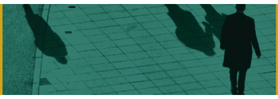
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)
180, *Part* 1, *pp.* 203–223

Adaptive and responsive survey designs: a review and assessment

Roger Tourangeau, J. Michael Brick, Sharon Lohr and Jane Li
Westat, Rockville, USA

[Received February 2015. Final revision December 2015]

Summary. The paper reviews the growing literature on responsive and adaptive designs for surveys. These designs encompass various methods for managing data collection, including front loading potentially difficult cases, tailoring data collection strategies to different subgroups, prioritizing effort according to estimated response propensities, imposing stop rules for ending data collection, monitoring key survey estimates throughout the field period, using two-phase or multiphase sampling for following up non-respondents and calculating indicators of non-response bias (such as the *R*-indicator) other than response rates to monitor and guide fieldwork. We give particular attention to efforts to evaluate these strategies experimentally or via simulations. Although the field seems to have embraced these new tools, most of the evaluation studies suggest they produce marginal reductions in cost and non-response bias. It is clearly difficult to lower survey costs without reducing some aspect of survey quality. Other issues limiting the effectiveness of these designs include weakly predictive auxiliary variables, ineffective interventions and slippage in the implementation of interventions in the field. These problems are not, however, unique to responsive or adaptive design. We give recommendations for improving such designs and for improving the management of data collection efforts in the current difficult environment for surveys.

Keywords: Adaptive design; Balance indicators; Non-response bias; Responsive design; *R*-indicator; Two-phase sampling

1. Introduction

Beginning with Groves and Heeringa (2006), many researchers have tried to implement 'responsive' and 'adaptive' designs for surveys. Groves and Heeringa (2006) envisioned a survey done in several phases (pages 440–441):

> '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".'

This paper reviews the growing literature on such designs and focuses on attempts to evaluate them.

Groves and Heeringa pointed to several examples of surveys using responsive designs, including the Chicago Mind and Body Study (Morenoff *et al.*, 2007), the National Comorbidity Study—Replication (Kessler *et al.*, 2004) and cycle 6 of the National Survey of Family Growth

Address for correspondence: Roger Tourangeau, Methods Unit, Westat, 1600 Research Boulevard, Rockville, MD 20850, USA.
E-mail: RogerTourangeau@westat.com

0964–1998/17/180203

(NSFG) (Groves *et al.*, 2005). Two of these included experiments done during an initial phase of data collection to inform key design decisions for later phases. For example, the National Comorbidity Study compared selecting one *versus* two household members from sample households in the first phase of data collection and, on the basis of the results, selected only one member from each sample household during the remainder of the field period. Two of the studies used classical two-phase sampling during the later stages of data collection to select cases that would receive additional interview attempts. In the Chicago Mind and Body Study, half of the active cases with low estimated response propensities and all the active cases with high estimated response propensities were retained for this second phase of data collection and the incentive for these cases was increased from $60 to $100. Finally, during a third phase, all remaining Chicago Mind and Body Study cases were slated for a final contact and offered $150 to complete the survey. Similarly, at the end of phase 2 of the NSFG, about a third of the cases were retained for further fieldwork. The final phase used the most productive interviewers from the prior phases, relaxed the rules for collecting screening information from proxy informants, used a prepaid incentive of $5 for the screener and also increased the incentive amount for the main interview (offering an additional $40 over the amount in phase 2).

Of course, experiments embedded in surveys are hardly new; nor is subsampling non-respondents to reduce the potential effects of non-response on the final survey estimates. Hansen and Hurwitz (1946) discussed the use of two-phase sampling for non-response in 1946. In fact, an earlier cycle of the NSFG, done in 1988, had also subsampled among the remaining non-responding cases in an attempt to reduce the effect of non-response on the final estimates (Judkins *et al.*, 1991). What make the studies that Groves and Heeringa (2006) cited examples of *responsive* design are several features. First of all, the experiments in the National Comorbidity Study and the NSFG were done as part of the main study, not as part of an earlier pretest or pilot study. The data from all these cases were included in the final data sets and used in making the final estimates from the survey. Second, the evaluation of the experiments utilized paradata (data collected about the process of data collection). For example, cost information (or proxies for cost, such as the number of interviewer hours) was used in evaluating National Comorbidity Study experimental results. Similarly, the subsampling of non-respondents in the Chicago Mind and Body Study and cycle 6 of the NSFG was based on propensity models that were run in realtime and that incorporated paradata (such as information gleaned from prior contacts with the sample case). More generally, Groves and Heeringa (2006) (see also Groves *et al.* (2005), Axinn *et al.* (2011) and Wagner *et al.* (2012)) defined responsive survey designs as

(a) identifying a set of design features that might affect survey costs and errors,
(b) identifying indicators of the cost and error properties of those features,
(c) monitoring those indicators during the initial phases of data collection,
(d) changing design features (as needed) on the basis of those indicators and the cost–error trade-offs that they imply and
(e) combining the data from the various phases into a final data set.

The changes in design features (the fourth point in their list) are what distinguish the different phases of data collection.

Later researchers have attempted to extend this basic framework in two directions. One approach has been to try to tailor the data collection from the outset, on the basis of frame or other data that are available on the sample units, with different subgroups assigned to different data collection protocols (see, for example, Calinescu *et al.* (2013) and Schouten *et al.* (2013)). If a survey has been conducted a number of times previously, it may be possible to determine different optimal designs for different subgroups on the basis of this past experience. Schouten and his colleagues labelled this approach ‘adaptive design’ (see also Couper and Wagner (2011) and Wagner (2013)). This approach contrasts with modifying the design after the survey has been fielded, with the modifications based on the early phases of the fieldwork. The other approach has been to discard the notion of discrete phases, making mid-course corrections and other adjustments throughout the field period, which is an approach that is sometimes labelled dynamic design.

Our review of efforts to implement responsive or adaptive designs for surveys begins by examining related approaches in other fields, such as adaptive sampling and adaptive treatment protocols. Next, we discuss attempts to use and evaluate responsive designs. There are several possible statistical goals that such survey designs might have (see Godbout *et al.* (2011) and Schouten *et al.* (2013)), such as minimizing data collection costs, minimizing variance or minimizing non-response bias. The different goals have different implications for managing fieldwork. A study by Peytchev and his colleagues, for example, attempted to use a type of adaptive design—case prioritization—to minimize non-response bias (Peytchev *et al.*, 2010). As Peytchev and his co-authors noted, adaptive designs require that the survey managers can identify some problem during data collection (such as underrepresentation of some subgroup of sample cases) and design some intervention to solve that problem. Various studies have attempted to evaluate the effectiveness of such survey designs, including that by Peytchev *et al.* (2010); we summarize these studies in Section 3. These empirical results are buttressed by several simulation studies, which are reviewed in Section 4.

2. Precedents from other fields

Adaptive designs have a long history in survey sampling and in the design of experiments. In this section, we describe some of the previous research on adaptive designs and discuss the potential relevance of this work for surveys. In all of the work that is discussed below,

(a) the design for later phases of the study is modified as a function of data collected in earlier phases of the study and
(b) the data collected in all phases of the study are combined to construct estimates of the quantities of interest.

These are also key features of responsive survey design.

2.1. *Adaptive design in survey sampling*

Two-phase sampling for non-response (Hansen and Hurwitz, 1946) is a form of adaptive design, where phase 1 consists of the initial contact attempts and phase 2 consists of subsampling the non-respondents from phase 1. Information that is collected about non-respondents at phase 1 may be used to stratify the phase 2 sample. If all the units that are subsampled at phase 2 respond to the survey, then this procedure eliminates the non-response bias. In the Hansen–Hurwitz type of adaptive sampling design, the selection probabilities of units in later phases depend on auxiliary information, but not directly on the values of the responses that have been observed in earlier phases. The notion of two (or more) phases is central to Groves and Heeringa’s (2006) definition of responsive design and they distinguished it from earlier efforts by the use of key indicators of data quality to that point in time in the decision to launch a new phase.

There are other precedents to adaptive and responsive design in the survey sampling literature. Adaptive and responsive designs often attempt to achieve ‘balanced’ or representative samples during data collection; balanced samples match the population distribution on one or more dimensions, such as sex or age (see Särndal (2011) and Särndal and Lundquist (2014a)). One early strategy for achieving balance during data collection was to use 'block quota' samples. In such samples, the first- and second-stage sampling units were selected by conventional probability sampling, but individual respondents were recruited by interviewers within the second-stage sample areas (typically, individual blocks or groups of adjacent blocks) to fill fixed quotas (e.g. a certain number of men under 30 years old and a certain number 30 years or older) within each sample area. These samples imposed balance on the respondents by region, level of urbanization, sex and age for men and employment status for women (see, for example, Sudman (1966) and Stephenson (1979)). Stephenson (1979) reported an experiment showing that the block quota sample gave results that are quite similar to those from a full probability design. More recently, Rivers (2006) has advocated sample matching for volunteer on-line panels; again, the goal is to achieve a sample that matches the population on its demographic make-up. And Valliant *et al.* (2000) advocated a form of probability sampling in which the sample that is selected matches the population mean (and variance) on one of more key auxiliary variables. Like quota samples, adaptive and responsive designs often attempt to achieve balance within the *responding* sample (e.g. Särndal and Lundquist (2014a); see also Schouten *et al.* (2014)).

2.2. Adaptive design in sequential analysis and clinical trials

Adaptive designs are often recommended for clinical trials because they may allow conclusions to be drawn by using fewer patients or a shorter study duration (Food and Drug Administration, 2010). Several types of adaptive design have been studied. In clinical trial work, adaptive designs may use continuous monitoring or they may update the design in phases.

The motivation behind using adaptive designs is the uncertainty about the optimal design when planning the trial. The size of the treatment effect is unknown in advance, so the initial design of a study may specify a sample size that is larger than needed to detect a significant difference. In other studies, the aim is to give the best treatment and dosage to as many patients as possible in the trial, but that 'best treatment' is unknown at the beginning of the study. Adaptive clinical trials take advantage of accumulating information to modify the design to achieve study goals at lower costs. A key to being able to analyse the results from the study is that the possible adaptations must be defined before the study begins.

Several types of adaptive design have been proposed in the sequential analysis and adaptive design literature.

(a) *Group sequential designs for early termination of study*: in clinical trials comparing a new drug with a control treatment, adaptive design allows the study to be terminated early if the new drug shows a strong effect. Jennison and Turnbull (2010) summarized methods that may be used in group sequential clinical trials.
(b) *Design modifications to obtain a desired power or predetermined precision*: these include the two-phase and three-phase methods that were described by Cox (1952), Hall (1981) and Lohr (1990), in which the estimated variance of the responses after each phase is used to calculate the remaining sample size that is needed to obtain a final estimate of fixed precision.
(c) *Adaptive assignment of future experimental units to treatments based on results to date*: Chow and Chang (2011) summarized methods for adaptive randomization of future units in an experiment. These methods attempt to assign patients sequentially to the treatment method that thus far is performing better in the trial. Alternative methods make adjustments to the proportion of patients who are allocated to each treatment in phases.

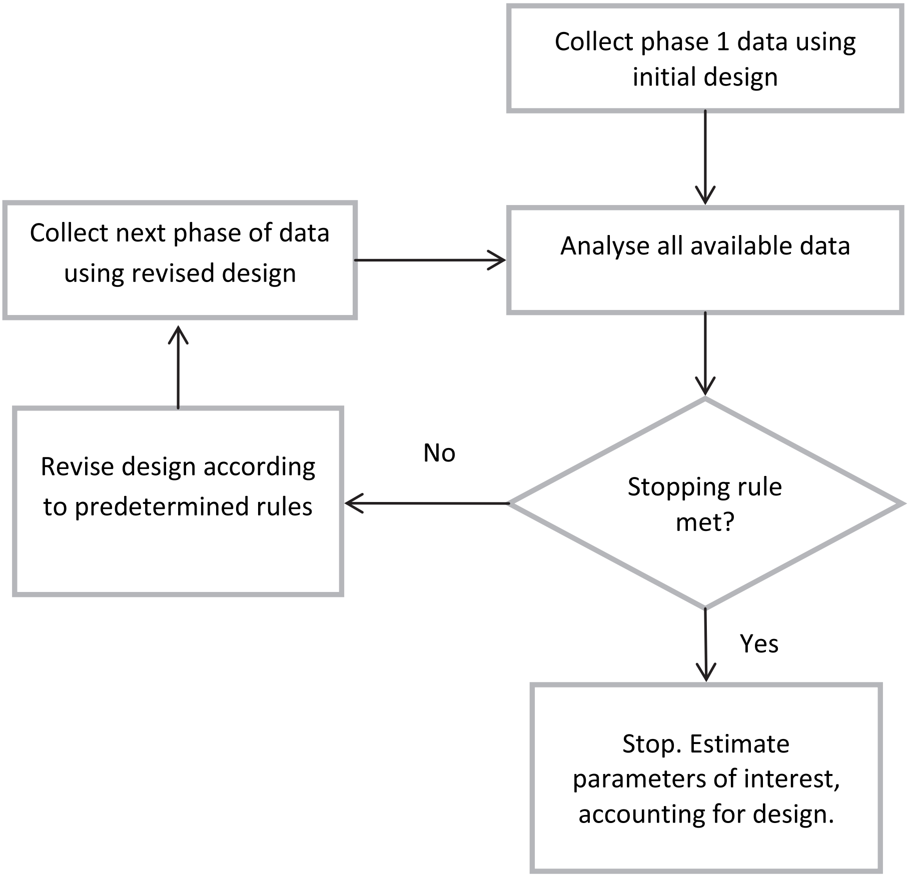

Collect phase 1 data using initial design
Analyse all available data
Stopping rule met?
No
Revise design according to predetermined rules
Collect next phase of data using revised design
Yes
Stop. Estimate parameters of interest, accounting for design.

Fig. 1. Structure of a phased adaptive clinical trial

(d) *Enrichment designs* (Wang *et al*., 2009) are designs in which the inclusion or exclusion criteria may be modified so that the study concentrates on populations that are most likely to benefit from the treatment.

The sequential methods that are used in clinical trials, thus, come in a variety of forms, but all fit within the basic structure that is shown in Fig. 1. A similar figure could be used to characterize responsive survey designs of the type described by Groves and Heeringa (2006).

The adaptive design literature in clinical trials emphasizes concern about bias that may result from the adaptive procedure itself. Group sequential stopping rules must be designed carefully to avoid bias. Korn *et al*. (2001) discussed possible biases that can result from enrichment designs. In the responsive and adaptive survey design work, the adaptations are made to increase response rates, to reduce non-response bias or to reduce cost, but little attention has been paid to the possibility that responsive design itself may introduce bias. Although a responsive design collects data in phases, using information collected in previous phases to change the design for the next phase, it does not necessarily follow the template in Fig. 1. In particular, a responsive design implementation might not necessarily have well-defined stopping rules or rules for revising the design as a function of the data observed to date.

An example shows how bias can arise from using the sample to date to determine the ultimate sample size for a study (see Lohr (1990)). In the Cox (1952) two-phase procedure for obtaining a fixed accuracy, a pilot sample is taken in phase 1. The variance is estimated from that pilot sample and used to calculate the size needed for the remaining sample to achieve a fixed width confidence interval. But the sample variance, calculated from the combined samples, will be biased downwards. This occurs because of the sequential nature of the procedure. If the sample variance from the pilot sample is too large, then the procedure will specify taking a large sample in phase 2, and the overestimate of the variance from the first stage will be corrected. If the sample variance is too small, however, then the sample size will be small and there will not be as many observations taken to correct the underestimate from phase 1.

The same sort of potential for bias exists with responsive designs. If non-response follow-up efforts are allocated on the basis of estimated precision or bias at early phases, those estimates have sampling error that may affect the design decisions. For example, the mean of the responses to date from a particular population subgroup might be thought to have low bias on the basis of the early data collection. This might then lead to fewer follow-up attempts in that subgroup, which would not give enough additional sample to overturn the initial judgement of low bias if in fact that judgement was faulty. Similarly, early indicators of sample representativeness may be subject to large sampling errors, leading to misallocation of field effort. Wagner and Hubbard (2014) discussed a related problem for responsive designs that rely on propensity models; they noted that models estimated early in the field period may produce biased estimates of the eventual response propensities.

3. Field evaluations of responsive or adaptive designs

In this section, we examine attempts to implement responsive or adaptive designs in surveys and then to evaluate their effectiveness. We have uncovered eight such efforts; in addition, we found three other studies that used simulations to explore the effects of implementing an adaptive design. Section 4 summarizes the results of the simulation studies.

Before we turn to the specific studies, it is worth pointing out that responsive and adaptive designs are a response to an increasingly difficult survey climate characterized by declining budgets, rising costs and falling response rates, both in the USA and elsewhere (Massey and Tourangeau, 2013). These designs, thus, represent an attempt to do more with less or at least to do as much as possible with less. More concretely, responsive or adaptive designs might try to achieve one or more of three statistical goals while remaining within ever-tighter survey budgets (see Beaumont *et al.* (2014)):

(a) minimizing variance,
(b) minimizing non-response bias or some proxy for it, such as sample imbalance (Särndal (2011); see also Schouten *et al.* (2009)) or
(c) maximizing response rates.

In addition, some researchers have begun to look at the use of such designs to reduce measurement errors (Calinescu *et al.*, 2013). Although some have criticized ‘blind’ maximization of survey response rates, we note that in many cases maximizing the overall response rate is equivalent to maximizing the final sample size, which in turn may be seen as essential to achieving the survey’s precision goals. The focus on response rates in the literature may obscure the importance of other closely related goals, such as the goal of hitting sample size or precision targets.

To date, studies have used three basic strategies to achieve one or more of these goals—multiphase designs (like that described by Groves and Heeringa (2006)), case prioritization (in which different cases are slated to receive different levels of effort) and tailoring of the fieldwork or mode of data collection (based on what is known about each case). We examine how successful each strategy has been.

3.1. Multiphase designs

There have been at least two major attempts to evaluate the effectiveness of multiphase designs: one based on cycle 6 of the NSFG and the other based on the National Intimate Partner and Sexual Violence Survey pilot. In addition, Chapman (2014) described two other studies using multiphase responsive designs but did not present evidence regarding their effectiveness.

3.1.1. Cycle 6 of the National Survey of Family Growth

Cycle 6 of the NSFG used three phases (Axinn *et al.*, 2011; Groves *et al.*, 2005; Groves and Heeringa, 2006). The first phase of cycle 6 of the NSFG lasted 4 months, the second, 7 months, and the final phase, 1 month. The final phase was 'designed explicitly to use analyses of paradata to ... improve representation of reluctant respondents' (Axinn *et al.* (2011), page 1128).

In all phases, the data were collected face to face. The fieldwork in phase 1 (which involved a subsample of the final sample) led to the decision to cap call attempts at 14 during the second phase; the phase 1 data indicated that, by 14 call attempts, the estimates were unlikely to change sufficiently to justify the added costs of further callbacks (Groves *et al.*, 2005). During phase 2, two propensity models were developed; these were discrete hazard models based on frame variables and variables derived from interviewer observations and other paradata. One model predicted the likelihood that a household that had not yet been screened would be screened on the next call; the other model predicted the likelihood that a sample case would complete a main interview on the next call. These models were used to group second-stage units, or segments, into those with low, medium and high overall estimated propensities. The grouping was intended to help field supervisors to direct the fieldwork. At the end of the second phase, a subsample of segments was selected for retention in phase 3. Groves *et al.* (2005) grouped segments into 16 strata, based on the number of active cases in the segment at the end of phase 2 and the total estimated propensities for those cases. Segments in the highest quartiles on each of these variables were sampled at higher rates than those in the lower quartiles (see Table Q in Groves *et al.* (2005)).

It is not clear exactly what statistical goal this sampling strategy was designed to achieve. Axinn *et al.* (2011) claimed that it was intended to improve the representation of 'reluctant respondents'. Groves *et al.* (2005) said that phase 3 was designed to produce a 'more representative' sample (page 38) by altering the data collection protocol in ways that might appeal to sample members who had failed to respond to the phase 2 protocol. Focusing data collection resources on segments with larger numbers of active cases seems likely to reduce the phase 3 data collection costs; Groves *et al.* (2005) noted that

> 'this design option placed large emphasis on the cost efficiency of the Phase 3 design to produce interviews, not on minimizing standard errors of the resulting data set'.

However, targeting areas with more cases with relatively high estimated phase 2 propensities might actually worsen any problems with representativeness by bringing in additional respondents who were similar to those who had already responded in phase 2. Still, the changes in the data collection protocol might have offset any effects from the targeting that is implicit in the subsampling. These changes included offering a prepaid incentive to households that had not yet completed a screener, allowing proxy reporters at the screening phase, boosting the incentive for a main interview from \$40 after completion to \$80 (\$40 of it prepaid) and retaining only the most successful phase 2 interviewers for phase 3.

Did the phase 3 strategy work? The final phase of data collection did raise the overall response rate to the survey from about 64% at the end of phase 2 to nearly 80% at the end of phase 3. However, Groves *et al.* (2005) did not report much gain in efficiency from the targeting of the sample in the final phase. For example, they looked at the number of call-backs per completed screener and did not find clear differences in this measure of efficiency across the segment level strata. They also examined the variation in response rates for 18 key subgroups of the sample; the coefficient of variation in the response rates for these 18 groups dropped from 7.6% at the end of phase 2 to 4.4% at the end of phase 3. In addition, Axinn *et al.* (2011) showed that the final phase also changed the demographic make-up of the NSFG sample, for example, bringing in a higher proportion of married respondents in the final phase than in the earlier phases. Finally, Axinn *et al.* (2011) also showed that the cases that were added in the final phase produced changes in the coefficients in multivariate models examining, for example, the probability of having relatively high numbers of sexual partners. Despite these signs that the final phase produced a more representative sample than the earlier phases, it is not clear how large the improvement was relative to, say, simply continuing phase 2.

3.1.2. National Intimate Partner and Sexual Violence Survey pilot

Peytchev *et al.* (2009) reported a telephone study that implemented a two-phase design, in which the second phase featured a shorter questionnaire and a larger incentive than the first. The initial sample was a list-assisted random-digit dialling sample supplemented by a sample of listed telephone numbers. Telephone numbers that could be matched to an address were sent an advance letter with a \$1 incentive and all cases were promised either \$10 or \$20. Apart from the amount of the incentive, the first phase also experimentally varied how the survey topic was described to potential respondents (‘crime’ *versus* ‘health’ *versus* ‘personal relationships’). Most cases received up to 20 calls during the first phase, but some received even more. In addition, refusal conversion was attempted during phase 1. Overall, this phase produced a response rate of 28.5%. In phase 2, the researchers selected a subsample of the remaining non-respondents, reduced the length of the questionnaire from 30 to 14 min, used a prepaid incentive of \$5 and offered a conditional incentive of \$20. This phase produced a response rate of 9.8% (or 35.5% overall).

Peytchev *et al.* (2009) compared the early and late respondents from phase 1 and the phase 1 and phase 2 respondents. They argued that the late respondents (who were interviewed after at least six call attempts) from phase 1 were likely to be similar on the key study variables to the early respondents (who were interviewed in five or fewer attempts) because they were recruited via the same protocol (and, in fact, the addition of cases who were interviewed after five attempts did not significantly change the estimates).

In contrast, Peytchev *et al.* (2009) believed that the phase 2 respondents were likely to differ from the phase 1 respondents, because the changes in protocol would attract different types of respondent. There was some support for this line of argument for the male respondents; the phase 1 male respondents were more likely to report victimizations than the phase 2 male respondents, with significant differences on four of six estimated 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’.

3.2. Case prioritization

Aside from the major changes in the data collection protocol like those used in the final phases of cycle 6 of the NSFG and in the National Intimate Partner and Sexual Violence Survey pilot, researchers have also investigated ‘case prioritization’ as an adaptive design strategy. In these studies, the goal has generally been to target low propensity cases in an effort to reduce the overall variation in response propensities. In the limit, if the actual underlying response propensities do not vary at all, the non-response bias will be 0 for means and proportions.

For example, Peytchev *et al.* (2010) reported a study in which the data collection protocol for different groups of cases was tailored from the outset. Their study involved a panel survey so that response propensities for each sample case could be estimated on the basis of information from the prior wave of data collection. Cases with low predicted response propensities were randomly assigned to an experimental or control treatment. For most of the data collection period, interviewers received 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 one phase, 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% for the control cases *versus* 90.8% for the experimental cases) or in the 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 non-response biases were *higher*. (As one of the reviewers of this paper noted, the non-response bias estimates in this study were based on the correlations between the survey variables and the fitted response propensities, which is an approach that relies on several assumptions.) Thus, interviewer incentives do not appear to have worked in lowering non-response bias.

Rosen *et al.* (2014) also reported a study that targeted low propensity cases in an effort to reduce non-response bias. Their study involved parents of students who were selected for the High School Longitudinal Study of 2009. Data were to be collected from the parents as well as the students, and Rosen and his colleagues fitted models predicting the response propensities of the parents. After an initial period of data collection by telephone and the Web, the remaining parents in the lowest propensity quartile were pursued by field staff for a face-to-face interview; data were still sought for parents in the other three propensity quartiles via telephone and the Web. This targeting of difficult cases boosted the response rate in the lowest propensity quartile and appeared to reduce the average relative bias (on five frame variables) from 7.6% to 7.2%.

Wagner *et al.* (2012) (see also Couper and Wagner (2011) and Lepkowski *et al.* (2013)) carried out a series of experiments involving case prioritization in cycle 7 of the NSFG. Cycle 7 consisted of 20 quarterly cross-sectional samples. Each quarterly sample was fielded in a two-phase design, with phase 1 lasting 10 weeks, and phase 2 two additional weeks. 16 randomized experiments were carried out over the course of cycle 7 to assess 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). The targeting varied across the experiments. Sometimes the goal was to improve response rates; sometimes it was to evaluate data that were available on the sampling frame; and sometimes it was to bring the characteristics of the respondent sample more closely in line with those of the original full sample (see Table 1 in Wagner *et al.* (2012)). In seven of the 16 experiments, the priority cases received significantly more calls than the control cases, but this led to a significant increase in response rates only twice in the 16 attempts. Additional experiments by Wagner and his colleagues attempted to shift the effort of NSFG interviewers from trying to complete main interviews to trying to achieve more screening. In each quarter, one week (the fifth) was designated as ‘screener week’; interviewers were instructed to give priority to screening households rather than completing main interviews during that week. This intervention seemed to lead to more screener calls than in prior or later weeks, but the effect on the number of *completed* screenings varied markedly across quarters. In both cases, the efforts to prioritize cases clearly affected interviewer behaviours but had less effect on the relevant survey outcomes, such as response rates.

Statistics Canada has begun to implement adaptive designs for its computer-assisted telephone interviewing surveys, partly in response to a new rule imposing a cap on the total number of calls to sample cases (Laflamme and Karaganis, 2010; Laflamme and St-Jean, 2011). Two computer-assisted telephone interviewing surveys each used three phases of data collection. During the first phase of data collection, cases were categorized by response propensities. During the second phase of data collection, cases were randomly assigned to the responsive collection or control conditions. The responsive collection cases were assigned priorities and high priority cases were slated to receive more calls; the control cases were not prioritized. The priority cases in the responsive collection group appeared to be cases that were predicted to have high response propensities. During the final phase of data collection, all remaining cases received the same treatment. During this phase, the goal was to equalize actual response propensities across key subgroups. Overall response rates were essentially unaffected by the case prioritization in the responsive collection group. In one survey, the response rates were 74.0% for the control group and 74.1% for the responsive collection group. In the second survey, the control group had a slightly higher response rate (73.0% *versus* 72.8%). The results were similar during the second phase of each survey, when the treatment of the two groups of cases differed. There is not much evidence that the new three-phase design or the responsive collection protocol increased the representativeness of the samples, but it may have decreased the number of interviewer hours (see Table 2 in Laflamme and St-Jean (2011)). Reducing costs without reducing representativeness clearly represents an advance.

Case prioritization may be more effective if the targeting is based on the likelihood that a case will affect the final estimates rather than based solely on estimated response propensities. Chapman (2014) reported an attempt to target ‘high value’ cases in a survey without discussing the exact measure of value used. It seems important to have a clearer notion of what cases are the most worth pursuing. We recommend this definition for the expected value of a case, V_i:

$$V_i = \hat{\rho}_i W_i \Delta_i.$$

The expected value of a case reflects its estimated response propensity $\hat{\rho}_i$, its design weight W_i (or the inverse of its selection probability) and its effect on the sample balance, Δ_i. The most valuable case will produce the largest expected reduction in the sample bias (as measured by the sample imbalance, or the distance of the sample means based on the current respondents from the corresponding population benchmarks).

A related approach is based on the distance of a given case from the current sample means, D_i:

$$V_i = \rho_i W_i D_i, \qquad D_i = \surd\{(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})\mathbf{S}^{-1}(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})\}.$$

More specifically, the value of the case reflects the Mahalanobis distance of its predicted values from the current sample means on a set of key survey variables (the *y*s). If it is not feasible to predict a case's values on the survey variables, D_i might reflect the case's distance from the sample means on a set of auxiliary variables (xs) that are known for both respondents and non-respondents (this is quite close to the balance measure that Särndal has advocated; see Särndal (2011)). If the covariance terms can be ignored, then the distance is just the sum across the ys or the xs of the standardized values for the case. It may also be useful to take into account the likely cost of obtaining data from the case, c_i—i.e. using $V_i/c_{i.}$. Thus, we recommend that, if possible, fieldwork should target the cases with the highest cost–benefit ratio, using one of the measures of value that are presented here. The benefit that is associated with a case should reflect its likely effect on the estimates.

3.3. Other approaches

3.3.1. Adaptive contact strategies

Although many researchers have examined optimal times for contacting sample members in telephone and face-to-face surveys (e.g. Greenberg and Stokes (1990) and Weeks *et al.* (1987)), relatively few have demonstrated the effectiveness of these 'optimal' call schedules empirically. Wagner (2013) presented the results from five experiments that used models to predict whether a given sample household would be contacted on the next call attempt for each of four call 'windows' (e.g. Tuesday–Thursday from 4 p.m. to 9 p.m.). Similar models were fitted for a telephone survey (Michigan's Survey of Consumer Attitudes (SCA)) and a face-to-face survey (cycle 7 of the NSFG). For the SCA, the model was based on census data for the zip code tabulation area linked to the phone number and on call history data; for the NSFG, the models also included variables based on interviewer observations about the sample dwelling. These models were used to identify the call window with the highest probability of a contact. 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 call time (in the NSFG).

In the first experiment on the SCA, although the proportion of calls resulting in a contact did increase for the experimental cases (to 12.0% *versus* 9.9% for the control cases), the contact rate was lower for experimental cases who had initially refused than for initial refusals in the control group. A later 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 appeared to ignore the recommended call windows; only 23.6% of the experimental cases were contacted in the recommended window (*versus* 23.0% in the control group).

3.3.2. Tailored fieldwork

Luiten and Schouten (2013) reported an experiment in which they tailored the data collection approach to different sample subgroups in the Dutch Survey of Consumer Sentiments. This survey consists of repeated cross-sections and, on the basis of earlier rounds of the survey, the researchers fitted contact and co-operation propensity models based on demographic characteristics of the sample members; the demographic variables were available from the population registry. The goal was to equalize response propensities across the subgroups.

In the experiment, there were two phases of data collection. In the initial phase, cases with low estimated co-operation propensities were sent a mail questionnaire; those with the highest estimated propensities were invited to complete a Web survey; and those with intermediate estimated propensities were given a choice between mail and Web. In the second phase, non-

Table 1. Contact and co-operation rates, by propensity quartile groups†

	Rates (%)	
	Experimental	*Control*
Contact propensity quartile		
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
Co-operation propensity quartile		
Lowest co-operation propensity	65.1	62.7
Second-lowest co-operation propensity	71.4	68.4
Second-highest co-operation propensity	72.8	75.3
Highest co-operation propensity	74.7	79.2

†Data from Luiten and Schouten (2013).

respondents were followed up by telephone. Cases in different contact propensity quartiles were assigned to different call schedules. Those in the quartile with the highest estimated contact propensities were fielded later and were called during the day; those in the second-highest quartile were called twice at night and then switched to a schedule that alternated daytime and night-time calls; those in the lowest two quartiles were called on every shift of every day. Finally, the best telephone interviewers were assigned to the cases with the lowest estimated co-operation propensities and the worst telephone interviewers were assigned to the cases with the highest estimated co-operation propensities. The control group for the experiment was the regular Survey of Consumer Sentiments sample, which is a computer-assisted telephone interviewing only survey.

Overall, the experimental, adaptive fieldwork group had a slightly higher response rate than the regular Survey of Consumer Sentiments cases (63.8% *versus* 62.8%: a non-significant difference). The representativeness of the experimental sample was significantly higher than that of the control sample (R-indicators of 0.85 and 0.77 respectively). Table 1 shows that the adaptive fieldwork did lower variability in both contact and co-operation rates. Across the four contact propensity quartiles, contact rates ranged from 84.2% to 96.9% in the control group; in the experimental group, the comparable figures were 87.1% and 95.3%. The adaptive design also lowered variation in the co-operation rates—as well as the overall co-operation rate (see Table 1). Still, the costs for the adaptive design were ‘marginally higher (2.6%) than those of the SCS’ (Luiten and Schouten (2013), page 185), and, although this cost difference was not statistically significant, the overall co-operation rate was significantly lower in the experimental sample.

3.4. Summarizing remarks

The studies that are reviewed in this section fall into two groups. Three of the studies (Groves and Heeringa (2006), Peytchev *et al.* (2009) and Rosen *et al.* (2014); see also a recent study by Roberts *et al.* (2014)) were non-experimental. They examined the effect of a change in protocol during the final phase of data collection on the composition of the final sample. Five others reported experiments (including Wagner *et al.* (2012), which summarizes multiple experiments).

The non-experimental studies compare the estimates that would have been obtained before the final phase of data collection with those obtained after the final phase, during which a major change in the data collection protocol was introduced. 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%, also decreased variation in the response rates across subgroups. In addition, Axinn *et al.* (2011) showed that the final phase of data collection changed the demographic make-up of the NSFG sample and produced changes in the coefficients of multivariate models for key study variables. Similarly, Peytchev *et al.* (2009) demonstrated that a major change in protocol (involving larger incentives and a shorter questionnaire) produced changes in the study estimates, at least for males. However, the changes were generally modest (less than 2 percentage points). Finally, in the study of Rosen *et al.* (2014), the switch to a much costlier method of data collection (face-to-face interviewing) produced a modest reduction in the bias on five frame variables (the average relative bias fell from 7.6% to 7.2%).

Table 2 summarizes the results from the experimental studies. In general, the studies demonstrate how difficult it is to raise response rates—a problem that is hardly unique to responsive and adaptive designs. For example, only two of the 16 experiments that were described by Wagner *et al.* (2012) significantly raised response rates in the NSFG. Three of the studies demonstrated reductions in *variation* in response rates across subgroups of the sample (see also Groves *et al.* (2005) and Rosen *et al.* (2014)), although this apparent reduction in the variation in estimated

Table 2. Selected study characteristics and outcomes, by experimental study

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	Significantly increased number of calls to priority cases in 7 of 16 experiments Significantly increased response rate in two experiments
		Screener week	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	Variance in response propensities lower 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% *versus* 9.9%) 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	Lower co-operation rate in adaptive group *R*-indicator significantly improved in adaptive group
		Difficult cases assigned to best telephone interviewers; easiest cases to worst telephone interviewers	Reduced variation in contact and co-operation rates in adaptive group No significant difference in costs or response rates

response propensities in the study by Peytchev *et al.* (2010) appeared to increase non-response bias rather than to reduce it. Laflamme and St-Jean (2011) reported that responsive designs reduced costs relatively to the standard protocol, but Luiten and Schouten (2013) reported that adaptive designs increased survey costs. Across all the studies (including cycle 6 of the NSFG), then, responsive and adaptive designs appeared to produce marginal gains in sample representativeness, but had little effect on overall response rates or overall costs.

4. Simulation studies of responsive or adaptive designs

Three additional studies have used simulations to explore the properties of responsive designs. In this section, we review these efforts.

4.1. Simulating the effect of stopping rules

Lundquist and Särndal (2013) examined the 2009 Swedish Living Conditions Survey, which follows a two-phase data collection strategy, with up to 20 telephone contact attempts in the first phase of data collection and an additional 10 (again by telephone) in the second. They noted that

> 'a data collection (including a follow-up) that proceeds according to an essentially unchanging form will produce very little change in the estimates beyond a certain "stability point" reached quite early in the data collection'

(page 561). The idea of a point of stability at which the data collection protocol should be altered or terminated is similar to Groves and Heeringa's (2006) notion of 'phase capacity', in which a given data collection protocol reaches its limit in terms of effectiveness. Lundquist and Särndal (2013) showed that the estimated non-response bias (based on three variables available for both respondents and non-respondents from the Swedish population register) in the Living Conditions Survey was lowest after 5–10 call attempts and grew progressively worse thereafter. The second phase of data collection, which increased the response rate from 60.4% to 67.4%, made the non-response biases worse for two of the three variables from the population register.

Lundquist and Särndal (2013) examined three alternatives to the strategy of continuing the same data collection protocol (i.e. repeated call-backs by telephone) up to a maximum of 30 attempts. They divided the sample into eight subgroups based on education (high *versus* low), property ownership (owner *versus* non-owner) and origin (born in Sweden *versus* born abroad). Alternative 1 would monitor response rates for the eight subgroups at call 12 of the initial phase of data collection and again at call 2 of the second phase; data collection would be ended at these points for subgroups with response rates of 65% or better. This strategy would have yielded a more balanced sample (one that more closely resembled the population on a vector of eight demographic characteristics) than the final Living Conditions Survey sample, though it would have lowered the overall response rate to 63.9%. The second alternative would have ended data collection for a subgroup as soon as its response rate reached 60%. This strategy would have produced even greater balance than the first alternative, despite a further drop in the response rate (to 58.9%). The final alternative would have terminated data collection for each group as soon the response rate for the subgroup reached 50%. This strategy would have produced the most balanced sample of all and it would have saved a large number of call attempts (reducing the total number of call attempts by 36.4%), while lowering the overall response rate to 50.3%. In part, the final strategy seems to have worked so well because it would have lowered the response rates in the high propensity subgroups so that they were closer to those in the low propensity subgroups (two of which never achieved a 50% response rate).

4.2. The effect of different case prioritization strategies

Särndal and Lundquist (2014b) simulated the effects of two methods for equalizing response propensities across cases, again using data from the Living Conditions Survey (see also Särndal and Lundquist (2014a)), as well as from a second survey, 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 that were examined in 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 top portion of the sample in terms of response propensities is set aside and fieldwork continues only for the remaining cases. In both surveys, both methods for equalizing the response propensities reduced imbalance (the distance between the respondents and the full sample on a set of auxiliary variables available for the full sample) compared with continuing to field all remaining non-respondents, as was done in the actual surveys. Another conclusion from this study is that calibrating the sample by using the auxiliary variables removed some of the non-response bias, but that bias was reduced even further when the set of respondents was better balanced. This is an important finding, since the same variables that are available for 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 (2014a, b) found gains for both.

Beaumont *et al.* (2014) reported another simulation study that examined the effect of case prioritization. In their simulation, they examined several quality indicators, including the variance of the non-response-adjusted estimator conditionally on the selected sample:

$$\text{var}_q(\hat{\theta}|s) \cong \sum_{g=1}^{G} \left(\frac{1}{p_g} - 1 \right) (n_g - 1) S^2_{y,g}$$

in which $\hat{\theta}$ is an estimated total (adjusted for non-response), p_g and n_g are the response rate and initial sample size for group g and $S^2_{y,g}$ is the variance on a variable y within group g. They contrasted four data collection strategies:

(a) *constant effort*, or no case prioritization,
(b) *optimal effort* (by reducing calls for groups approaching their target response rate),
(c) adjusting effort to *equalize response rates* across groups (by concentrating calls on low response propensity groups) and
(d) adjusting effort to *maximize the overall response rate* (by concentrating calls on high propensity groups).

Their simulations also assumed three 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. These scenarios correspond to the missing completely at random, missing at random and not missing at random missing data mechanisms respectively.

On the basis of their simulation, Beaumont *et al.* (2014) reached three major conclusions. First, when response propensities are constant overall or constant within group, all the effort scenarios produce unbiased estimates; however, when the propensities are strongly related to the survey variable, all the effort scenarios produce bias. Second, neither the R-indicator nor the non-response rate is a good indicator of non-response bias or non-response variance. Finally, when response propensities are known, the optimal effort strategy produced somewhat lower root-mean-square error RMSE than the other strategies (see Table 2 in Beaumont *et al.* (2014))

and the strategy that attempts to maximize response rates produced the worst RMSE. Of course, part of the problem is that response propensities are *not* known and may not be accurately estimated from the available auxiliary variables (although see Durrant and Steele (2009) and Durrant and D'Arrigo (2014), for some examples of apparently good prediction of response propensities when paradata are included in the propensity models).

4.3. Summarizing remarks

This section has reviewed the results of the studies simulating the effect of implementing various stop rules or case prioritization strategies. The primary objective of the simulations was the reduction of non-response bias, although the researchers also examined the cost implications of the various strategies. The simulation study by Lundquist and Särndal (2013) suggests that *lowering* the response rates in high propensity groups may be the best strategy; it produces more nearly equal response rates across subgroups, greater balance in the samples and lower data collection costs. Similarly, Särndal and Lundquist (2014a, b) showed that diverting effort from high propensity groups to low propensity groups yielded greater balance in the respondent sample on the auxiliary variables that were available from the frame and lower non-response bias on three register variables that were not used either in the stop rules or the post-survey calibration scheme. But the study by Beaumont *et al.* (2014) demonstrates that all data collection strategies, whether responsive or not, have the potential to increase non-response bias. As they showed, attempting to maximize response rates (by focusing on high propensity cases) may make the bias worse. When the objective is reduction of non-response bias, the options may be limited,

Table 3. Selected study characteristics and outcomes, by simulation study

Study	*Statistical goal*	*Simulated intervention(s)*	*Results*
Lundquist and Särndal (2013)	Increase sample balance, reduce non-response bias	Stopping data collection for a subgroup once a target response rate achieved for that subgroup	Lowest response rate threshold produced the highest balance Lowest threshold also achieved lowest non-response bias (on three registry variables)
Särndal and Lundquist (2014a, b)	Increase sample balance, reduce non-response bias	12 stopping rules (3 target response rates plus the equal proportions method × 3 vectors for defining subgroups)	Lowest response rate threshold again produced the highest balance Lowest threshold also achieved lowest non-response bias on three registry variables Both balance in data collection and calibration reduce non-response bias
Beaumont *et al.* (2014)	Optimal effort, equalize response rates, maximize overall response rate	4 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 RMSE; maximizing the response rate, the worst

since few design features have been consistently found to increase responses from low propensity cases (Brick, 2013). Furthermore, unless there are large increases in the response rates for the low propensity cases, the change in the bias will be small. Unfortunately, Lundquist and Särndal's (2013) finding that it is easier to increase non-response bias than to reduce it substantially may reflect the current difficult climate for conducting surveys.

Table 3 presents a summary of the simulation studies that were reviewed here.

5. Discussion

Our review summarized the studies that have examined the effectiveness of responsive or adaptive designs. These studies examined the effect of a change in protocol during the final phase of data collection or presented the results from experiments or simulations.

Several conclusions emerge from these studies. First, major changes in survey protocol (e.g. shorter questionnaires, much larger incentives or switching to face-to-face data collection) are more likely to reduce non-response bias than simply continuing with the original data collection protocol. This is one of the clearest contributions to emerge from the literature on responsive and adaptive design. The non-experimental studies (e.g. Axinn *et al.* (2011), Peytchev *et al.* (2009) and Rosen *et al.* (2014)) all point to this conclusion, although in most cases the gains in terms of bias reduction are modest. Still, anything that reduces non-response bias while simultaneously lowering cost is clearly an advance. The advantages of two-phase (or multiphase) designs that focus resources on a subsample of respondents have been apparent for almost 70 years, although here also the reductions in non-response bias are not always dramatic. What is new in this literature is the finding that designs that begin and end with a single protocol for all cases—continuing with more of the same—may actually make things worse (as in Lundquist and Särndal (2013)). The findings to date involve relatively dramatic shifts in protocol during the final phase, such as much larger incentives or shifting from telephone to face-to-face data collection, and these moves may not be feasible for many studies (e.g. because face-to-face interviewing is just too expensive). It is difficult for *all* surveys—whether they use responsive or adaptive designs or more conventional approaches—to reduce non-response bias, particularly in an era of shrinking survey budgets.

Second, despite numerous attempts to use propensity models to improve the efficiency of data collection, the gains in response rates or reductions in variation in response propensities have typically been very modest (see our summary in Table 2). It is difficult both to raise overall response rates to achieve precision goals and to equalize response propensities across subgroups to reduce non-response bias. These empirical findings tend to be confirmed by the simulation studies, which suggest that reducing variation in response propensities may not have a very large effect on the bias. The most practical way to reduce variation in response propensities and to reduce costs may be to stop or reduce effort on high propensity cases or subgroups and to focus instead on lower propensity cases. Variations on this strategy are examined by all three simulation studies, and they consistently improve sample balance, reduce bias or lower RMSE. For many survey sponsors who are concerned about precision, giving up or reducing effort on the (relatively) easy cases, and therefore reducing overall response rates and sample sizes, may not be a palatable strategy. Nonetheless, we believe that stopping rules of the type that was explored by Särndal and Lundquist (2014a) are likely to be useful tools for managing surveys. If estimates with the same mean-square errors can be obtained for less cost, this is surely a positive development. There is no sense in throwing good money after bad. We believe that further work on optimal stopping rules is definitely warranted.

The simulation studies also confirm that raising response rates may not lower bias and indeed may make it worse (Beaumont *et al.*, 2014; Lundquist and Särndal, 2013; Särndal and Lundquist, 2014a, b). Bringing in more of the same—additional cases who resemble the existing respondents—may only make matters worse (as we argued earlier). This conclusion about the relationship between non-response rates and non-response bias echoes findings from earlier studies (e.g. Curtin *et al.* (2000), Groves and Peytcheva (2008) and Keeter *et al.* (2000)), which demonstrate that response rates are at best poor indicators of non-response bias. Our recommendation is that survey managers monitor a range of indicators during fieldwork, including balance indicators, full sample and partial R-indicators, key survey estimates, response rates and subgroup response rates, and costs per case. All surveys—whether or not they explicitly adopt responsive or adaptive designs—attempt to deploy field resources as efficiently as possible and it is important to take into account as much information as possible in making allocation decisions. Our proposed measures of the value of a case require that response propensities be estimated, that predicted values be imputed for key variables and that key estimates be computed and monitored throughout the field period. An efficient allocation of effort would take into account both the value of the case (by one of our measures) and the likely cost of collecting data from the case.

Finally, the work by Särndal and Lundquist (2014a, b) and Lundquist and Särndal (2013) addresses an important issue—whether it is useful to achieve balance on auxiliary variables during data collection when after-the-fact weighting adjustments can have similar effects on the estimates. Their results consistently support the value of achieving balance during data collection for reducing bias (see especially Särndal and Lundquist (2014b); see also Schouten *et al.* (2014), who give theoretical arguments pointing to the same conclusion). Achieving balance during data collection is also likely to have the added benefit of reducing the effect of weighting on the variance of the survey estimates.

If the results to date indicate that the gains from responsive or adaptive survey designs have not been dramatic, it is worthwhile to speculate on factors that might be limiting their potential effect. At least four factors come to mind. First, in the current survey climate, it is extremely difficult to make dramatic gains. Perhaps the most that can be reasonably hoped for is to limit the losses. Second, to make these approaches work, researchers need auxiliary variables that are related to both response propensities and the key survey variables. Without this, the predicted response propensities may bear little relation to the actual response propensities. Managing data collection based on inaccurate propensities seems unlikely to produce large increases in efficiency or large reductions in bias. Still, the inclusion of paradata in propensity models may improve the fit of propensity models (Durrant and Steele, 2009; Durrant and D'Arrigo, 2014). In addition, to produce gains at the estimation stage, the auxiliary variables also must have a strong relationship to the survey variables (Särndal and Lundquist, 2014a). Unfortunately, the auxiliary variables, including paradata, that are available in many cases may not have the necessary predictive power. Part of the problem is that both records-of-calls data and interviewer observations are prone to measurement error (Biemer *et al.*, 2013; West, 2013), attenuating any relationship that they might have to the survey variables or the response propensities. Again, this issue is not unique to surveys employing responsive and adaptive designs; non-predictive auxiliary variables will also limit the effectiveness of post-survey weighting adjustments.

Third, it is not always clear how to intervene to obtain cases, particularly cases with low underlying propensities, to respond even when the model accurately predicts propensities. The surest methods are to offer much larger incentives or much shorter questionnaires, but often neither of these methods is feasible. Increasing the value of incentives has consistently been shown to produce diminishing returns; reducing the number of items means that some data items are not available for the full analysis. Other techniques generally produce minimal gains.

As we have already noted, it may be easier to lower the actual propensities of high propensity cases (by reducing the effort for these cases) than to raise the propensities of the difficult cases.

Fourth, as Wagner's work has shown (e.g. Wagner (2013)), interventions may not always be carried out as planned. Designating a case 'high priority' may not have much effect on how much effort it receives. It may be that interviewers are already allocating their effort optimally (or near optimally); for instance, having gone into the field to work a high priority case, an interviewer may correctly recognize that there is little marginal cost in also attempting to contact a nearby case of lower priority. It is often difficult in practice to persuade field interviewers to follow dictates from the central office. Again, this is a general issue for managing face-to-face surveys, not just for those using responsive or adaptive designs. In this respect, telephone surveys may offer a more promising platform for testing mid-course corrections involving interventions on interviewers. In addition, it seems likely that face-to-face interviewers will come under greater scrutiny in the future as survey managers adopt computer-assisted recording of interviews and Global Positioning Systems to monitor fieldwork. These developments may augur well for adaptive and responsive designs of the future.

Despite these limitations, we believe that the literature on responsive and adaptive design has some useful lessons for survey managers:

(a) it is important to clarify statistical priorities at the outset of the survey and to monitor measures of quality related to these priorities;
(b) it is useful to identify difficult cases (or subgroups) on the basis of previous rounds of the survey or based on propensity models fitted in the current round;
(c) no single indicator gives a complete picture of the risk of error in a survey and multiple indicators should be monitored during the field period;
(d) tailoring the data collection protocol to different subgroups may increase sample representativeness (one size may not fit all);
(e) if changes in protocol are made in later phases, these should be large and should produce markedly different types of respondents; continuing the same protocol from the outset may worsen non-response bias;
(f) face-to-face interviewers may ignore instructions about how they should do their work and close monitoring (or large incentives) may be needed to ensure that they implement changes in protocol;
(g) sometimes the best strategy is to cease further efforts by imposing stopping rules.

Many questions about these designs remain unresolved, and these suggest areas where further research is needed. One such area is the effect of adaptive and responsive designs on statistical properties of the estimators, such as their bias and variance. Responsive and adaptive designs sometimes use interim information from the data to modify the later phases of data collection. The design modifications arising from these multiple looks at the data may affect the properties of the estimators, as occurs in adaptive clinical trials.

For us, one of the central issues is whether it is generally better to achieve balance during data collection or afterwards via weighting. Various considerations must be taken into account here—the higher cost of continuing fieldwork to achieve balance, the possible dangers of achieving balance by reducing effort on 'easy' cases (those with high underlying response propensities), the likely effectiveness of the weighting variables in removing any non-response biases and the increased variance that is associated with more extreme adjustments to the design weights. It would be helpful if there were additional empirical demonstrations of the utility of increasing balance during data collection and if there were also good theory-based guidelines for achieving the best trade-off. For example, it may be that weighting makes more sense in countries with good population registries than in countries, like the USA, that lack such registries (although Särndal and Lundquist (2014a, b) still found benefits for increasing balance during data collection even when good registry data are available).

Another important question that is raised by this literature is what measures during data collection are the most effective at achieving balance? When interventions are taken or mid-course corrections made, which (increased incentives, changes in the mode of data collection, shortening the questionnaire, refusal conversion, further call-backs, and so on) are more likely to increase balance (and under what circumstances) and which are likely to exacerbate imbalances by bringing in more of the same? We look forward to the next round of studies that help to resolve these and other important issues as we try to bring more science to bear in managing survey data collection.

References

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. *J. Off. Statist.*, **30**, 607–621.

Biemer, P. P., Chen, P. and Wang, K. (2013) Using level-of-effort paradata in non-response adjustments, with application to field surveys. *J. R. Statist. Soc.* A, **176**, 147–168.

Brick, J. M. (2013) Unit nonresponse and weighting adjustments: a critical review. *J. Off. Statist.*, **29**, 329–353.

Calinescu, M., Bhulai, S. and Schouten, B. (2013) Optimal resource allocation in survey designs. *Eur. J. Oper. Res.*, **226**, 115–121.

Chapman, C. (2014) National Center for Education Statistics adaptive design overview. *Federal Committee on Statistical Methodology Conf., Washington DC, Dec. 16th*. The Hague: International Statistical Institute.

Chow, S.-C. and Chang, M. (2011) *Adaptive Design Methods in Clinical Trials*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Couper, M. P. and Wagner, J. (2011) Using paradata and responsive design to manage survey nonresponse. In *Proc. 58th Wrld Statistical Congr.*, pp. 542–548. (Available from `http://2011.isiproceedings.org/papers/450080.pdf`.)

Cox, D. R. (1952) Estimation by double sampling. *Biometrika*, **39**, 217–227.

Curtin, R., Presser, S. and Singer, E. (2000) The effects of response rate changes on the Index of Consumer Sentiment. *Publ. Opin. Q.*, **64**, 413–428.

Durrant, G. B. and D'Arrigo, J. (2014) Doorstep interactions and interviewer effects on the process leading to cooperation or refusal. *Sociol. Meth. Res.*, **43**, 490–518.

Durrant, G. B. and Steele, F. (2009) Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *J. R. Statist. Soc.* A, **172**, 361–381.

Food and Drug Administration (2010) *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*. Washington DC: Food and Drug Administration. (Available from `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf`.)

Godbout, S., Beaucage, Y. and Turmelle, C. (2011) Achieving quality and efficiency using a top-down approach in the Canadian Integrated Business Statistics Program. *Conf. European Statisticians, Ljubljana, May 9th–11th*.

Greenberg, B. S. and Stokes, S. L. (1990) Developing an optimal call scheduling strategy for a telephone survey. *J. Off. Statist.*, **6**, 421–435.

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.

Groves, R. M. and Heeringa, S. G. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. R. Statist. Soc.* A, **169**, 439–457.

Groves, R. and Peytcheva, E. (2008) The impact of nonresponse rates on nonresponse bias. *Publ. Opin. Q.*, **72**, 167–189.

Hall, P. (1981) Asymptotic theory of triple sampling for sequential estimation of a mean. *Ann. Statist.*, **9**, 1229–1238.

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

Jennison, C. and Turnbull, B. W. (2010) *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: CRC Press.

Judkins, D. R., Mosher, W. D. and Botman, S. (1991) *National Survey of Family Growth: Design, Estimation, and Inference*. Hyattsville: National Center for Health Statistics.

Keeter, S., Kohut, A., Miller, C., Groves, R. and Presser, S. (2000) Consequences of reducing nonresponse in a large national telephone survey. *Publ. Opin. Q.*, **64**, 125–148.

Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B.-E., Walters, E. E., Zaslavsky, A. and Zheng, H. (2004) The US National Comorbidity Survey Replication (NCS-R): design and field procedures. *Int. J. Meth. Psychiatr. Res.*, **13**, 69–92.

Korn, E. L., Arbuck, S. G., Pluda, J. M., Simon, R., Kaplan, R. S. and Christian, M. C. (2001) Clinical trial designs for cytostatic agents: are new approaches needed? *J. Clin. Oncol.*, **19**, 265–272.

Laflamme, F. and Karaganis, M. (2010) Development and implementation of responsive design for CATI surveys at Statistics Canada. *European Quality Conf., Helsinki*. (Available from `https://www.researchgate.net/publication/228583181_Implementation_of_Responsive_CollectionDesign_for_CATI_Surveys_at_Statistics_Canada`.)

Laflamme, F. and St-Jean, H. (2011) Highlights and lessons from the first two pilots of responsive collection design for CATI surveys. *In Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 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*. Hyattsville: National Center for Health Statistics.

Lohr, S. (1990) Accurate multivariate estimation using triple sampling. *Ann. Statist.*, **18**, 1615–1633.

Luiten, A. and Schouten, B. (2013) Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *J. R. Statist. Soc.* A, **176**, 169–189.

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

Massey, D. and Tourangeau, R. (2013) New challenges to social measurement. *Ann. Am. Acad. Polit. Socl Sci.*, **645**, 6–22.

Morenoff, J. D., House, J. S., Hansen, B. B., Williams, D. R., Kaplan, G. A. and Hunte, H. E. (2007) Understanding social disparities in hypertension prevalence, awareness, treatment, and control: the role of neighborhood context. *Socl Sci. Med.*, **65**, 1853–1866.

Peytchev, A., Baxter, R. K. and Carley-Baxter, L. R. (2009) Not all survey effort is equal: reduction of nonresponse bias and nonresponse error. *Publ. Opin. Q.*, **73**, 785–806.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010) Reduction of nonresponse bias in surveys through case prioritization. *Surv. Res. Meth.*, **4**, 21–29.

Rivers, D. (2006) Sample matching: representative sampling from internet panels. *White Paper Series*. Polimetrix, Palo Alto.

Roberts, C., Vandenplas, C. and Stähli, M.E. (2014) Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs. *Surv. Res. Meth.*, **8**, 67–80.

Rosen, J. A., Murphy, J., Peytchev, A., Holder, T., Dever, J. A., Herget, D. R. and Pratt, D. J. (2014) Prioritizing low-propensity sample members in a survey: implications for nonresponse bias. *Surv. Pract.*, **7**, no. 1.

Särndal, C.-E. (2011) The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *J. Off. Statist.*, **27**, 1–21.

Särndal, C.-E. and Lundquist, P. (2014a) Balancing the response and adjusting estimates for nonresponse bias: complementary activities. *J. Soc. Statist.*, **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. *J. Surv. Statist. Methodol.*, **2**, 361–387.

Schouten, B., Calinescu, M. and Luiten, A. (2013) Optimizing quality of response through adaptive survey design. *Surv. Methodol.*, **39**, 29–58.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**, 101–113.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2014) Theoretical and empirical support for adjustment of nonresponse by design. *Discussion Paper*. Statistics Netherlands, The Hague. (Available from `www.cbs.nl/NR/rdonlyres/AD8BC9BC-38B4-4BC9-B79C-6DF5A2E1B34B/0/201415x10pub.pdf`.)

Stephenson, C. B. (1979) Probability sampling with quotas: an experiment. *Publ. Opin. Q.*, **43**, 477–496.

Sudman, S. (1966) Probability sampling with quotas. *J. Am. Statist. Ass.*, **61**, 749–771.

Valliant, R., Dorfman, A. and Royall, R. M. (2000) *Finite Population Sampling and Inference: a Prediction Approach*. New York: Wiley.

Wagner, J. (2013) Adaptive contact strategies in telephone and face-to-face surveys. *Surv. Res. Meth.*, **7**, 45–55.

Wagner, J. and Hubbard, F. (2014) Producing unbiased estimates of propensity models during data collection. *J. Surv. Statist. Methodol.*, **2**, 323–342.

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. *J. Off. Statist.*, **28**, 477–499.

Wang, S. J., Hung, H. M. J. and O'Neill, R. T. (2009) Adaptive patient enrichment designs in therapeutic trials. *Biometr. J.*, **51**, 358–374.

Weeks, M. F., Kulka, R. A. and Pierson, S. A. (1987) Optimal call scheduling for a telephone survey. *Publ. Opin. Q.*, **51**, 540–549.

West, B. T. (2013) An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *J. R. Statist. Soc.* A, **176**, 211–225.