Survey Research Methods (2016)
Vol. 10, No. 1, pp. 35-47
doi:10.18148/srm/2016.v10i1.6157

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

Adaptive survey designs for nonresponse and measurement error in multi-purpose surveys

Melania Calinescu
Department of Mathematics
VU University Amsterdam
The Netherlands

Barry Schouten
Statistics Netherlands and
Utrecht University
The Netherlands

Recently, survey methodology literature has put forward responsive and adaptive survey designs as means to make efficient tradeoffs between survey quality and survey costs. The designs, however, restrict quality-cost assessments to nonresponse error, while there are various design features that impact also measurement error, e.g. the survey mode, the type of questionnaire (long or condensed) and the type of reporting (self or proxy). Extension of adaptive survey design to measurement error is, however, not straightforward when a survey has many and diverse survey items. An adaptive survey design needs to make an overall choice of design features that applies to all survey items simultaneously. In this paper, we investigate adaptive survey designs that account for both nonresponse and measurement error. In order to do so, we model the underlying causes for differences in measurement error between design features. This leads to response styles or response latencies. We tailor efforts so that either response style propensities are minimized or constrained. We illustrate the ideas with a case study on the 2008 Dutch Labor Force Survey. The design features in this study are the type of reporting (self-reporting only versus proxy-reporting allowed), and the number of face-to-face calls.

Keywords: Responsive survey design; Response error; Mixed-mode surveys; Response latencies

1 Introduction

To date, virtually all literature on adaptive and responsive survey design has focused on bounding the impact of nonresponse error using trade-offs between costs and indirect measures of nonresponse bias. We refer to Calinescu (2013), Calinescu, Bhulai, and Schouten (2013), Chesnut (2014), Coffey, Reist, and White (2013), Laflamme and Karaganis (2010), Luiten and Schouten (2013), Lundquist and Särndal (2013), Peytchev, Riley, Rosen, Murphy, and Lindblad (2010), Schouten, Calinescu, and Luiten (2013), Wagner (2013), Wagner et al. (2012), and Rosen et al. (2014). Adaptive survey designs allocate different data collection strategies to different population strata by optimizing explicit quality and cost measures. The characteristics that form strata come from the sampling frame, from other linked administrative data or from paradata. Design features like the advance letter, questionnaire, mode, contact procedure, interviewer and/or incentives may be varied over the strata. Currently, the designs assume that measurement error is affected in the same way by all design features under consideration. While this assumption may be acceptable for some design features, like the number of calls, it is certainly not valid for design features like the survey mode (Web, Mail, telephone, face-to-face), the type of reporting (self-reporting or proxy reporting), and the type of questionnaire (long form or condensed basic question form). In this paper, we attempt to generalize adaptive survey design to design decisions where both nonresponse and measurement error are affected. Such a generalization is both promising and urgent; many survey institutes are redesigning their surveys to multi-mode designs or are considering such redesigns.

In this paper, we apply mathematical programming to select optimal adaptive survey designs. Although mathematically elegant and transparent, such an approach is quite demanding in terms of the availability and accuracy of estimated design parameters like response propensities. A less strict trial-and-error approach may be adopted along the same lines; the optimal designs following from mathematical programming may inform such an approach.

When the survey has only one or a very small number of key survey variables, like the consumer trust index in consumer sentiments surveys, or the total distance travelled in travel surveys, then optimization of the adaptive survey design can focus entirely on that variable. Calinescu and Schouten (2015) propose to minimize the nonresponse-adjusted method effect for the key survey variable relative

Contact information: Barry Schouten, Statistics Netherlands, Department of Methodology PO Box 24500, 2490HA The Hague ,The Netherlands (jg.schouten@cbs.nl)

to some benchmark design and subject to a number of constraints like costs and precision. The non-response adjusted method effect then is the net effect of unadjusted nonresponse bias and measurement bias differences between design choices. This approach is not feasible when a survey has many and/or diverse key survey variables, like crime victimization surveys, health surveys, living conditions surveys, omnibus surveys or time and budget expenditure surveys.

The literature has put forward several options to signal, detect or directly quantify measurement error. One option is to use validation data or record check data that are considered error-free. The answers to the survey items are compared to these gold standard data and provide a direct measure of error on the items. See, for example, Bakker (2012). A second option, commonly applied in psychometric literature, e.g. Weijters, Schillewaert, and Geuens (2008) and Klausch, Hox, and Schouten (2013), is to perform factor or latent class analyses and to investigate the reliability and validity of the survey items in the analyses towards latent variables. Such models may even explicitly include factors or classes that represent certain forms of measurement error, see, for example,Van Rosmalen, Van Herk, and Groenen (2010) and Heerwegh and Loosveldt (2011). If one believes in the imposed latent variable configuration, then the analysis again reveals direct measures of error over items. A third, and final, option is to summarize answering behavior that is deemed undesirable or prone to error, because they signal a potential deficiency in the answering process. Such summaries may be based on paradata or process data from the answering process, like the durations per completed item or observations from interviewers. However, summaries may also be based on the response data alone, like the frequency of do-not-know answers.

The results of such assessments may be summarized as response quality indicators, e.g. an indicator for a difference to linked validation data, an indicator for a large variation in answers to multiple scale questions, or an indicator for a fast responder. For an insightful recent use of such indicators in a measurement error analysis, see Medway and Tourangeau (2015). Medway and Tourangeau (2015) also include an overall indicator, measuring whether at least one of the response quality indicators signaled potential measurement error.

We propose to account for measurement error over multiple survey items in adaptive survey design by including propensities for the occurrence of response quality indicators in the mathematical optimization problems. We call such propensities response quality propensities. They are analogous to response propensities that correspond to the occurrence of a unit response. We investigate two approaches to include them and demonstrate the approaches on a simple case study based on real survey data from the 2008 Dutch Labor Force Survey (LFS). In the LFS case study, we analyze the type of reporting and the number of calls as design features. We evaluate response quality indicators based on differences to linked tax data on jobs. Details of the LFS case study can be found in Calinescu, Schouten, and Bhulai (2012).

In section 2, we construct the framework for adaptive survey design accounting response propensities and response style propensities. In section 3, we discuss the LFS case study. We end with a discussion in section 4.

2 An adaptive survey design framework for nonresponse and measurement error

The design choices behind adaptive survey design (ASD) can be formulated as a mathematical optimization problem in which a quality objective function is maximized given cost constraints and additional constraints on quality. In this section, we extend the ASD framework for nonresponse error and to measurement errors. We adapt the mathematical optimization problem so that it accounts explicitly for measurement error.

2.1 Adaptive survey design for measurement error

A key question, when extending ASD to measurement error, is whether the survey has a few or many variables of interest. When a survey has only one or a few key variables, then ASD may focus directly on the key variables and attempt to maximize accuracy. When a survey has many and/or diverse variables of interest, then such an approach cannot be adopted.

We see two options to deal with measurement error for multiple key survey variables in adaptive survey design: One option is to choose a distance function that transforms the vector of method effects for the individual survey variables to a single value. Method effects are then defined as the net effect of unadjusted nonresponse bias and measurement bias differences between design choices. Another option is to define and apply indicators of response quality.

The first option reduces the multi-dimensionality by a multivariate distance function. Although appealing at first, this option has a number of methodological problems. It is less straightforward how to choose a distance function for survey variables with different measurement levels, the method effect for a categorical survey variable must be derived itself from a vector of methods effects per category, and the resulting mathematical optimization problems are likely to become very hard to handle due to the complex, non-linear and non-convex objective function.

The second option is to employ indicators of response quality as summaries of response error over multiple survey items. Examples of such indicators are given in table 2.1. The list in table 2.1 is not exhaustive, but it does provide examples for the basic types of indicators: gold standard data, latent variable models, answering behaviors and composite. Gold standard indicators employ linked data that are considered error-free and summarize differences to the gold standard data. Latent variable model indicators are based on models that impose a structure of latent factors or classes on a series of survey items. If the models hold, then they allow for conclusions about measurement error. These conclusions may apply to random measurement error, i.e. to reliability, but also to systematic measurement error, i.e. to loadings on factors representing certain response styles or response latencies. We refer to Baumgartner and Steenkamp (2001), Saris and Gallhofer (2007) and Weijters et al. (2008) for a detailed account of such models. The answering behavior indicators attempt to signal deficiencies in the answering process. Tourangeau and Rasinski (1988) introduced four phases followed by respondents in answering survey questions: Interpretation and comprehension, Information retrieval, Judgment, and Reporting. Measurement error may be seen as a deficiency in this process for a survey item, either deliberate or subconscious. Answering behavior indicators focus on one or more phases. For instance, indicators based on durations over completed survey items focus on short-cutting the four phases. Composite indicators combine several indicators in a single overall indicator, e.g. the respondent showed at least one negative score on any of the indicators. We will not discuss the various examples in table 2.1, as this would go beyond the scope of this paper. The response quality indicators can be included in the optimization by estimating their propensities to occur for different population subgroups. We call these, response quality propensities.

The option of response quality indicators also has methodological and conceptual drawbacks but the advantages are twofold and, in our opinion, outweigh that of the multivariate distance function option: First, the resulting optimization problems are more tractable as the number of response quality indicators is, generally, much smaller than the number of survey variables. Second, a close analogy is reached to traditional adaptive survey designs for nonresponse error that are based on response propensities only; they may be seen as special cases where response quality propensities are set to zero. We, therefore, opt for the response quality indicator option.

It is important to note that, for adaptive survey design, it is sufficient to consider response quality indicators that are known or conjectured to be affected differently by the design features under consideration. A choice between a long and condensed form of a questionnaire or between different survey modes will have different impacts on the quality of the response. Although, the number of potential response quality indicators is large, only a subset may have to be employed.

2.2 Adaptive survey designs for nonresponse error

ASD optimization for nonresponse error is the optimal allocation of design features or strategies to population subgroups such that specified indirect measures for nonresponse error are optimized given a set of constraints, the most prominent being costs (e.g. Schouten et al., 2013). One may also minimize costs subject to constraints on the indirect measures for nonresponse error and subject to other constraints, but this dual problem is less frequently described and analysed in the literature. The differentiation of strategies over subgroups is what separates ASD from traditional uniform designs. However, in ASD there is also a more explicit focus on indirect measures for nonresponse error and costs. The subgroups or ASD strata are formed using administrative data, frame data or paradata. The set of allocation probabilities from subgroups to strategies form the decision variables in the optimization problem, e.g. Calinescu et al. (2013).

We introduce some notation: Sample units are clustered into homogeneous subgroups $\mathcal{G} = \{1, \ldots, G\}$. The relative sizes of the subgroups are denoted as w_g, with $\sum w_g = 1$. The set of available strategies is represented by $\mathcal{S} = \{1, \ldots, S\}$, and $\{p_g(s)\}_{s\in\mathcal{S}, g\in\mathcal{G}}$ is the set of allocation probabilities.

In this paper, we focus on the maximization of quality given a certain budget. The (expected) response rate may function as indirect measure for nonresponse error. It is defined as

$$\text{Response Rate:} \quad \max_{p_g(s)}(\rho) = \sum_{g\in\mathcal{G}, s\in\mathcal{S}} w_g p_g(s)\rho_g(s) \quad , \tag{1}$$

where $\rho_g(s)$ is the response propensity of group g to strategy s. Various constraints may be added. We describe a few, but our exposition is not exhaustive.

Let $c_g(s)$ be the costs incurred by allocating strategy s to subgroup g and B be the total available budget. The cost constraint is

$$\text{Cost Constraint:} \quad n \sum_{g\in\mathcal{G}, s\in\mathcal{S}} w_g p_g(s) c_g(s) \leq B \quad , \tag{2}$$

with n the sample size.

One may like to further constrain the impact of nonresponse error by bounding the sample variance, S^2, of response propensities on relevant subgroups. Here, we use the R-indicator, which is a transformation of the sample variance to the [0, 1] interval, where a value one represents optimal representativeness and a value zero represents maximal deviation from representativeness. See Schouten, Cobben, and Bethlehem (2009). We set a lower threshold α to the R-indicator, i.e. demand a minimal representativeness on the auxiliary variables included in the response propensity model.

$$\text{R-Indicator Constraint:} \quad 1 - 2S \leq \alpha \tag{3}$$

Table 1
Examples of response quality indicators based on gold standard data, latent variable models, and paradata on answering behavior.

Type	Response quality indicator
Gold standard data	Difference to validation data Difference to audit or record check data
Latent variable models	Amount of random measurement error in scale items (reliability) Loading on common factors/classes representing response styles or latencies
Answering behavior	Average duration per completed item Variance in durations over completed items Average decrease in duration per completed item over course of interview Percentage of items with missing data Percentage of items with do-not-know answers Percentage of items with order effects Percentage of items with agree answers Occurrence of rounding of answers (continuous measurement levels) Percentage of items with answers in non-sensitive categories Percentage of items with answers that skip filter questions Variance of responses to batteries of items

with

$$S^2 = \sum_{g \in \mathcal{G}} w_g \left(\sum_{s \in \mathcal{S}} p_g(s) \rho_g(s) - \rho \right)^2 \tag{4}$$

The response propensities in (4) are based on the same subgroups as the subgroups in the allocation of strategies. However, the choice of subgroups may, in general, be taken differently. It is important to remark that this constraint is non-linear and non-convex and, consequently, complicates the optimization problem when it is included.

An important quality constraint is the precision, which we operationalize as lower bounds β_g to the expected number of respondents

$$\text{Precision Constraint:} \quad n \sum_{s \in \mathcal{S}} w_g p_g(s) \rho_g(s) \geq \beta_g, \forall g \in \mathcal{G} \tag{5}$$

Again, the subgroups for precision constraints may be taken differently from the subgroups that are assigned strategies. Furthermore, the precision constraint may also be simplified to an overall constraint on the expected number of respondents.

For logistical or practical reasons, one may limit the number of switches in design features on some of the subgroups, e.g. allow for only one mode-switch or only one switch from self-reporting to proxy reporting. This constraint implies that some allocation probabilities are set to zero

$$\text{Strategy Switching Constraint:} \quad p_g(s) = 0 \text{ for } s \in \mathcal{S}_g \tag{6}$$

There are also technical constraints as allocation probabilities need to take values in the interval $[0, 1]$ and need to sum up to one.

$$\text{Regularity Constraint:} \quad \sum_{s \in \mathcal{S}} p_g(s) = 1, \forall g \in \mathcal{G}$$
$$\text{and } p_g(s) \in [0, 1], \forall g \in \mathcal{G}, s \in \mathcal{S} \tag{7}$$

The ASD optimization problem is constructed by selecting a number of the constraints and optimizing the quality objective function. Such optimization may be done in several ways; see Lundquist and Särndal (2013), Wagner et al. (2012), and Schouten and Shlomo (2014). In this paper, we apply mathematical optimization; see Calinescu (2013).

2.3 Combining nonresponse and measurement error into an adaptive survey design

We incorporate measurement error into the ASD framework through response quality propensities. We define response quality propensities as follows: First, a lower and/or upper limit is set to the response quality indicator, e.g. to the duration per completed item or to the magnitude of the difference to validation data. Second, for each respondent, his/her response quality, e.g. average completion time or difference to a linked validation record, is compared to the threshold(s) and a 0-1 score is derived. The response quality propensity then follows from modeling these 0-1 scores using auxiliary variables.

Assume that, in the comparison of the selected design features, M response quality indicators, labeled $m = 1, 2, \ldots, M$, are employed. Let $\left(A_{1,g}, \ldots, A_{M,g}\right)' \in \{0, 1\}^M$ be the vector containing the 0-1-values for exceeding the thresholds per indicator, and let

$$\theta_g\left(a_1, \ldots, a_M; s\right) = P\left[A_{1,g} = a_1, \ldots, A_{M,g} = a_M; s\right]$$

be the joint probability distribution for subgroup g, given strategy s is applied. $\theta_g\left(0, \ldots, 0; s\right)$ represents the propensity that none of the response quality indicator exceeds a threshold, while $1 - \theta_g\left(0, \ldots, 0; s\right)$ is the propensity that at least one exceeds a threshold.

We see two main approaches to include response quality propensities into the ASD: modify the response rate by the response quality rate, or constrain the response quality rate. We label them as Approaches I and II, in the following. Under Approach I, a response is only counted when a respondent exceeds at most a certain number of response quality indicator thresholds. Under Approach II, the proportion of respondents exceeding thresholds is constrained. We elaborate the two approaches.

Approach I uses a modified response rate. It is defined as

$$\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s) \theta_g(0, \ldots, 0; s) \quad , \tag{8}$$

and in the optimization the response rate objective in (1) is replaced by

Modified Response Rate:

$$\max_{p_g(s)} \left(\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s) \theta_g(0, \ldots, 0; s) \right) \quad . \tag{9}$$

An example of an optimization under Approach I is

$$\max_{p_g(s)} \left(\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g\left(s\right) \rho_g\left(s\right) \theta_g\left(0, \ldots, 0; s\right) \right)$$

subject to

$$n \sum_{g\in\mathcal{G},s\in\mathcal{S}} w_g p_g(s) c_g(s) \leq B$$

$$n \sum_{s\in\mathcal{S}} w_g p_g(s) \rho_g(s) \geq \beta_g, \forall g \in \mathcal{G}$$

$$\sum_{s\in\mathcal{S}} p_g(s) = 1, \forall g \in \mathcal{G} \text{ and } p_g(s) \in [0,1], \forall g \in \mathcal{G}, s \in \mathcal{S}$$

Approach II adds an additional constraint on the response quality rate instead of modifying the response rate as the objective function. The response quality rate is defined as

$$\frac{\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s) (1 - \theta_g(0, \ldots, 0; s))}{\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s)} \tag{10}$$

and an upper limit γ is set

Response Quality Constraint:

$$\frac{\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s) (1 - \theta_g(0, \ldots, 0; s))}{\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s)} \leq \gamma \quad . \tag{11}$$

By multiplying both sides of (11) by the denominator of response quality rate and by subtracting the resulting right-hand side from the left-hand side, the constraint can be rewritten to

$$\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s) (1 - \theta_g(0, \ldots, 0; s) - \gamma) {\leq} 0 \quad , \tag{12}$$

to preserve linearity of the problem in the allocation probabilities.

An example of optimization under Approach II is

$$\max_{p_g(s)}(\rho) = \sum_{g\in\mathcal{G},s\in\mathcal{S}} w_g p_g(s) \rho_g(s)$$

subject to

$$n \sum_{g\in\mathcal{G},s\in\mathcal{S}} w_g p_g(s) c_g(s) \leq B$$

$$1 - 2\mathcal{S} \geq \alpha$$

$$\frac{\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s) (1 - \theta_g(0, \ldots, 0; s))}{\sum_{s\in\mathcal{S},g\in\mathcal{G}} w_g p_g(s) \rho_g(s)} \leq \gamma$$

$$\sum_{s\in\mathcal{S}} \rho_g(s) = 1, \forall g \in \mathcal{G} \text{ and } p_g(s) \in [0,1], \forall g \in \mathcal{G}, s \in \mathcal{S}$$

where the response rate is maximized given constraints on cost, the R-indicator and given again the regularity constraints.

Of course, in the Approach I and II examples other constraints can be added. For details about algorithms and software for optimization problems sketched in this section, we refer to Calinescu (2013).

Both (8) and (10) can be altered to relax the impact of the response quality indicators; one may, for instance, allow for one indicator to exceed its threshold but not for more than one. $\theta_g\left(0, \ldots, 0; s\right)$ is then replaced by the sum over the probabilities for all vectors $(a_{,} \ldots, a_M)'$ with at most one $a_m = 1$.

The two approaches have different advantages. Approach I is attractive from a framework point of view: The response quality propensities are incorporated in the response propensities and the whole ASD framework for nonresponse can simply be applied. However, it seems to suggest also that respondents that show a deficiency on one or more aspects of response quality will be discarded, which may not be realistic in practice. Approach II leads to an additional constraint and, hence, a more complex optimization problem. It does, however, allow for tuning of the maximal average response quality propensity; it offers more flexibility and it is closer to practice.

Obviously, the input parameters to the ASD are subject to imprecision and to bias. The parameters need to be estimated based on historical survey data which has sampling error and may be outdated or incomplete.Schouten, Calinescu, and Burger (2014) discuss analyses that assess the sensitivity and robustness of adaptive survey designs to inaccuracy in the estimated response propensities and cost parameters. We return to this issue in the discussion of this paper.

The numerical optimization of non-linear problems presented here can be done in standard statistical software like SAS or R. The case study of this paper has a relatively small number of strata and the optimization problem was solved through an extensive grid search. In general, the number of decision parameters is too large to apply a direct search or a grid search, although large sections of the solution space can, usually, be discarded based on deductive logic. In such situations, packages like nloptr in R can be used, and we recommend to use various starting values in the optimization. The allocation probabilities of current non-adaptive designs should be one of the starting values, so that any optimum, even if it is a local optimum, is an improvement.

3 A case study: the Labor Force Survey

We perform one case study based on the 2008 Dutch Labor Force Survey (LFS). For this survey, we consider response quality indicators based on linked validation data.

3.1 The LFS data

The Dutch LFS is a monthly household survey using a rotating panel with five waves. After the first wave, households are asked whether they are willing to participate in the subsequent waves. The time lag between the waves is three months, so that any household stays in the panel for at most one year. The first wave is the longest and most detailed and may take up to 45 minutes. It contains questions about the number of employments, the number of hours worked, the desire to work more hours, any activities conducted to find employment, any barriers in trying to find employment, occupation and profession, and educational level. In 2008, this wave was administered by a face-to-face interviewer. Subsequent waves are much shorter and mostly ask for changes in employment and educational status. These waves were administered by telephone. Since the telephone waves are short and are far less costly, we focus completely on the first wave in the case study.

The LFS target population consists of persons between 15 and 65 years of age. The 2008 LFS uses a two stage sampling design with municipalities as primary units and addresses as secondary units. Households are self-weighting, except households with at least one person of 65+. All members in the household 15 years of age and older are interviewed. Proxy interviewing is allowed by members of the household. There is one restriction on proxy reporting: Children in the household are not allowed to provide proxy answers for their parents, regardless of their age.

Face-to-face interviewers have one month to make contact and complete the interview. They are instructed to spread visits over the month and over days and evenings. After the second failed contact attempt they leave a card with their name and telephone number, and after the third failed contact attempt they are allowed to call the household (when a phone number is available) to make an appointment. In 2008, a maximum of six visits to the address was allowed. If no contact was made at the sixth visit, then the address was processed as a noncontact.

For 2008, the total LFS sample that entered the first wave consisted of 135,332 persons. In the first wave, the interviewers obtained responses for 78,321 persons. Since the number of sample units is very large, we ignore any imprecision of the estimates for response and response quality propensities in this study.

The main publication cells for the LFS monthly statistics are gender and age groups 15-25, 26-55 and 56-65. Gender and age are both available from the sampling frame. We focus on the three age groups in the ASD optimization. Employment status and educational level differ significantly over these three subgroups. Their relative size in the sample is 19.6%, 62.4% and 18.0%, respectively.

3.2 Design features in the case study

We restrict ourselves to two design features: the type of reporting (self-reporting only vs. proxy-reporting allowed) and the number of face-to-face visits. We consider a design where only self-reporting is allowed and a design where also proxy reporting is possible. Furthermore, we allow for a maximum number of up to ten visits.

The type of reporting affects both nonresponse and measurement; proxy reports increase response rates, especially contact rates, but may introduce additional measurement error. The type of reporting has been investigated in the context of the LFS, see Lemaitre (1988), Moore (1988) and Thomsen and Villund (2011), and is known to have an impact on employment statistics.

The number of visits is clearly linked to the contact rate, and, hence, response rate, but there is no clear link of the number of visits to measurement error. This link would exist if harder to contact persons provide more or less measurement error. Harder to contact persons are known to have higher employment rates, but there is no evidence that they produce more measurement error for the LFS, see Schouten, Van der Laan, and Cobben (2014).

For these design features, we consider two main ingredients to the optimization problem: the response propensity per population stratum and the costs of a sample unit per population stratum. In the next section, we present the third ingredient: the response quality propensities per population stratum.

We start with the response propensities. Figure 1 presents estimated proportions of the sample responding per visit and per age group for self-reporting only (left panel) and proxy-reporting allowed (right panel). The estimation of the response propensities is partly model-based. The response propensities for a design with proxy reporting allowed are estimated directly from the survey data for visits one to six. For visits seven to ten, we fitted a geometrical distribution on the first six visits and extrapolated response propensities. Since the LFS allows for proxy reporting, we based contact propensities for self-reporting on the Dutch Health survey in which persons are sampled. However, we used the participation propensities of the LFS which is a simplification. We could have performed a regression to predict what family member is reporting him/herself and acts as proxy for others, and calibrate participation propensities. In estimating the response propensities for different numbers of visits, we ignore the impact of timing of visits and the strategy interviewers may adopt when they have smaller maximal numbers of visits. We simply simulated a cap on the number of visits, but ideally response propensities should be estimated based on real caps. Such data are hard to come by, of course, and with good reason. It must be expected that response propensities are underestimated for smaller caps, therefore, because interviewers adapt their strategy. Furthermore, there may be errors in the paradata on visits, see Biemer, Chen, and Wang (2013).

Figure 1 shows that there is potential gain in an adaptive survey design as response propensities are clearly different between age groups and between strategies. In both strategies, at lower number of visits, age group 56-65 has the highest proportions and age group 25-25 the lowest. These differences are mostly caused by the difference in contact rates, but also participation rates are highest for 56-65 and lowest for 15-25. Allowing for proxy reporting clearly raises the proportions of response at early visits, and also tends to bring the proportions closer to each other for the age groups.

Next, we move to the costs. The cost per sample unit for group g and strategy s, $c_g(s)$, is approximately equal to

$$c_g(s) = \mathrm{IPR} \times \mathrm{NV}_g(s) \times \mathrm{TT}_g(s) + \rho_g(s) \times \mathrm{ID}_g(s) \quad , \quad (13)$$

where IPR is the hourly interviewer pay rate, $\mathrm{NV}_g(s)$ is the expected number of visits in stratum g for strategy s, $\mathrm{TT}_g(s)$ is average travelling time per visit in stratum g for strategy s, and $\mathrm{ID}_g(s)$ is the average interview duration in stratum g for strategy s. $\rho_g(s)$ is the corresponding response propensity. Calinescu and Schouten (2015) use (13) to estimate costs for the LFS. They found (Table A5 in Calinescu & Schouten, 2015) that costs per sample unit are almost constant over population strata based on age, household size, ethnicity and registered unemployment. The underlying data showed that the average travelling time and interview duration varied only very little over the strata. However, the constant travelling time hold, because Statistics Netherlands interviewers handle multiple surveys at the same time; the contact strategy to the LFS gets subsumed in regular interviewer workloads. In the optimization, we make a simplification and use only the expected number of visits as proxy for costs, i.e. we bound the expected number of visits from above. In the same LFS historic survey data, it was found that the average interview duration is approximately four times longer than the average travelling time per visit. In other words, an increase of a half visit in the expected number of visits has the same impact on costs as an increase in the response propensity of 12.5%.

Also in terms of costs there is a clear potential gain for adaptive survey designs. The expected number of visits and the response propensity vary over the age groups and, obviously, depend strongly on the maximal number of visits.

3.3 Response quality indicators for the LFS

In this section, we consider the third ingredient to the adaptive survey design, the response quality propensities. Before we do, we first explain how we measure response quality for the LFS.

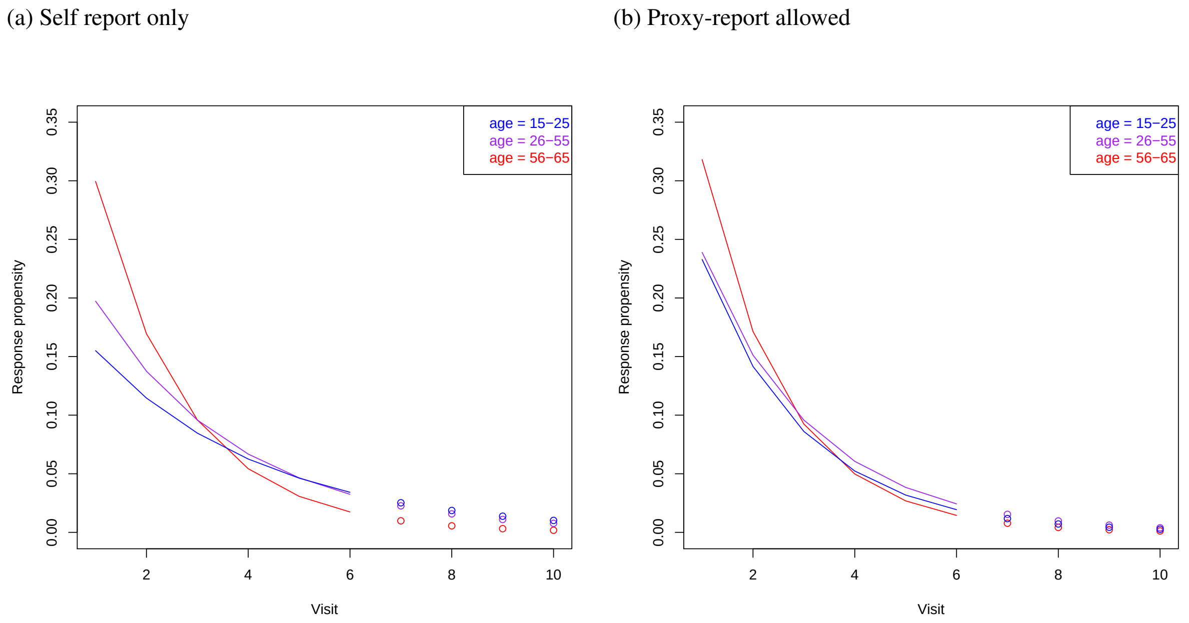
For our case study, we did not have the availability of paradata and the LFS questionnaire is not designed around latent constructs, but we did have linked administrative data. Various government administrative data about employment and registered unemployment could be linked to the sampling frame directly. We augmented the LFS with data from the Dutch Employment register (to determine whether a person is working in employment and the number of jobs this person has) and the Dutch Unemployed register (to determine whether a person is registered at an employment office in order to find a job). We defined three types of differences between the survey data and the administrative data:

- Difference 1: Not employed in register, but employed in LFS;
- Difference 2: Not employed in register and no employment office registration, but subscription to employment office in LFS;
- Difference 3: Employed in register, but not employed in LFS.

Since the two administrative data sets are essential to various ministries, they receive a lot of attention and quality checks. For this reason, we assume that there will be relatively few errors in the administrative data.

The type of reporting is expected to have an influence on the occurrence of measurement error, but the number of visits is not, or only mildly. A proxy reporter may be less knowledgeable and/or less motivated to provide answers, and, hence, may be more likely to fail to report one or more of the jobs.

The three differences between register data and LFS data may have various causes. A possible cause is socially desirable answering, i.e. the respondent chooses an answer category that displays him/her in a more favorable position. LFS respondents may feel that they need to have a job or may need to search for one. Another possible cause is motivated underreporting, i.e. the respondent deliberately reports a smaller



(a) Self report only

(b) Proxy-report allowed

Figure 1. Estimated proportions of the sample responding per visit and per age group. The left panel shows the proportions for self-reporting only and the right panel for a design where proxy reporting is allowed. The proportions for visit 7 to 10 are extrapolated using the first six visits (dotted line).

Table 2
Estimated response quality propensities per age group and overall for type 3 differences.

	Age group			
	15-25	26-55	56-65	All
Design feature	%	%	%	%
Self-reporting only	6.0	2.8	4.1	3.7
Proxy allowed	7.8	3.5	4.7	4.6

amount of items or events, e.g. Eckman et al. (2014). LFS respondents have to report details on all jobs they have and may decide not to as it is burdensome.

In the 2008 LFS, 8,0% of the respondents showed one of the three differences. This relatively large proportion of differences explains the difficulty Statistics Netherlands national accounts department has with integrating the two sources at a macro-level. The number of visits turned out to be unrelated to the occurrence of all three differences. Furthermore, the first and second difference are unrelated to the type of reporting. However, the third difference increases from 3.7% to 4.6%, when proxy reporting is allowed. Despite this relatively modest increase of 0.9%, we decided to employ a response quality indicator for a type 3 difference. Table 2 contains the propensities for the three age groups for difference 3.

The potential gain of adaptive survey designs for response quality are smaller than for response and costs, but there are clear differences between age groups and strategies. Table 2 shows that the largest differences are found for the youngest group; the proportion of respondents that deviate from administrative data increases from 6.0 to 7.8%.

It is important to note that we use only a single response quality indicator in our study. Clearly, measurement error in the LFS is much more diverse and high-dimensional than missing employments. More response quality indicators should and could be included, preferably based on paradata about answering behavior. We leave this to future research.

3.4 Results

We demonstrate the two approaches to include response quality propensities, as described in section 2.3, for the LFS. We do this by answering two questions that relate to realistic design decisions. Under both approaches, we include the cost constraint and the R-indicator constraint (apart from the standard regularity constraints), as presented in section 2.2. We do not include a precision constraint per subgroup, for the sake of simplicity and ease of presentation. In the following, we define the response quality rate as the proportion of respondents for who employments are missing, as explained in section 3.3.

In the regular, non-adaptive LFS no explicit constraints were set on the R-indicator or response quality rate. Table 3 contains the response rates when the average number of visits per sample unit is constrained to 2, 2.5 and 3 for the

Table 3
Optimal response rate for different cost constraints on the average number of visits per address for a design with self-reporting only and with proxy reporting allowed. Corresponding R-indicator values and response quality rates are given.

Cost constraint in average number of visits per address	Optimal adaptive survey design		
	Response rate %	R-indicator on age	Response quality rate %
Self-reporting only			
3	60.9	0.851	3.5
2.5	52.6	0.629	3.3
2	43.4	0.548	3.2
Proxy-reporting allowed			
3	64.2	0.939	4.5
2.5	62.4	0.924	4.5
2	50.5	0.718	4.2

design with self-reporting only and for the design in which proxy-reporting is allowed. Resulting R-indicator and response quality rates are computed. From table 3 it is clear that allowing for proxy reporting has a strong impact on the response rate. Surprisingly, also the R-indicator on the three age groups improves strongly. However, the proxy reporting comes at a cost; the response quality rates go up by 1% point. The number of visits also has a strong impact on the response rate and R-indicator, but less so on the response quality rate.

Now, we present the two examples of design decisions that could be performed, assuming that the current design is self-reporting only:

1. Can the response rate be improved for a budget level of 3 visits per address by allowing for proxy reporting while maintaining the R-indicator and response quality rate?

2. Can the R-indicator be improved for a budget level of 2.5 visits per address by allowing for proxy reporting while maintaining the precision and response quality rate?

Hence, in the first design decision, we take the self-reporting only design with a budget of 3 visits per address as benchmark. In the second design decision, this benchmark is the self-reporting only design with 2.5 visits per address.

For the first analysis, improving the response rate, we start by adopting Approach I, in which the modified response rate is maximized subject to constraints on cost and the R-indicator. Table 4 presents optimal modified response rates for different values of the two constraints. Also given are the corresponding, realized response rates, R-indicator values and response quality rates. For a maximum average of three visits per address, the optimization does lead to an increase in response rate of 3.1% and the R-indicator is significantly improved from 0.851 to 0.928, but an increase of the response quality rate of 0.5% has to be accepted. The response quality rate cannot be controlled. We, therefore, move to Approach II, in which the response rate is maximized subject to constraints on the cost, R-indicator and response quality rates. Table 5 contains the optimal response rates for various levels of the constraints plus their corresponding, realized R-indicator values and response quality rates. It turns out that the design without proxy reporting is optimal and, hence, cannot be further improved unless the response quality rate constraint is abandoned or relaxed. Hence, from the optimization it must be concluded that we cannot improve the response rate unless we accept an increase in response quality rate of 0.5%.

For the second analysis, improving the R-indicator, we assume a lower budget level of on average 2.5 visits per address. The R-indicator is then 0.629 and the response quality rate is 3.3%. Again, we start by looking at Approach I, assuming constraints on the R-indicator and precision. Table 4 tells us that the R-indicator can indeed be improved significantly from 0.629 to 0.905. Simultaneously, the response rate is increased considerably. The latter means that the sample size could be reduced, and, hence, the total cost, when proxy is allowed. However, again the response quality rate is not maintained, it increases considerably by 1.1%. Approach II may offer a solution. Table 5 shows that we can increase the R-indicator to 0.852 while keeping the response quality rate at 3.5%, which may be an acceptable increase. However, it is not possible to maintain the response rate; at best we get a drop of 3.3%, which is likely not acceptable. We must conclude that it is not possible to increase the R-indicator unless we are willing to lose some response rate and precision. Note that the R-indicator values in tables 4 and 5 are the same for constraints 0.80 and 0.85; the R-indicator constraint is not affecting the optimization until it increases to 0.90.

In the two design decision examples, we allowed for switches in reporting type for the same sample unit, e.g. the first visit may be self-report only, while for subsequent visits proxy reporting may be allowed. This may lead to impractical designs. As an example, table 6 displays the optimal contact protocol constraining the response quality rate to 3.5% and the budget to an average of 2.5 visits per address. The optimization leads to the following solution: For age group 15-25 one self-report visit is made, for age group 26-55 nine visits are made with a switch to proxy after the third visit, and for age group 56-65 seven visits are made with a switch after the second visit. The number of switches can be constrained to zero. We will not elaborate this here, but refer to Calinescu et al. (2012), where constraints on reporting type are discussed.

Effective survey design for nonresponse and measurement error in a setting with multiple survey variables of interest, is a complex high-dimensional optimization problem. The reduction of this dimensionality is, therefore, crucial. This

Table 4
Optimal modified response rates (Approach I) for different levels of the cost constraint on the average number of visits per address and the R-indicator when maximizing the modified response rate. The resulting response rate, R-indicator and response quality rate are given.

	Optimal adaptive survey design				
Cost constraint in average number of visits per address	R-indicator constraint	Optimal modified response rate (in%)	Response rate (in%)	R-indicator on age	Response quality rate (in%)
3	0.80	61.5	64.0	0.928	4.0
2.5	0.80	59.6	62.1	0.875	4.4
3	0.85	61.5	64.0	0.928	4.0
2.5	0.85	59.6	62.2	0.875	4.4
3	0.90	61.5	64.0	0.928	4.0
2.5	0.90	59.5	62.2	0.905	4.4

Table 5
Optimal response rates for different levels of the cost constraint on the average number of visits per address and the R-indicator, when the response quality constraint is set at 3.5% (Approach II). The resulting R-indicator and response quality rate are given.

	Optimal adaptive survey design			
Cost constraint in average number of visits per address	R-indicator constraint	Optimal modified response rate (in%)	R-indicator on age	Response quality rate (in%)
3	0.80	60.9	0.851	3.5
2.5	0.80	49.3	0.852	3.5
3	0.85	60.9	0.851	3.5
2.5	0.85	49.3	0.852	3.5
3	0.90	–	–	–
2.5	0.90	–	–	–

Table 6
Optimal design when maximizing the response rate and constraining the budget to 2.5 visits per address and the response quality rate to 3.5%. "S" = self-report only and "P" = proxy allowed.

	Optimal protocol per visit number									
Group	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
15-25	S	–	–	–	–	–	–	–	–	–
26-55	S	S	S	P	P	P	P	P	P	–
56-65	S	S	P	P	P	P	P	–	–	–

holds true for both non-adaptive and adaptive designs. However, in adaptive survey design, there is usually a more explicit focus on indirect measures for non-sampling errors. In this paper, we extend adaptive survey designs to measurement error.

We chose to reduce the dimensionality of the measurement error on the survey variables through, so-called, response quality indicators, more specifically through the propensities that they occur for different population subgroups. We compared two approaches to include response quality propensities into an adaptive survey design framework: modifying the response rate by the response quality rate and adding a constraint on the response quality rate. We strongly favor the second approach for two reasons. First, adding constraints on the response quality rates offers much more flexibility. Second, given that we may not be able to identify and employ all possible response quality indicators, it seems wiser to gradually constrain relevant response quality aspects than to treat respondents that show no lack of response quality as perfect respondents.

Our extension of adaptive survey design is novel and has two implications for survey design. First, the extension to response quality and measurement error allows for a wider range of design features that can be included in adaptive survey design. The survey mode is perhaps the most influential survey mode in terms of both quality and costs but needs a combined view on nonresponse and measurement error. Second, the use of response quality propensities allows for optimization in a way that is very similar to handling just nonresponse error in adaptive survey design; the response propensities are simply supplemented by propensities to find reduced data quality.

We ignored two methodological complications: correlations between nonresponse and measurement error, and accuracy of design input parameters. The two errors may correlate so that investigating their impact into separate indicators or constraints may be too naive. There is some literature on the interaction between the two errors, see Fricker and Tourangeau (2010), Olson (2006) and Olson (2012). Future research should look into this. Clearly, the input parameters to the optimization have a certain bias and precision given that they need to be based on historical survey data and expert knowledge. So do input parameters to non-adaptive designs, but, obviously, the level of detail is higher in adaptive designs because of the stratification. It is imperative that analyses of the sensitivity of the designs to inaccuracy in response propensities, response quality propensities and cost parameters are performed. Two types of such sensitivity analyses are useful: 1) the changes in format of the optimal design and corresponding quality and costs for variations of the parameters, and 2) the performance of the optimized design for variations of the parameters in terms of quality and costs. Variations in parameters can be achieved in many ways; one can both add noise as well as trends and seasonal patterns. Important is that the variations respect any correlation in the parameters like generally decreasing response propensities. Future research should devote attention to robustness of designs.

It is important that our study is evaluated, replicated and improved by others. The case study in this paper is based on real survey data. We made, however, some simplifications that warn against direct implementation into practice. The study is mostly meant as a demonstration of the impact of various constraints and choices, and as a stepping stone for discussion. How can our study be replicated: In the first step, a set of design features (e.g. survey modes) needs to be selected. In the second step, for each design feature, it needs to be decided whether it affects both nonresponse and measurement error. If that is the case, then, third, a list of response quality indicators needs to be made that are conjectured to be affected by some of the design features and not by others. In the paper, we mention various examples, but there are likely to be more and we recommend to consult the literature on measurement error related to the design features of interest. In the fourth step, the population needs to be stratified into relevant subgroups. It is most straightforward to start from the auxiliary variables in the weighting or nonresponse adjustment. The fifth step is to estimate response propensities and response quality propensities given historic survey data. The last step is to formulate the adaptive survey design optimization problem and to solve it numerically. Many software packages support linear optimization problems. Some software packages (like R) also support non-linear problems, which is needed when constraints like the R-indicator are included.

4 Acknowledgements

The authors would like to thank Sandjai Bhulai (VU University Amsterdam) for his constructive comments on the mathematical framework presented in the current paper. The authors also like to thank the associate editor and two referees for their useful comments.

References

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*, 8–17.

Baumgartner, H. & Steenkamp, J. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, *28*, 143–156.

Biemer, P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in nonresponse adjustment with application to field surveys. *Journal of the Royal Statistical Society, Series A*, *176*(1), 147–168.

Calinescu, M. (2013). Optimal resource allocation in adaptive survey designs. PhD Thesis, VU University Amsterdam. Retrieved from http://dare.ubvu.vu.nl/

Calinescu, M., Bhulai, S., & Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operations Research*, *226*(1), 115–121.

Calinescu, M. & Schouten, B. (2015). *Adaptive survey designs to minimize mode effects. A case study on the dutch labour force survey*. Forthcoming in Survey Methodology.

Calinescu, M., Schouten, B., & Bhulai, S. (2012). *Adaptive survey designs that minimize nonresponse and measurement risk*. Discussion paper 201224, Statistics Netherlands.

Chesnut, J. (2014). *Model-based switching from Internet to mail in the American Community Survey*. Paper presented at JSM, August 2–7, Boston, USA.

Coffey, S., Reist, B., & White, M. (2013). Monitoring methods for adaptive design in the National Survey of College Graduates. In *JSM proceedings, survey methods research section* (pp. 3085–3099). Alexandria, VA: American Statistical Association.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., & Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, *78*(3), 721–733.

Fricker, S. & Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national surveys. *Public Opinion Quarterly*, *74*(5), 934–955.

Heerwegh, D. & Loosveldt, G. (2011). Assessing mode effects in a National Crime Victimization Survey using structural equation models: social desirability bias and acquiescence. *Journal of Official Statistics*, *27*(1), 49–63.

Klausch, L., Hox, J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of ordinal rating scale questions. *Sociological Methods and Research*, *42*(3), 227–263.

Laflamme, F. & Karaganis, M. (2010). *Implementation of responsive collection design for CATI surveys at statistics canada*. Paper presented at Q2010, 3–6 May, Helsinki, Finland.

Lemaitre, G. (1988). A look at response errors in the Labor Force Survey. *The Canadian Journal of Statistics*, *16*, 127–141.

Luiten, A. & Schouten, B. (2013). Adaptive fieldwork design to increase representative household survey response. a pilot study in the Survey of Consumer Satisfaction. *Journal of Royal Statistical Society, Series A*, *176*(1), 169–190.

Lundquist, P. & Särndal, C. (2013). Aspects of responsive design for the Swedish Living Conditions Survey. *Journal of Official Statistics*, *29*, 557–582.

Medway, R. & Tourangeau, R. (2015). Response quality in telephone surveys. Do pre-paid cash incentives make a difference? *Public Opinion Quarterly*, *79*(2), 524–543.

Moore, J. (1988). Self/proxy response status and survey response quality. A review of literature. *Journal of Official Statistics*, *4*(2), 155–172.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias and total bias. *Public Opinion Quarterly*, *70*(5), 737–758.

Olson, K. (2012). Do non-response follow-ups improve or reduce data quality? A review of the existing literature. *Journal of the Royal Statistical Society A*, *176*(1), 129–145.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., & Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, *4*(1), 21–29.

Rosen, J. A., Murphy, J. J., Peytchev, A., Holder, T. E., Dever, J. A., Herget, D. R., & Pratt, D. J. (2014). Prioritizing low-propensity sample members in a survey: implications for nonresponse bias. *Survey Practice*, *7*(1), 1–8.

Saris, W. & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*, 29–43.

Schouten, B., Calinescu, M., & Burger, J. (2014). *Adaptive mixed-mode survey designs accounting for mode effects. A case study on the LFS*. Paper presented at the Joint Statistical Meetings, August 2-7, Boston, USA.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, *39*(1), 29–58.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativity of survey response. *Survey Methodology*, *35*(1), 101–113.

Schouten, B. & Shlomo, N. (2014). *Selecting adaptive survey design strata with partial R-indicators*. Discussion paper, Statistics Netherlands. Retrieved from http://www.cbs.nl

Schouten, B., Van der Laan, J., & Cobben, F. (2014). *The impact of contact effort on mode-specific selection and measurement bias*. Survey Methods: Insights from the Field, retrieved from. Retrieved from http://surveyinsights.org/?p=3629

Thomsen, I. & Villund, O. (2011). Using register data to evaluate the effects of proxy interviews in the Norwegian labor force survey. *Journal of Official Statistics*, *27*(1), 87–98.

Tourangeau, R. & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299–314.

Van Rosmalen, J., Van Herk, H., & Groenen, P. (2010). Identifying response styles: a latent-class bilinear multinomial logit model. *Journal of Marketing Research*, *47*, 157–172.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, *7*(1), 45–55.

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, *28*(4), 477–499.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Aacademy of Marketing Science*, *36*, 409–422.