect on the representativeness of the samples, but may have decreased the number of interviewer hours (see Table 2.2 in Laflamme and St-Jean (2011)). Still, reducing costs without reducing representativeness may represent a worthwhile, if modest, advance.

Adaptive contact strategies. Can survey managers improve the rate at which sample members are contacted by modelling the best time to contact them? Although many papers have explored optimal times for contacting sample members in surveys, few have examined whether these "optimal" call schedules produce gains empirically. Wagner (2013) is an exception. He reported five experiments that used models to predict whether a given sample household would be contacted on the next call attempt in each of four call "windows" (e.g., Tuesday through Thursday from 4 p.m. to 9 p.m.). Similar models were used in telephone (the Survey of Consumer Attitudes, or SCA) and face-to-face (Cycle 7 of the NSFG) surveys. The models were used to identify the best call window (the one with the highest probability of a contact) for each sample household. In the experimental groups, cases were moved to the top of the list for calling in that window (in the SCA) or field interviewers received that window as the recommended time to contact the household (in the NSFG).

Three experiments involved the SCA. In the first, the proportion of calls producing a contact was higher for the experimental cases than for the controls (12.0 percent versus 9.9 percent), but the strategy seemed to backfire for cases who had initially refused, with *lower* contact rates among the initial refusals in the experimental group. A second experiment varied the call window for experimental cases after an initial refusal but this strategy lowered the overall proportion of calls producing a contact. The final SCA experiment still found that the contact rate for refusal conversion calls was lower in the experimental group than in the control group. The results in the NSFG were also somewhat disappointing. The field interviewers apparently ignored the recommended call windows; only 23.6 percent of the experimental cases were contacted in the recommended window (versus 23.0 percent in the control group). We had a similar experience in our effort to get interviewers to follow an optimal route in their trips to the field (see Section 3.1 below).

Tailored field work. Luiten and Schouten (2013) report an experiment that tailored the data collection approach to different subgroups in the Dutch Survey of Consumer Sentiments (SCS). The goal was to equalize response propensities across the subgroups. The SCS consists of repeated cross-sectional surveys and, based on earlier rounds, Luiten and Schouten fit contact and cooperation propensity models based on

demographic characteristics of the sample members; these variables were available for the entire sample from the population registry. There were two phases of data collection. In the initial phase, cases with lowest estimated cooperation propensities were sent a mail questionnaire; those with the highest estimated propensities were invited to complete a web survey; and those in the middle were given a choice between mail and web. The second phase consisted of following up nonrespondents by telephone. Cases in different contact propensity quartiles were assigned to different call schedules. Those with the highest estimated contact propensities were fielded later in the field period and called during the day; those in the second highest quartile were called twice at night and then switched to a schedule alternating daytime and nighttime calls; and those in the lowest two contact propensity quartiles were called on every shift of every day. Finally, the best telephone interviewers were assigned to the cases with the lowest estimated cooperation propensities and the worst telephone interviewers were assigned to the cases with the highest estimated cooperation propensities. The control group for the experiment was the regular SCS, which is a CATI-only survey.

Although the adaptive field work group had only a slightly higher response rate than the regular SCS (63.8 percent versus 62.8 percent, a non-significant difference), the representativeness of the experimental sample, as measured by the R-indicator, was significantly higher than that of the control sample. (The R-indicator, introduced by Schouten, Cobben and Bethlehem (2009), is based on the variation in the estimated response propensities. A higher number indicates less variation and therefore a more representative sample.) Table 2.2 below shows that the adaptive field work did lower the variation in both contact and cooperation rates. Across contact propensity quartiles, the contact rates ranged from 84.2 percent to 96.9 percent in the regular SCS; in the experimental sample, the range was from 87.1 to 95.3. The adaptive design also lowered variation in the cooperation rates. Still, the costs for the adaptive design were marginally higher than those of the SCS and the overall cooperation rate was significantly *lower* in the experimental sample. Unfortunately, as this study illustrates, reducing the variability in the response propensities often means not trying as hard to get the easiest cases and this may lower the overall response rate.

Table 2.2
Contact and cooperation rates, by propensity quartile groups

	Contact rates	
Contact propensity quartile	**Experimental**	**Control**
Lowest Contact Propensity	87.1	84.2
Second Lowest Contact Propensity	96.6	94.5
Second Highest Contact Propensity	93.7	95.7
Highest Contact Propensity	95.3	96.9
	Cooperation rates	
Cooperation propensity quartile	**Experimental**	**Control**
Lowest Cooperation Propensity	65.1	62.7
Second Lowest Cooperation Propensity	71.4	68.4
Second Highest Cooperation Propensity	72.8	75.3
Highest Cooperation Propensity	74.7	79.2

Source: Tourangeau et al. (2017); data from Luiten and Schouten (2013).

2.4 Simulation studies

Besides the experiments discussed in the previous section, three additional studies have used simulations to explore the properties of responsive and adaptive designs.

Stopping rules. Lundquist and Särndal (2013) used data from the 2009 Swedish Living Conditions Survey (LCS) to explore the impact of various "stopping rules", rules for ending data collection. The LCS follows a two-phase data collection strategy, with up to 20 telephone contact attempts in the first phase of data collection followed by ten more in the second phase. They noted that continuing to follow the same data collection protocol "will produce very little change in the estimates beyond a certain 'stability point' reached quite early in the data collection" (page 561). This is quite similar to Groves and Heeringa's (2006) notion of "phase capacity", or the point at which a given data collection protocol begins to achieve diminishing (or vanishing) returns. Sturgis, Williams, Brunton-Smith and Moore (2017) present results suggesting that this stability point may be reached quite early during the field period. They examined estimates derived from 541 questions from six face-to-face surveys in the U.K. They found that the expected proportions were, on average, only 1.6 percent from the final estimate after a single contact attempt and were off by only 0.4 percent after five attempts. These results suggest that, from the vantage point of reducing bias, a lot of field effort is wasted.

Lundquist and Särndal show that the estimated nonresponse bias (based on three variables available for both respondents and nonrespondents from the Swedish population register) in the LCS was lowest after five to ten call attempts and actually got progressively worse thereafter. The second phase of data collection, which increased the response rate from 60.4 percent to 67.4 percent, made the nonresponse biases worse for two of the three register variables. They examined three alternatives to continuing the same protocol up to 30 attempts. They divided the sample into eight subgroups based on education, property ownership, and national origin. Under the first alternative response rates for each of eight the subgroups would be checked at call 12 of the initial phase of data collection and again at call 2 of the second phase; data collection would end for subgroups with response rates of 65 percent or better at these points. This strategy would have yielded a lower response rate (63.9 percent) than the actual protocol but a sample that was more closely aligned with the population on eight demographic characteristics. The second alternative they examined would have ended data collection for a subgroup as soon as its response rate reached 60 percent and the third alternative, as soon as the subgroup response rate reached 50 percent. The 50 percent strategy would have produced the most balanced sample of all and would have reduced the total number of call attempts by more than a third. In part, this strategy worked so well because it would have lowered the response rates in the high propensity subgroups so they were closer to those in the low propensity subgroups. As in the study by Peytchev and colleagues (Peytchev et al., 2009), continuing with the same data collection protocol seemed to do little improve the representativeness of the sample, and may in fact have reduced it.

In a related effort, in 2017, the Medical Expenditure Panel Survey (MEPS) used a stopping rule based on a propensity model. MEPS is a rotating panel study. Each year a new panel of about 10,000 addresses

is selected from a sample of households that completed National Health Interview Survey the previous year. Sample households were asked to complete two MEPS interviews in their first year, two in the second, and a fifth in the third year. The survey is continuous, with interviews conducted throughout the year. The stopping rules were applied in two stages in the first half of 2017: first to cases in their third round (a relatively soft start, since most Round 3 interviews were scheduled by telephone and most respondents were cooperative, having already participated twice), and then to Round 1 cases. Interviewers are often reluctant to comply with directions to stop contacting a case after a specific number of attempts. The MEPS approach was to remove low propensity cases with too many attempts – generally six – from the interviewer assignment and have a supervisor review them. Supervisor could move a case back into the interviewer's assignment if there was some reason to believe the case might be completed, but most of the time these cases were closed out (Hubbard, 2018). Overall, implementing the stopping rule reduced the number of in-person attempts by 8,500, producing a large saving in field costs.

Different case prioritization strategies. In a later paper, Särndal and Lundquist (2014b) simulated the effects of two methods for equalizing response propensities across cases, using data from the Living Conditions Survey and the Party Preference Survey. Under the first method (the *threshold* method), no further follow-up attempts are made to cases whose response propensities have reached some threshold (lower than the overall target response rate). This is similar to the strategies examined in their earlier paper (Lundquist and Särndal, 2013). Under the other method (the *equal proportions* method), at various points during the field period (e.g., after three, six, or nine call attempts), the portion of the sample with the highest response propensities is set aside and field work continues only for the remaining cases. In both surveys, both methods for equalizing the response propensities reduced the distance between the respondents and the full sample on a set of auxiliary variables, as compared to continuing to field all remaining nonrespondents, as was done in the actual surveys. Another conclusion from this study is that calibrating the sample using the auxiliary variables removed some of the nonresponse bias, but that bias was reduced even further when the set of respondents was more closely aligned with the population in the first place. This is an important finding, since the same variables available for fitting propensity models are also available for post-survey adjustments, and it is not clear whether equalizing response rates (or response propensities) during data collection is more effective than simply adjusting the case weights afterwards. Särndal and Lundquist (2014b) find gains for both.

Beaumont, Bocci and Haziza (2014) report another simulation study that examines the impact of case prioritization. They contrasted four strategies: 1) *constant effort* (no case prioritization); 2) *optimal effort* (by reducing calls to members of groups approaching their target response rate); 3) *equalizing response rates* across groups (by concentrating calls on low response propensity groups); and *maximizing the overall response rate* (by concentrating calls on high propensity groups). The simulations by Beaumont and his colleagues assumed three different scenarios – uniform response propensities, uniform response propensities within groups, and response propensities that are highly $(r = 0.67)$ correlated with the survey variable of interest. (In addition, the simulation assumed that the sample consisted of three

subgroups, that calls yielding an interview were 25 times more expensive than ones that didn't, that calls to a case were capped at 25, and the survey had a fixed data collection budget.)

The simulations supported three major conclusions. First, when response propensities are constant overall or constant within each group, all the effort strategies produce unbiased estimates, but when the propensities were strongly related to the survey variable, all of them produced bias. Second, neither the R-indicator nor the nonresponse rate was a good indicator of nonresponse bias or nonresponse variance. Finally, when response propensities were known, the optimal effort strategy produced somewhat lower root mean square error than the other strategies (see Table 2.2 in Beaumont et al. (2014)) and the strategy that attempted to maximize response rates produced the worst. The optimal effort strategy resembles the approaches explored by Lundquist and Särndal (2013). Of course, a practical difficulty is that response propensities are not known with real surveys, and they may not be accurately estimated from the available auxiliary variables.

2.5 Summary

Table 2.3 summarizes the results from the experimental and simulation studies. In general, they show how hard it is to raise response rates in the current environment. For example, only two of the 16 experiments described by Wagner and his colleagues significantly raised response rates in the NSFG (Wagner et al., 2012). Some studies (e.g., Luiten and Schouten, 2013) demonstrate reductions in variation in response rates across subgroups of the sample, although in one study (Peytchev et al., 2010) this apparent reduction in the variation in estimated response propensities appeared to *increase* nonresponse bias rather than reduce it. Laflamme and St-Jean (2011) reported that responsive design reduced costs relative to the standard protocol, but Luiten and Schouten (2013) reported that an adaptive design increased the costs per case. Across all the studies (including Cycle 6 of the NSFG), then, responsive and adaptive designs appeared to produce some gains in sample representativeness, but had little effect on overall response rates or overall costs.

Several non-experimental studies come to similar conclusions. These studies compare the final survey estimates with those that would have been obtained without the final phase of data collection, when a major change in the data collection protocol was introduced. For example, Groves and his colleagues (Groves et al., 2005) showed that the final phase of data collection in Cycle 6 of the NSFG, which boosted the overall response rate from 64 to 80 percent, also decreased variation in the response rates across subgroups (see also Axinn, Link and Groves, 2011). This is similar to the experimental results reported by Peytchev, Baxter and Carley-Baxter (2009) who found that major changes in protocol (larger incentives and a shorter questionnaire) produced changes in the study estimates, at least for males. However, the changes were generally small – less than two percentage points.

Table 2.3
Selected study characteristics and outcomes, by study

Experimental Study	Statistical Goal	Intervention	Results
Peytchev et al. (2010)	Equalize response propensities	Bonus for interviewers for completing high priority cases	• Variance in response propensities lower in experimental group • Response rate 1.5% higher in experimental group • Estimated bias *higher* in experimental group
Wagner et al. (2012)	Increase response rates, improve representativeness	Case prioritization Screener week	• Significantly increased number of calls to priority cases in seven of 16 experiments • Significantly increased response rate in two experiments • Increased number of screening calls
Laflamme and St-Jean (2011)	Increase response rates (Phase 2), equalize response propensities (Phase 3)	Categorization and prioritization of cases	• Less variance in response propensities in experimental group • Response rate 1.5% higher in experimental group
Wagner (2013)	Increase contact rate per call	Models used to assign cases to optimal call window	SCA • Contact rate improved (12.0 vs. 9.9 percent) • No change in response rate NSFG • Interviewers did not follow recommended call window
Luiten and Schouten (2013)	Equalize response propensities	Initial mode (mail versus Web) varied by propensity quartile Hard cases assigned to best telephone interviewers, easiest to worst telephone interviewers	• Lower cooperation rate in adaptive group • R-indicator significantly improved in adaptive group • Reduced variation in contact and cooperation rates in adaptive group • No significant difference in costs or response rates
Simulation Study	**Statistical Goal**	**Intervention**	**Results**
Lundquist and Särndal (2013)	Increase sample balance, reduce nonresponse bias	Stopping data collection for a subgroup once a target rate achieved for that subgroup	• Lowest response rate threshold produced the highest balance • Lowest threshold also achieved lowest nonresponse bias (on three registry variables)
Särndal and Lundquist (2014a, b)	Increase sample balance, reduce nonresponse bias	12 stopping rules	• Lowest response rate threshold again produced the highest balance • Lowest threshold also achieved lowest nonresponse bias on three registry variables • Both balance in data collection and calibration reduce nonresponse bias
Beaumont, Bocci and Haziza (2014)	Optimal effort, equalize response rates, maximize overall response rate	Four case prioritization strategies (constant effort, reduce effort for groups approaching target response rate, prioritize low propensity cases, prioritize high propensity cases)	• With uniform response propensities, all four strategies yield unbiased estimates • When response propensities strongly related to survey variables, all strategies produce biased estimates • With known propensities, optimal strategy yields best root mean square error (RMSE); maximizing the response rate, the worst

3. Optimal routing

Case prioritization, like adaptive design more generally, is an example of an intervention in the data collection process intended to reduce error, costs, or both. In the next two sections, we examine two other interventions designed to improve data collection outcomes – optimal routing and rapid feedback to

interviewers. The first uses a variation on case prioritization; the second focuses on reduction of measurement error.

Prioritizing high-value cases. One problem with the existing studies on case prioritization is that they have all used estimated response propensities as the basis for prioritization. The response propensity may be a useful summary of the variables used to model the propensity but may not fully reflect the researchers' priorities. In an earlier paper (Tourangeau et al., 2017), we proposed a different basis for case prioritization. Under our scheme, the cases receiving the highest priority should be the ones with the highest ratio of anticipated value to anticipated cost:

$$B_i = \frac{\hat{\rho}_i W_i V_i}{C_i} \tag{3.1}$$

where B_i represents the benefit-to-cost ratio for case i; the numerator is the product of the case's estimated response propensity $(\hat{\rho}_i)$, its weight (W_i), and some measure of its value for the research (V_i); and the denominator represents the likely cost of completing the case (C_i). For example, the value assigned to a member of a rare subgroup may be higher than that assigned to a member of a larger group. Or the value of a case may be an estimate of its impact on reducing the distance between the current sample from a vector of population benchmarks. Because it includes the estimated propensity and the weight in the numerator, the scheme in (3.1) may result in giving priority to "easy" cases or give lower priority to cases from oversampled subgroups, which would have lower weights. Thus, a lot hinges on how the value of a case is assessed. We attempted to apply a version of (3.1) in conducting a pilot test of a strategy that we call *optimal routing*.

The pilot test. Our pilot test was done as part of the Population Assessment of Tobacco and Health (PATH) Reliability and Validity Study (the PATH-RV Study; Tourangeau, Yan, Sun, Hyland and Stanton, 2019), a study designed to assess the reliability and validity of answers to the Wave 4 PATH Study questionnaires. (The PATH Study is a major longitudinal study of tobacco use, and the study sponsors wanted to be sure the questions yielded reliable responses.) In the PATH-RV study, a sample of 524 respondents completed the PATH questions twice, roughly two weeks apart. There were two questionnaires, one for adults (18 years old and older) and one for youths (12 to 17 years old). Given the aims of the reliability study, we deemed youth cases to be twice as valuable as adult cases (because youths were rarer and harder to interview than adults) and reinterviews 1.5 times more valuable than initial interviews. Thus, an initial interview with an adult was worth a value of 1; an adult reinterview, 1.5; an initial youth interview, 2; and a youth reinterview, 3. We used these values in the place of V_i in (3.1). Because the sample for the PATH-RV study was nearly equal probability, we ignored the weights. However, we did incorporate an estimate of the likelihood that the case would cooperate on the next contact attempt. We also developed a program that calculated an optimal route for contacting a set of cases on a given day, partly in an effort to minimize travel and interviewer time – that is, to minimize C_i in (3.1).

The system we developed had two components. The first one estimated the likelihood that each remaining case would cooperate on the next contact attempt. The models for this first component used sociodemographic information from the Census Planning Database for block groups and the history of previous contact attempts whenever at least one contact attempt was available. For cases with no prior contact attempts, we used a logistic regression model to estimate the cooperation propensity; for cases with prior contact attempts, we used a proportional-hazards Cox regression model. The second component was a routing system that reviewed respondent-level information and produced an interviewer's schedule for a given day. The set of cases in the day's assignment reflected the anticipated value of the cases. All the sample cases were geolocated, allowing us to estimate the travel time between each pair of cases for a given hour of the day. The routing system took as input the feasible tasks for a given case (e.g., it did not schedule a reinterview until the initial interview had been completed), along with the case's geographical location, estimated duration of the task, case value, and response propensity. It then computed the shortest driving route with the highest possible expected value and selected a set of tasks that could fit in a working day for a given interviewer. The route delivered to interviewers included the sequence of cases and tasks that we expected interviewers to attempt. It took appointments into account, and the route was constructed to ensure that interviewer could arrive at their appointments on time.

The experimental design. We conducted an experiment that compared interviewer performance on "treatment" days when we gave them the list of cases to try to get along with a suggested route to follow in pursuing those cases with "control" days when we gave them no special instructions about which cases to work or how to work them. The data collection for the experiment took place between October and December, 2017.

Before the start of data collection, interviewers selected at least six days during which they would work only on the PATH-RV study. We then randomly allocated three of those days to the control arm of the experiment and three to the treatment arm. On control days, we sent interviewers an email in which we asked them to "use their best judgement on how to contact" their caseload. On treatment days, we sent them an email that included a list of cases that we wanted them to work and the route they were to follow. Interviewers were told to follow our recommendations "if at all possible". During the training sessions and in the email accompanying the selected route, we discouraged deviations from the instructions, but allowed them if the interviewers judged them necessary to account for unforeseeable events, such as traffic accidents.

Fifty-three interviewers participated in the experiment. Changes to the days the interviewers worked on our study, together with the depletion of the pool of open cases in the final days of the study, produced a reduction in the number of treatment and control days actually available for the interviewers. Ultimately, we had a total of 220 observations.

Interviewer compliance and interviewer efficiency. Did the interviewers follow the instructions we sent them in the treatment email? Well, they did some of the time. There was an average overlap of 62 percent between the cases we recommended for a given treatment day and the cases the interviewers

actually worked that day. What is particularly striking is that there was, on average, a 52 percent overlap on the cases selected by our model and the cases selected by the interviewers on the control days, when we didn't give them any instructions. This small difference between the treatment and control days partly reflects the limited number of cases that could be worked on any given day. As a result, the decisions that interviewers would have made on their own were often close to what we thought would have been optimal, putting a low ceiling on the possible impact of the treatment.

Still, there was only moderate compliance with instructions by the interviewers. A Census Bureau test had similar results. The test was done in eight areas in Philadelphia, Pennsylvania (Walejko and Miller, 2015). In some areas, interviewers were assigned seven high priority cases each day; these high priority cases were those deemed most likely be interviewed on the next contact attempt, according to response propensity models. As with our experiment, interviewer compliance was an issue. As Walejko and Miller (2015) put it: "The ability of response propensity models to identify promising cases for daily contact, however, remains unclear after this pilot test because interviewers did not dutifully work priority cases."

Was there any sign in our study that the optimal routing treatment improved interviewer efficiency? We examined five outcomes of interest:

1) The number of miles interviewers traveled;
2) The hours they spent;
3) The number of contacts per completed interview;
4) The number of completed cases; and
5) The average value of the cases completed.

The first two variables reflect the impact of the treatment on the costs of collection. We also wanted to assess whether our routing system reduced the number of contact attempts needed to complete a case – that is, whether it made the interviewers more productive. Similarly, we examined whether the treatment increased the number of completes and whether the completed cases had higher values on average on the treatment days than on the control days. Our analyses of the effects of the treatment are shown in Table 3.1. The models include random effects for each interviewer and pool the effect of the treatment across interviewers. The top two panels show the estimates for the intercept and treatment effects under an intent-to-treat model (ignoring whether the interviewers actually followed our instructions), and the second panel incorporates a measure of the interviewers' compliance with the instructions. None of the outcome measures shows a significant treatment effect, although there were significant compliance main effect for miles and contacts – interviewers traveled fewer miles and made fewer contacts when they did what we suggested (whether we conveyed those suggestions to them or not). Although there was a treatment by compliance interaction effect on contacts, the net effect of the treatments seems to have been negative.

Table 3.1
Estimated intercepts and effects (and standard errors), by outcome and model

	Miles	Hours	Contacts	Completes	Value
Intent-to-treat					
Intercept	76.8 (7.0)	5.26 (0.37)	5.40 (0.53)	0.81 (0.17)	1.73 (0.26)
Treatment	8.06 (5.6)	0.04 (0.27)	-0.28 (0.48)	-0.15 (0.16)	-0.24 (0.30)
Incorporating compliance					
Intercept	89.2 (8.99)	5.69 (0.46)	7.66 (0.62)	0.98 (0.21)	1.84 (0.40)
Treatment	0.30 (9.87)	-0.30 (0.48)	-1.35 (0.79)	0.15 (0.28)	0.25 (0.53)
Compliance	-28.9 (13.4)	-1.01 (0.65)	-5.32 (1.06)	-0.03 (0.37)	-0.28 (0.70)
Treatment x compliance	20.5 (17.4)	0.86 (0.85)	3.07 (1.38)	-0.55 (0.48)	-0.87 (0.92)

Note: Results based on 53 interviewers and 220 total observations.

Interviewer reactions. Debriefings with the interviewers revealed some of the reasons for their relatively low levels of compliance with our recommendations. Although the interviewers were generally positive about the routing system, they had several reservations about it. The behavior of interviewers reflects the goal of getting completed interviews, but their implicit assumption is that all completes are equally valuable. However, our routing system reflected a specific definition of the expected value of a case and also an estimate of its cost. As a result, it sometimes omitted cases that were close to the households on the recommended route. Interviewers indicated that a priority list or a scoring of the cases by their value would have made the decisions of the automatic system more comprehensible and also would have allowed them to incorporate those values into their own workday planning. In addition, interviewers sometimes disagreed with the suggested routes because of circumstances that could not be observed by our routing system. With any adaptive design strategy (or, more generally, with any planning system), there is the risk of missing some useful information and this may undercut compliance.

The debriefing also called attention to some of the assumptions embedded in the model. For instance, we established a single time window for all interviewers as the most likely time they would be working. This allowed us to account for daily traffic patterns in our recommendations. But a different route might have been better than the one we recommended for a different time of day when traffic was lighter or heavier. All the interviewers who took part in the experiment were experienced field interviewers, and some reported they felt that detailed routing instructions were tantamount to discounting their abilities and experience. In their opinion, the system might be a good tool for novice interviewers, but, for them, it signaled a lack of confidence on the part of the survey managers. Finally, they all reported that one reason they worked as field interviewers was being able to plan their own workday. Many of these same factors doubtless played a role in the limited success of the attempts by Wagner and his colleagues (see Section 2.3) and the Census test to change interviewer behavior.

Despite these obstacles to compliance, research has shown that interviewers are sensitive to incentives. Tourangeau, Kreuter and Eckman (2012) demonstrated that interviewers in a telephone study completed more screeners when they were given a bonus for each screener they completed and they completed more main interviews when they were given a bonus for each completed main interview. Perhaps similar incentives could be used to encourage interviewers to complete high priority cases or to minimize travel

time. For example, interviewers could receive a small bonus for every high priority case they contact. Clearly, we need to figure out how to get interviews to follow instructions if our interventions are going to have any impact.

Other studies of interviewer travel. More recently, Wagner and Olson (2018) carried out an extensive analysis of interviewer travel in two face-to-face surveys, the National Survey of Family Growth (NSFG) and the Health and Retirement Survey (HRS). Both surveys feature national area probability samples and the Survey Research Center at the University of Michigan carries out the field work for both. The surveys have different target populations – people from 15-44 years old in the NSFG and from 51-56 years old in the HRS. The authors examined how far interviewers travelled and how many sample areas they visited on each day they worked. In both studies, interviewers visited about two areas, on average, on each day they worked but they travelled about 30 miles more in average in the NSFG than in the HRS (85.4 versus 53.4). Wagner and Olson found that travelling to more areas was associated with more contact attempts, but with fewer contacts made and fewer interviews completed (see their Table 4.1). Although theirs is an observational study and not an experiment, it is consistent with the results of our pilot study; more travel seems to reduce the number of contacts made and interviews completed. However, the causal direction of this finding is quite ambiguous. It could be that travel time reduces the time interviewers have left to contact and interview sample cases, but it also could be that interviewers keep going when their contact attempts don't yield a positive outcome, moving on to different sample areas.

4. Rapid CARI feedback

Interviewers can contribute in several ways to the total error of a survey estimate, affecting coverage, nonresponse, and measurement errors (Schaeffer, Dykema and Maynard, 2010; West and Blom, 2017). There can be complex interactions among these different interviewer-related error sources. For example, there may be a tradeoff between coverage and nonresponse errors (Tourangeau et al., 2012); in our study, the interviewers with the highest response rates also found the fewest eligible households. In a series of papers, West and his colleagues (West and Olson, 2010; West, Kreuter and Jaenichen, 2013; West, Conrad, Kreuter and Mittereder, 2018) have shown that different interviewers may elicit different answers because of differences in the respondents they recruit (e.g., some interviewers may be better than others at recruiting older respondents) but also because of differences in their levels of measurement error. As anyone who has ever listened to CARI recordings can testify, interviewers do not always stick to the script and their improvisations can sometimes elicit poor quality responses.

Pilot study. Having listened to recordings of field interviewers as part of the field test for a major national study, we designed an experiment to test the hypothesis that providing timely feedback to interviewers about their reading of the questions would improve the quality of the answers they elicited. (At the client's request, we do not divulge the name of the study.) This particular survey was a good test bed for assessing the effects of rapid feedback because the interviewers administered a short screening questionnaire to a household informant and then similar questions were administered to each sample

member via audio computer-assisted self-interviewing (ACASI). As a result, we could compare the screening data collected by each interviewer with a "gold standard" for several of the key items in the survey. Of course, the ACASI data are not error-free, but we regarded them as less error-prone than the screener data for two reasons: Each person reported for himself or herself whereas the screener was administered to a single household informant; and the questions were self-administered rather than administered by an interviewer and self-administration was likely to reduce any social desirability bias in the responses.

The experiment included 291 interviewers. Half were assigned to receive rapid feedback and half were assigned to the control group. Every fifth screener done by interviewers in the rapid feedback group was CARI-coded to identify departures from standardized interviewing. Figures 4.1 displays the questions coders answered for each screening interview. After a screener was coded, interviewers (and their supervisors) were sent a report with their performance and a link to the question recordings. For their first coded screener, interviewers were instructed to schedule a feedback session with a central office "mentor", who reviewed the results and provided guidance for improvement. For their second coded screener, interviewers were sent only the report and a link to the recordings. For subsequent screeners, interviewers were only instructed to schedule a feedback session with their mentor if the coding identified problems; otherwise, they were only sent the written report.

Figure 4.1 Coding questions for rapid feedback pilot study. The questions were repeated for each member of the household.

Q1. How clearly can you hear the interviewer on this recording? [HEARINT]
- Very clearly (4)
- Somewhat clearly (3)
- Not very clearly (2)
- Cannot hear the interviewer (1)

Q2. How clearly can you hear the respondent on this recording? [HEARRESP]
- Very clearly (4)
- Somewhat clearly (3)
- Not very clearly (2)
- Cannot hear the interviewer (1)

Q3. Did the interviewer read the question exactly as worded? [EXWORD]
- Yes (1)
- No (2)

Q4. [IF NO TO Q3] How did the interviewer change the wording of the question? Pick all that apply
- Did not read lead-in or introductory text before the question [NOINTRO]
- Did not read "Please look at this picture" [NOPIC]
- Did not read "Please look at this list" [NOLIST]
- Did not read all brand names or product examples [NONAMES]
- Did not read response options correctly [NORESP]
- Did not read "choose all that apply" [NOCHOOSE]
- Omitted, added, or changed other words within the question [NOREADOTH]

Q5. Did the interviewer correctly enter the respondent's answer? [ENTERANS]
- Yes (1)
- No (2)

The experiment was conducted from May to August, 2014, with 1,729 respondents interviewed by the feedback group and 1,717 interviewed by the control group.

To evaluate the effects of rapid feedback, we compared three variables derived from the screening items to the corresponding variables from the ACASI interviews. In principle, the two should match. Table 4.1 shows the proportion of respondents in the treatment and control groups who were classified the same way in the screener and the ACASI data. For all three, the match rate was significantly higher for respondents who were interviewed by interviewers getting feedback. (We used a Rao-Scott F test that took into account the clustering of the sample by areas. All three F -values were significant at $p < 0.01$.) Kappa values measuring the chance corrected agreement between screener and ACASI responses are substantially higher for interviewers in the rapid feedback group as well.

Table 4.1
Agreement between screener and ACASI responses, by condition and variable

	Rapid feedback	Control	Rao-Scott F value (1 and 230 df)
Composite			
% Agree	93%	88%	15.5***
Kappa	0.85	0.76	
Variable 1			
% Agree	95%	92%	8.8**
Kappa	0.89	0.83	
Variable 2			
% Agree	89%	85%	7.7**
Kappa	0.76	0.69	

** $p < 0.01$;

*** $p < 0.001$.

Note: The composite was a summary variable derived from variables 1 and 2.

MEPS study. Based on the success of this initial study, Edwards, Sun and Hubbard (2019) undertook a replication. In 2018, the Medical Expenditure Panel survey had implemented a major upgrade of the CAPI system and had simplified some sections of the questionnaire. Two question series were of particular interest because they were asked in all interviews, always recorded in CARI (almost all respondents gave consent to record), and were critical for producing data on the use and cost of health care services, key MEPS statistics. These were the questions on the use of calendars or other records of medical care during the interview and "provider probes", filter questions that prompt the respondent to recall services from various types of medical providers. The calendar series asked whether various records were available during the interview (e.g., a calendar with entries for medical visits, insurance statements, etc.), and who in the household was associated with each record type. The CAPI entry area for these items was a grid, with each household member listed on a row and each record type a column header. Interviewers could enter answers in any order, by person or by record type. The objective was to encourage respondents to bring records for all family members into the interview and to structure the questioning so that the records could be incorporated into the interview in any order. The provider probes

consisted of 15 questions about various kinds of health care providers. They were re-ordered in the technical upgrade to begin with three that accounted for the highest expenditures.

Audio-recordings of the calendar series and the provider probes series were reviewed by two behavior coders. The coding system allowed coders to call up specific interviewers or questions. Coders evaluated the overall quality of the interview and of each instance of asking the calendar series and the provider probes. The inter-coder agreement rate was 0.82. Verbal and written feedback was provided to the interviewer quickly (ideally within 72 hours of the interview). The next interview conducted by the interviewer was also coded, so that each interviewer had a pair of interviews in the data set, one just before and one just after feedback. Because the process was implemented in late fall only a subset (122) of the MEPS interviewers were available to participate in the study, resulting in 244 interviews in the data set. Data about the feedback interaction was also captured (such as whether the interviewer agreed with the feedback or asked for clarification). Again, we expected that interviewer behavior more consistent with the study protocol would be observed after feedback, both for overall interview quality and for each question series.

Table 4.2 shows the rapid feedback results for each question series. Interviewers maintained the meaning of the questions but did not follow the protocol exactly in the majority of instances $(n = 5{,}259)$, both before and after feedback. Still, question-asking behavior that followed the protocol exactly increased from 33.4 percent before feedback to 43.4 percent after feedback; failing to maintain the question meaning decreased from 9.8 percent before feedback to 3.7 percent after feedback. An F-test that took into account the clustering of the observations by interviewer found a significant overall difference between interviewer behavior before and after feedback, both overall $(F(2{,}118) = 3.86,\ p < 0.05)$ and for the provider probes $(F(2{,}118) = 5.71,\ p < 0.01)$. The differences for the calendar series was in the same direction but not statistically significant. These results, like those of the pilot study, indicate that rapid feedback to the interviewers can lead to marked improvements in how they administer the questions.

Table 4.2
Interviewer behaviors, before and after feedback, by question series

Interviewer Behavior	Calendar series		Provider probes		Both series	
	Before	After	Before	After	Before	After
Followed protocol exactly	18.6%	27.9%	43.3%	51.7%	33.4%	43.4%
Maintained meaning but did not follow protocol exactly	68.9%	65.1%	48.7%	46.4%	56.8%	52.9%
Did not maintain meaning	12.5%	7.0%	8.0%	2.0%	9.8%	3.7%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
n	1,240	759	1,832	1,428	3,072	2,187

5. Conclusions

The three main methods reviewed here have a mixed record of success. What lessons can we draw from these efforts to substitute data for intuition in the management of surveys?

The literature on responsive and adaptive design leads to several conclusions. First, it is important to clarify the statistical goals for the design at the outset of the survey and to monitor measures of quality related to these goals. Different strategies serve different goals. For example, equalizing response propensities may reduce nonresponse bias at the expense of a smaller sample size and increased sampling variance. It is essential to acknowledge such tradeoffs. Second, both the overall response rate and the variation in response propensities contribute to the average nonresponse bias. As a result, no single indicator gives a complete picture of the risk of error in a survey and survey managers should monitor multiple indicators, including changes in a set of key survey estimates. Advances in "dashboard" design (Mohadjer and Edwards, 2018) make it easier for central office staff and field supervisors to monitor a large number of indicators of how the field work is going. Third, simply continuing a given data collection protocol may not change the estimates much (Sturgis et al., 2017) and, in some cases, may decrease the representativeness of the sample (Lundquist and Särndal, 2013; Särndal and Lundquist, 2014). Under a given data collection protocol, the respondents recruited late in the field period are not likely to differ much from the ones recruited earlier. The sample will continue to overrepresent the cases with higher propensities under that protocol. To change the mix of respondents – and to improve the overall representativeness of the sample – may require major changes in the data collection protocol, such as much larger incentives, a switch to a different mode of data collection, or a much shorter questionnaire. These strategies all have their drawbacks, leading to the conclusion that sometimes the best strategy is just to cease further efforts by imposing stopping rules. Continuing to pursue cases with very low response propensities to respond is a formula for driving up costs without really improving the statistical properties of the final estimates.

Both the literature on responsive and adaptive designs and the study on case prioritization and optimal routing discussed in Section 3 above indicate that one factor limiting the effectiveness of central office interventions on field work is resistance by the interviewers. We need more research on how to improve interviewer compliance and on the impact of closer monitoring (or larger incentives) to ensure interviewers implement the desired changes in protocol. The studies on rapid feedback to the interviewers are encouraging in this regard. Both studies I reviewed in Section 4 indicate that when interviewers are given timely feedback on their administration of the questions they do a better job, and this reduces the level of measurement error in the answers they elicit.

One thing is certain. In an increasingly difficult climate for surveys, efforts to improve the management of surveys and to apply as much as science as possible in that endeavor will surely continue.

Acknowledgements

This paper was written in response to my receiving the Waksberg Award. I am grateful to Brad Edwards, Gonzalo Rivero and Tammy Cook for their helpful suggestions on this paper; to Aaron Maitland and Gonzalo Rivero for their help in designing and carrying out some of the studies described here; and to Statistics Canada for giving me the award.

References

Atrostic, B.K., Bates, N., Burt, G. and Silberstein, A. (2001). Nonresponse in U.S. government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, 17, 209-226.

Axinn, W.G., Link, C.E. and Groves, R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*, 48, 1127-1149.

Beaumont, J.-F., Bocci, C. and Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Brick, J.M., and Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752.

Brick, J.M., and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, 645, 36-59.

Calinescu, M., Bhulai, S. and Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, 226, 115-121.

Chun, A.Y., Heeringa, S.G. and Schouten, B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, 34, 581-597.

de Leeuw, E.D., and de Heer W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In *Survey Nonresponse*, (Eds., R.M. Groves, D. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc, 41-54.

Edwards, B., Sun, H. and Hubbard, R. (2019). Behavior change techniques for reducing interviewer contributions to total survey error. Unpublished manuscript.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 439-457.

Groves, R.M., and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72, 167-189.

Groves, R.M., Singer, E. and Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.

Groves, R.M., Benson, G., Mosher, W.D., Rosenbaum, J., Granda, P., Axinn, W., Lepkowski, J. and Chandra, A. (2005). *Plan and Operation of Cycle 6 of the National Survey of Family Growth*. Hyattsville: National Center for Health Statistics.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Hicks, W.D., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L.D. and Moss, A.J. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, 74, 985-1003.

Hubbard, R. (2018). That wasn't part of the plan! Reducing effort through stopping rules to place CAPI cases on hold and work plans to set them free. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Denver, Colorado, May 17, 2018.

Laflamme, F., and Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada. Paper presented at the European Quality Conference, Helsinki, Finland.

Laflamme, F., and St-Jean, H. (2011). Highlights and lessons from the first two pilots of responsive collection design for CATI surveys. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 1617-1628.

Lepkowski, J.M., Mosher, W.D., Groves, R.M., West, B.T., Wagner, J. and Gu, H. (2013). *Responsive Design, Weighting, and Variance Estimation in the 2006–2010 National Survey of Family Growth*. Vital and Health Statistics, Series 2, No. 158. Hyattsville, MD: National Center for Health Statistics.

Luiten, A., and Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society A*, 176, 169-189.

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, 557-582.

Mohadjer, L., and Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education*, 26, 263-277.

Peytchev, A., Baxter, R.K., and Carley-Baxter, L.R. (2009). Not all survey effort is equal: Reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, 73, 785-806.

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.

Särndal, C.-E. (2011). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.-E., and Lundquist, P. (2014a). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société de Statistique*, 155, 28-50.

Särndal, C.-E., and Lundquist, P. (2014b). Accuracy in estimation with nonresponse: A function of the degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2, 361-387.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). Interviews and interviewing. In *Handbook of Survey Research*, Bingley, UK, Emerald Group Publishing, 437-470.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*, Boca Raton, FL: CRC Press.

Steeh, C., Kirgis, N., Cannon, B. and DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *Journal of Official Statistics*, 17, 227-247.

Sturgis, P., Williams, J., Brunton-Smith, I. and Moore, J. (2017). Fieldwork effort, response rate, and the distribution of survey outcomes: A multilevel meta-analysis. *Public Opinion Quarterly*, 81, 523-542.

Tourangeau, R. (2017). Presidential address: Paradoxes of nonresponse. *Public Opinion Quarterly*, 81, 803-814.

Tourangeau, R., Kreuter, F. and Eckman, S. (2012). Motivated underreporting in screening surveys. *Public Opinion Quarterly*, 76, 453-469.

Tourangeau, R., Brick, J.M., Lohr, S. and Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society A*, 180, 203-223.

Tourangeau, R., Yan, T., Sun, H., Hyland, A. and Stanton, C.A. (2019). Population Assessment of Tobacco and Health (PATH) reliability and validity study: Selected reliability and validity estimates. *Tobacco Control*, 28, 663-668.

U.S. Census Bureau (2014). *American Community Survey Design and Methodology*. Downloaded February 19, 2016 from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7, 45-55.

Wagner, J., and Olson, K. (2018). An analysis of interviewer travel and field outcomes in two field surveys. *Journal of Official Statistics*, 34, 211-237.

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G. and Kruger Ndiaye, S. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28, 477-499.

Walejko, G., and Miller, P. (2015). The 2013 census test: Piloting methods to reduce 2020 Census costs. *Survey Practice*, 8.

West, B.T., and Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 175-211.

West, B.T., and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 1004-1026.

West, B.T., Kreuter, F. and Jaenichen, U. (2013). Interviewer effects in face-to-face surveys: A function of sampling, measurement error or nonresponse? *Journal of Official Statistics*, 29, 277-297.

West, B.T., Conrad, F.G., Kreuter, F. and Mittereder, F. (2018). Nonresponse and measurement error variance among interviewers in standardized and conversational interviewing. *Journal of Survey Statistics and Methodology*, 6, 335-359.

Williams, D., and Brick, J.M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 6, 186-211.