

A new stopping rule for surveys

James Wagner^{*†} and Trivellore E. Raghunathan

Non-response is a problem for most surveys. In the sample design, non-response is often dealt with by setting a target response rate and inflating the sample size so that the desired number of interviews is reached. The decision to stop data collection is based largely on meeting the target response rate. A recent article by Rao, Glickman, and Glynn (RGG) suggests rules for stopping that are based on the survey data collected for the current set of respondents. Two of their rules compare estimates from fully imputed data where the imputations are based on a subset of early responders to fully imputed data where the imputations are based on the combined set of early and late responders. If these two estimates are different, then late responders are changing the estimate of interest. The present article develops a new rule for when to stop collecting data in a sample survey. The rule attempts to use complete interview data as well as covariates available on non-responders to determine when the probability that collecting additional data will change the survey estimate is sufficiently low to justify stopping data collection. The rule is compared with that of RGG using simulations and then is implemented using data from a real survey. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: stopping rule; sample surveys; responsive design; non-response bias

1. Introduction

The problem of non-response has become ubiquitous for sample surveys. Non-response rates have been steadily increasing [1–4]. Surveys that achieve 100 per cent response rates are very rare. In the presence of non-response, which has the potential to lead to bias, the decision to stop data collection is a complex one. In practice, many factors affect this decision—the completed interview and response rate targets, specific deadlines, the budget for the survey, or targets for specified subgroups. Most of these reasons for stopping do not involve the analysis of the accumulated survey data. However, non-response bias is the product of the non-response rate and the differences between responders and non-responders. Monitoring the response rate makes the strong assumption that higher response rates have a lower risk of non-response bias. Theoretically, this need not be true. In addition, recent empirical studies provide evidence that have called this assumption further into question [5–9]. A stopping rule based on the response rate ignores the accumulating survey data that could be used to inform the decision about when to stop collecting data.

A recent article by Rao, Glickman, and Glynn (RGG) [10] proposed a set of stopping rules for surveys that are based on the accumulating data. The present article builds on this result, and proposes a new stopping rule. The rules proposed by RGG examine changes in the estimate from the accumulating data. If additional data do not change the estimate, then the rules suggest that data collection can be stopped. The rule proposed here looks at the current estimate, and estimates the probability that collecting data from current non-responders with known covariates will change the current estimate. Once the probability that data from current non-responders will change the estimate becomes sufficiently small, then data collection is stopped. Such a rule would provide an improved context for survey data collections. Seeking to add the case that is easiest to interview is the logical strategy when maximizing the response rate is the goal. Seeking to interview the case that reduces our uncertainty about the remaining non-responders is the most logical strategy when satisfying a stopping rule based on the survey data is the goal.

Section 2 provides some background on current practices with regard to stopping survey data collections, as well as RGG's rules. In Section 3, a new stopping rule is proposed and compared with the rule of RGG. In Section 4, the results of a simulation study comparing this new rule with that of RGG will be presented. Finally, in Section 5, the proposed stopping rule is implemented using data from an actual survey.

University of Michigan, Ann Arbor, MI 48104, U.S.A.

*Correspondence to: James Wagner, University of Michigan, Institute for Social Research, G373 Perry, Ann Arbor, MI 48104, U.S.A.

†E-mail: jameswag@isr.umich.edu

2. Background

Stopping rules have long been a key feature of monitoring clinical trials. However, other than the article of RGG, there has been no attempt to apply this approach to surveys. In some instances, surveys have operational constraints that determine when data collection will stop. For example, a survey may have a fixed field period; or a survey may stop collecting data when the allocated budget has been spent. Often the statistic used in decisions to stop is the target response rate. In this case, the stopping rule involves the response indicator, but ignores the accumulated data on survey variables of interest. Some surveys may have response rate targets for important subgroups (for example, regions or demographic subgroups). Such stopping rules involve the response indicator and data from the sampling frame, but still ignore the data on survey variables collected so far. Groves and Heeringa [11] define an approach they call 'responsive design' that uses the accumulating survey data to modify the design. One suggestion they make is to move to another phase of data collection when the estimate derived from data collected under the current protocol stops changing. They do not, however, provide a rule for when to stop data collection.

RGG propose four rules to be used in determining when data collection should be stopped. The type of survey for which they have designed their rules is a mailed survey. They consider 'waves' of data collection; that is, each wave is a new mailing to non-responders from the previous waves. The problem, as they define it, is to determine after which wave to stop collecting data.

Their rules are defined for binary survey outcomes, Y_i , where i denotes the i th subject. The survey is conducted in waves 1 to k . The first two rules they propose do not attempt to account for non-response. They also propose two rules that attempt to account for non-response through imputation. The latter two rules are recommended by RGG. For both rules, they use a logistic regression model to develop imputations for the missing data. In the first rule, which they call Rule 3.1, one set of imputations is prepared using the data from all waves prior to the current wave (1 to $k-1$). The estimate of the proportion with $Y_i=1$ based on these imputed data sets they call \hat{P}_{k-1} . A second set of imputations is prepared using the data from all waves including the current wave (waves 1 to k). The proportion estimated using these imputed data sets they call \hat{P}_k . The rule (Rule 3.1) is that if the two sets of imputations provide different estimates of the proportion, then data collection should continue. RGG propose two different tests to see if there is a difference between these two sets of imputations. Rule 3.1a is the first test. For Rule 3.1a, the difference in the estimated proportions ($\hat{P}_k - \hat{P}_{k-1}$) is standardized and data collection is continued if the standardized difference is larger than a percentile of the standard normal distribution. They also suggest another test of this rule (Rule 3.1b) which would stop data collection if the difference in the estimated proportions ($\hat{P}_k - \hat{P}_{k-1}$) is less than a specified proportion (.01, for example). A second rule (Rule 3.2) of this type is proposed in which the first set of imputations is based on the previous wave only (wave $k-1$), and not all previous waves. The second set of imputations is based only on the current wave (wave k). In their simulations, these two rules performed nearly equally well with relatively low root mean-squared error (RMSE).

RGG's rules are flexible enough to be applied to surveys other than mailed ones. Although the rules were developed for a mailed survey conducted in waves, it could easily be applied to either a telephone or a face-to-face survey. Instead of comparing waves of response, the rule could be implemented after each call, day in the field period, or other specified points during data collection. The rules can also easily be generalized to situations other than binary variables.

A weakness of the approach is that it requires at least two waves of data collection. For the estimate from prior waves, we only include part of the data (waves 1 to $k-1$). This may lead to some loss of efficiency. In this sense, the method is retrospective. The rule stops data collection only after the latest 'wave' fails to contribute anything new to the estimate. In other words, we would have had the same estimate had we stopped at least one wave earlier.

Another weakness is that the rule assumes that a similar relationship among covariates and the survey variable obtains for waves beyond the current wave. Of course, this might not be the case and stopping might leave us with biased estimates. However, such an argument can be used against any stopping point short of a 100 per cent response rate. Very few surveys have any possibility of achieving 100 per cent response. The current approach is to specify a response rate at which to stop. This approach ignores all data except for the response indicator. It is also vulnerable to the same criticism—that the relationship between the covariates and the survey variable is different for non-responders. The model for non-response offered by RGG's Rules 3.1 and 3.2 at least attempts to place the decision to stop on a reasoned and statistical basis that uses all the available data by modeling the impact of non-response and creating a statistical rule for when to stop. A response rate, when used as a stopping rule, does neither.

In Section 3, an alternative stopping rule will be proposed. This rule will make use of all the data. Simulations will be used to compare the results to rules suggested by RGG, and then the rule will be implemented with data from an actual survey.

3. A proposed stopping rule for surveys

3.1. The 'stop and impute' rule

The rule that will be proposed is based on imputation methods. It involves the comparison of two estimates. The first estimate is the one we would have if we were to stop collecting data right now and impute the missing values. The second estimate is the one we would have if we were to continue collecting data until we had achieved a specified number of additional interviews and then we were to impute the remaining missing data. If these two estimates are likely to produce the same results, then we

should stop collecting data. For estimating the population mean under SRS, these two estimates can be denoted:

$$e_1 = \frac{\left(\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^n \hat{y}_i\right)}{n}$$

and

$$e_2 = \frac{\left(\sum_{i=1}^{n_1+n_2p} y_i + \sum_{i=n_1+n_2p+1}^n \hat{y}_i\right)}{n},$$

where y_i is the observed survey variable for case i , \mathbf{Z}_i is a vector of m covariates for case i (which is available for *all* sampled cases) the first element of which is 1 with an approximately linear relationship to y_i , \hat{y}_i is the predicted value derived from the model $\hat{y}_i = \beta \mathbf{Z}_i'$, β is a vector of estimated coefficients for a linear regression model with y regressed on \mathbf{Z} , n_1 is the current set of responders, n ($n = n_1 + n_2$) is the total sample, and p is the proportion of the remaining n_2 sampled units that is expected to be collected with continued data collection (e.g. interviews expected on the next call or next few calls). The error term ε_i in the model $y_i = \beta \mathbf{Z}_i' + \varepsilon_i$ is distributed $N(0, \sigma_\varepsilon^2)$.

In this setting, a rule can be used to determine when to stop data collection. In order to do so, we evaluate the probability that these two estimates are the same. The investigator must specify a value δ that is acceptably small such that a difference between e_1 and e_2 that is less than this value is not meaningful. If the probability

$$Pr(|e_1 - e_2| < \delta | \mathbf{Z}, p, \beta, \sigma_\varepsilon)$$

is sufficiently large, then option e_1 —where we stop data collection now and impute the remaining missing values—is preferred.

In order to determine this probability, the variance of $e_1 - e_2$ is needed. The difference between e_1 and e_2 can be expressed as

$$\frac{\sum_{i=n_1+1}^{n_1+n_2p} \hat{y}_i - \sum_{i=n_1+1}^{n_1+n_2p} y_i}{n}.$$

This is the difference between the imputed and observed values for the n_2p cases we are considering adding to our data. If we substitute the regression equation for the predictions and the mean value for the cases that would be interviewed, then we have the following reformulation of the previous expression:

$$\text{Var} \left(\frac{\sum_{i=n_1+1}^{n_1+n_2p} (\beta \mathbf{Z}_i') - n_2p\bar{y}}{n} \right)$$

The next step is to note that the expected value of the quantity in parentheses is 0. If we can derive the variance of the quantity in parentheses, then we can estimate the probability that this quantity is close to zero. If we denote the cases that would respond (i.e. the n_2p cases) with further effort using the subscript wr and the estimate of β derived from the current set of respondents, then this can be shown to be equivalent to the following:

$$\frac{n_2p}{n} (\hat{\beta}_r \bar{\mathbf{Z}}_{wr} - \bar{y}_{wr}).$$

The variance of the difference between e_1 and e_2 is:

$$\hat{\sigma}_\varepsilon^2 \left(\frac{n_2p}{n} \right)^2 \left[(1 + \bar{\mathbf{Z}}_{wr}' (\mathbf{Z}_r' \mathbf{Z}_r)^{-1} \bar{\mathbf{Z}}_{wr}) + \frac{1}{n_2p} \right].$$

This can be described as the prediction variance for the cases that would be collected plus the conditional variance of the mean of y . This variance estimate for the difference between e_1 and e_2 can be used to standardize the results for comparison with the standard normal distribution. If the probability

$$Pr(|e_1 - e_2| < \delta | \mathbf{Z}, p, \beta, \sigma_\varepsilon)$$

is large enough, then data collection can be stopped. In other words, if the probability that additional data (n_2p cases) will not substantially change our current estimate—conditional on our model, the covariates, and the residual variance—is very high, we should stop collecting data.

This rule is cost efficient as it allows us to stop data collection when the imputation model is precise enough that we can have nearly the same certainty in our estimate as we would if we were to collect a specified number of additional cases. The rule does, however, assume a model that relates the covariates \mathbf{Z} to y . If this model is incorrect, or if the estimated coefficients $\hat{\beta}$ are different among responders than among non-responders, then it is possible that we will stop too early.

One practical issue is determining which n_2p cases the researcher believes will be collected. This can be done by taking a random sample of the remaining n_2 cases, or by sampling using an estimated probability of response to select a sample of cases likely to respond given a specified protocol (this protocol could be tailored to the case in order to improve the probability of response among relatively 'difficult' to interview sampled units). Another practical issue is how to set p . This can be set either by

empirical observation from other surveys about the expected number of interviews to be completed with additional effort, or as a number that is meaningful for continued data collection.

This rule is certainly more focused on the risk of bias than on meeting targets for sampling error. Although it is true that the rule requires some minimum size in order to decide when to stop, this minimum can vary quite a lot depending on the specific interrelationships among y , Z , and the propensity to respond (this will be seen in the simulations in Section 4). One simple solution is to have a second rule that the sampling error (estimated using multiple imputation-appropriate methods) must be within a specified limit before stopping. Then both this rule about sampling error and the ‘stop and impute’ rule must be met before stopping.

Working out the required sample size before collecting any data is surely more difficult in this circumstance. Simulations may be helpful in this regard. Simulations could be used to determine what sample sizes might result under various assumptions about the interrelationships of y , Z , and the propensity to respond. These simulations could help project a range of outcomes. Then the investigators can vary the value of the parameters (δ and the cutoff probability for defining a ‘high probability’ that $e_1 - e_2$ is small). These parameters can be set to help tune the expected sampling error.

3.2. Simulations

The simulations will emulate the conditions of the simulations of RGG—that is, a mailed survey conducted in waves (k). There are two key conditions that impact our estimates of the survey variable (y). The first is the relationship between the propensity to respond (here denoted r) and y . If the two are correlated, this can lead to bias when non-response occurs. Data collected at different waves are likely to lead to different estimates of \bar{y} . The second condition is the relationship of covariates (z), which are available for all sampled units, to the survey variable y . If the propensity to respond (r) and y are correlated, but we still have a covariate (z) that predicts well the survey variable, then we may still be able to adjust our estimates (either through imputation or weighting) with this variable and still produce unbiased estimates. If, however, the correlation of z and y is confounded with r , then we are back in the difficult situation where z is not useful for adjustment purposes.

Following RGG, a simulation study was developed to test the performance of this rule while varying these two relationships—the correlation between the z and y variable and the relationship between z and the propensity to respond r . One thousand data sets were generated under the cross-classification of the following conditions:

- (i) Correlations of z and y either does not or does depend on r ,
- (ii) Wave of response (r) either does not or does depend on z .

Crossing these two conditions produces the four simulations detailed in Table I.

The goal has been to follow the key features of the simulation study of RGG in order to be able to compare the results of their stopping rules to those of the rule proposed here. RGG’s rules were also generalized to the normally distributed variable to which the ‘stop and impute’ rule applies. In these simulations, z is normally distributed with a mean of 10 and a standard deviation of 1. The propensity to respond r is formulated (following RGG) as the wave of response, which is modeled as a Poisson distribution with a mean of 1 when r does not depend on z . The waves are numbered starting with 0 as the first wave. In the condition where r does depend on z , r has a mean of 1 when $z < 10$ and a mean of 5 when $z \geq 10$. When the z - y correlation does not depend on r , it is fixed at a single value for all waves. When the correlation does depend on r , the correlations increase with the wave of response. In addition, in order to demonstrate the impact of the variability of z , z is also simulated with different standard deviations. As the ‘stop and impute’ rule depends on the choice of a suitably small δ , a δ of 1 per cent of the mean was chosen. The impact of the magnitude of δ on the stopping wave will be considered further in the fourth simulation. In addition, if the probability that the difference between e_1 and e_2 is less than δ is greater than .95, data collection was stopped. In order to ensure comparability to RGG’s rules, p was selected such that n_{2p} was equal to the number of respondents at the next wave. Another important choice is the number of imputations. In the simulations reported here, 5 imputations were used. This follows what RGG did in their simulations. Rubin [12] suggested that if the fraction of missing information was small enough, then 5 imputations should be a sufficient number. In practice, if the fraction of missing information is not small, then a larger number of imputations may be needed [13].

3.2.1. Simulation 1. In the first simulation, r does not depend on z , and the correlation of z and y is independent of r . This corresponds to the missing completely at random assumption described by Little and Rubin [14]. Under these conditions, the responders are effectively a random sample of the total sample (including non-responders). This is shown by Figure 1, which

Simulation	Wave of response (R)	Correlation Z, Y	Type of missingness
1	Does not depend on Z	Does not depend on R	MCAR
2	Depends on Z	Does not depend on R	MAR
3	Does not depend on Z	Depends on R	NMAR
4	Depends on Z	Depends on R	NMAR

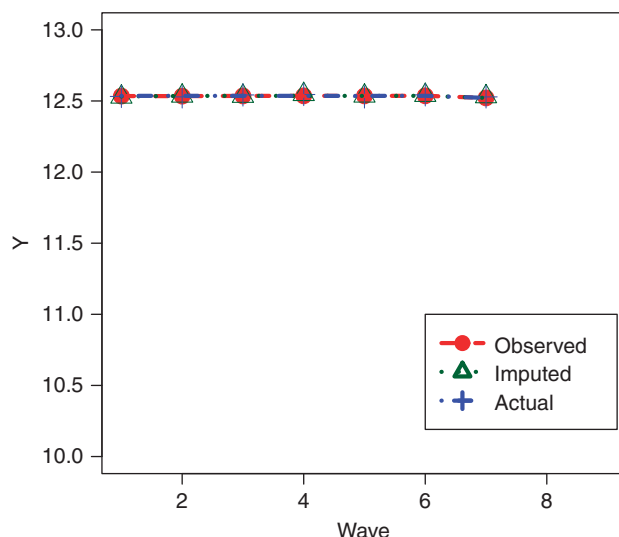


Figure 1. Observed, imputed, actual mean Y for simulation 1.

Table II. Simulation 1 results.			
	RGG 3.1a	RGG 3.1b	Stop and impute rule
Mean number of waves	2.10	2.18	1.22
Mean number of waves std dev	0.30	0.39	0.41
Var(\hat{y})	0.007	0.007	0.012
Mean sample size	150.7	154.0	90.9
Bias(\hat{y})*100	-0.03	-0.03	-0.14
RMSE(\hat{y})*100	8.28	8.19	11.14
Coverage	99.9	99.9	98.0

plots the observed \bar{y} , the fully imputed \bar{y} , and the true \bar{y} by wave. The three points are virtually identical at each wave. The wave two respondents are equivalent to the wave one respondents, and so on.

Table II presents the results of 1000 simulations of these conditions (y depends only on z , r does not depend on z). One thousand data sets were created at each of nine different correlations between z and y . In the table, only the correlation for .1 is presented as stronger correlations have a similar impact across all three rules (see Wagner [15] for additional details). The row 'mean stopping wave' reports the mean number of waves completed before stopping. The next row is the standard deviation of the stopping wave. The next row is the variance of \hat{y} . This variance is calculated using methods for combining multiple imputations [14]. The bias of the estimate is multiplied by 100, as is the RMSE. The final row is the proportion of the time that the 95 per cent confidence intervals attain the nominal coverage of the population mean.

In this situation, it should be clear that stopping earlier is preferred as the outcome variable is uncorrelated with the response propensity and the z variable is uncorrelated with the wave. Under these conditions, the 'stop and impute' rule performs better than either rule of RGG. Mainly, this is because RGG's rules require at least two waves of data collection, whereas the 'stop and impute' rule can stop after the first wave. As the correlation between z and y increases, this early stopping after the first wave is more likely to happen.

3.2.2. *Simulation 2.* In the second simulation, the wave of response (r) is a function of z . In the simulation, this effect was implemented by having r drawn from a Poisson distribution with mean 1 when $z < 10$. When $z \geq 10$, then r was drawn from a Poisson distribution with mean 5. In this case, there is a risk of bias. This can be seen from Figure 2. The observed \bar{y} are different than the true \bar{y} . Fortunately, the z variable allows us to correct for this bias through the imputation model that correctly relates z and y even at wave 1. This is why the actual and impute lines are nearly identical. These conditions correspond to the missing at random assumption described by Little and Rubin [14]. One thousand data sets were created under these conditions (r depends on z , the correlation of z and y does not depend on r). The results are presented in Table III.

In this situation, RGG's rules are more conservative. They generally result in a smaller bias of the survey statistic than the 'stop and impute' rule, but the bias is still relatively small for the 'stop and impute' rule. However, it is large enough that the nominal coverage for the 95 per cent confidence interval is not quite attained. Again, the 'stop and impute' rule stops much earlier than either of RGG's rules.

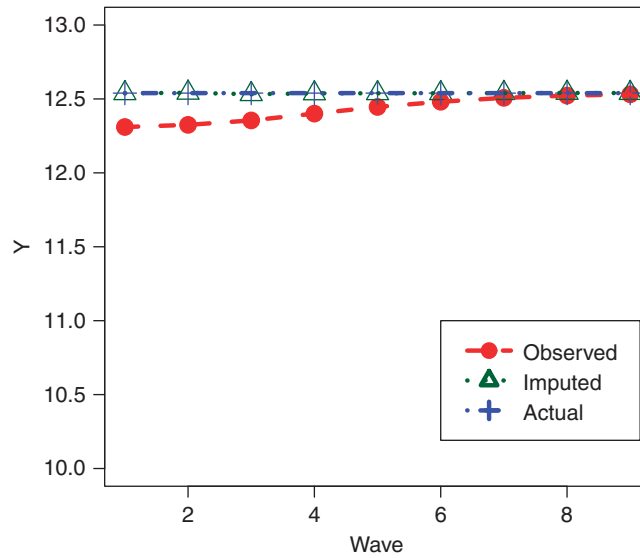


Figure 2. Observed, imputed, actual mean Y for simulation 2.

Table III. Simulation 2 results.			
	RGG 3.1a	RGG 3.1b	Stop and impute rule
Mean number of waves	2.52	2.78	1.41
Mean number of waves std dev	0.70	0.81	0.49
Var(\hat{y})	0.024	0.022	0.053
Mean sample size	91.1	97.1	54.9
Bias(\hat{y})*100	-0.17	-0.36	-0.09
RMSE(\hat{y})*100	15.57	14.72	22.98
Coverage	94.6	95.2	91.8

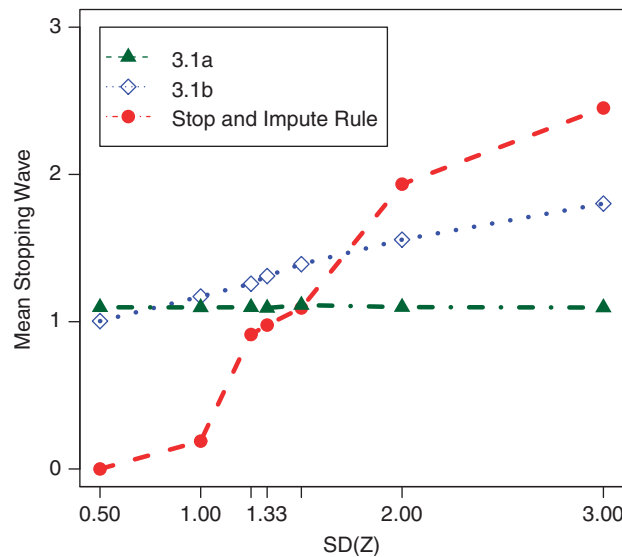


Figure 3. Mean stopping wave by SD(Z).

As the 'stop and impute' rule is more directly dependent on the variance of z , this parameter setting can make a large difference in when stopping will occur. Figure 3 shows the impact of different standard deviations of z on the mean wave of stopping. With a mean of 10, a standard deviation of 1 still represents a large variation relative to the mean.

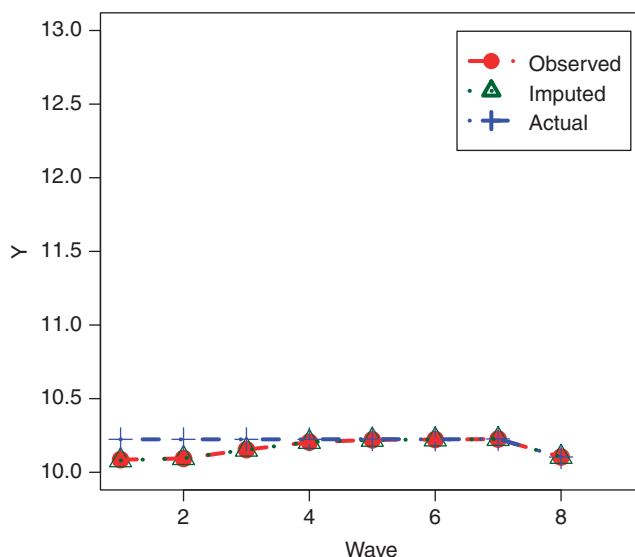


Figure 4. Observed, imputed and actual mean Y for simulation 3.

	RGG 3.1a	RGG 3.1b	Stop and impute rule
Mean number of waves	2.09	3.14	3.88
Mean number of waves std dev	0.30	0.91	0.35
Var(\hat{y})	0.063	0.053	0.047
Mean sample size	150.6	178.3	195.3
Bias(\hat{y})*100	-12.3	-7.00	-2.28
RMSE(\hat{y})*100	27.98	23.98	21.70
Coverage	99.4	99.8	100.0

It is clear from the figure that the ‘stop and impute’ rule is much more susceptible to variation in z . This should be expected as the rule directly incorporates this variation. RGG Rule 3.1a involves the difference in the imputed means between two waves. This difference is very insensitive to variation in z . RGG Rule 3.1b, on the other hand, is somewhat sensitive to variation in z . In sum, high variances on the covariate may reduce its utility for the ‘stop and impute’ rule.

3.2.3. *Simulation 3.* Simulation 3 is the situation where r does not depend on z , but the correlation of z and y is a function of r . This simulation was implemented by specifying the correlation for each value of r (0.01, 0.01, 0.04, 0.1, 0.1, 0.2, 0.2, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3). The correlation of z and y is an increasing function of r . Figure 4 depicts this situation. The imputed and observed \bar{y} depart from the true value until the response rate becomes relatively high. This is the not missing at random assumption (at least for the first four waves) described by Little and Rubin [14].

The results of this simulation are presented in Table IV. As the correlation of z and y varies with r , different correlations (e.g. 0.1 to 0.9) were not tried as they were with simulations 1 and 2.

In this situation, it appears that the ‘stop and impute’ rule and RGG 3.1b are more conservative and produce less biased estimates of \bar{y} . The RMSE of the estimate using the ‘stop and impute’ rule is lower than either of the other rules. If limiting bias is more important than the overall RMSE, the more conservative rule might be preferred for this situation. However, it is also much more costly, requiring on average 1.8 more calls than RGG Rule 3.1a.

3.2.4. *Simulation 4.* In the final simulation, the wave of response (r) depends on z and the correlation of z and y is a function of r . Figure 5 shows the observed, imputed, and true \bar{y} by wave for this simulation. Again, these conditions produce the not missing at random situation.

The bias is a function of the wave. Figure 6 shows the average proportionate bias over the 1000 simulations at each wave for the imputed mean (as opposed to the mean derived from the observed data alone). From Figure 6 it can be seen that the bias is still relatively high until after the sixth wave (wave 5). The results for the three stopping rules under simulation 4 are presented in Table V.

In this situation, the bias begins to dominate the RMSE for RGG Rule 3.1a. The ‘stop and impute’ rule does much better in terms of the bias and is the only rule to attain anything close to the nominal coverage. Of course, much greater effort is required.

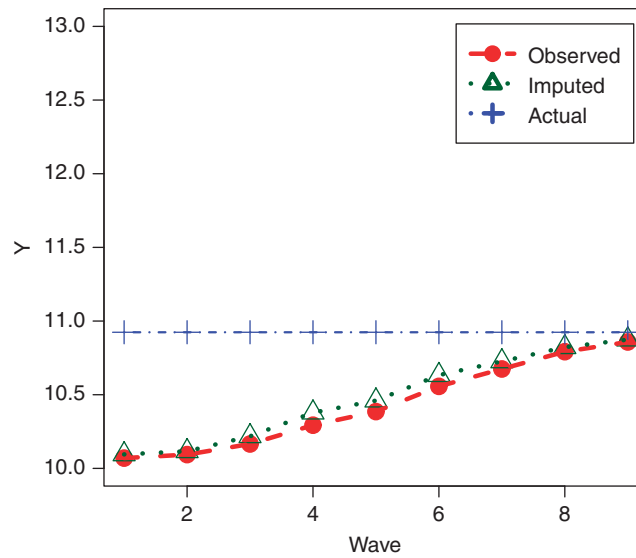


Figure 5. Observed, imputed and actual mean Y for simulation 4.

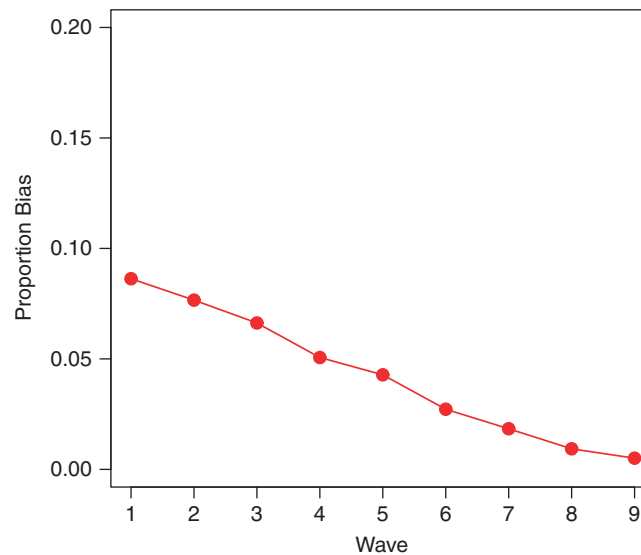


Figure 6. Simulation 4 proportion bias by wave.

Table V. Simulation 4 results.			
	RGG 3.1a	RGG 3.1b	Stop and impute rule
Mean number of waves	2.56	5.00	7.94
Mean number of waves std dev	0.75	1.94	0.75
Var(\hat{y})	0.231	0.107	0.052
Mean sample size	91.7	139.6	186.8
Bias(\hat{y})*100	-76.29	-45.24	-10.41
RMSE(\hat{y})*100	90.20	55.86	25.06
Coverage	59.1	72.9	93.4

As the choice of δ can have a large impact on when stopping occurs, Simulation 4 was run with various values for δ . As the expected mean of y was approximately 11, the values of δ were 0.1, 0.2, 0.3, 0.5, 1.0, 2.5, and 5.0. Figure 7 presents the results of these simulations.

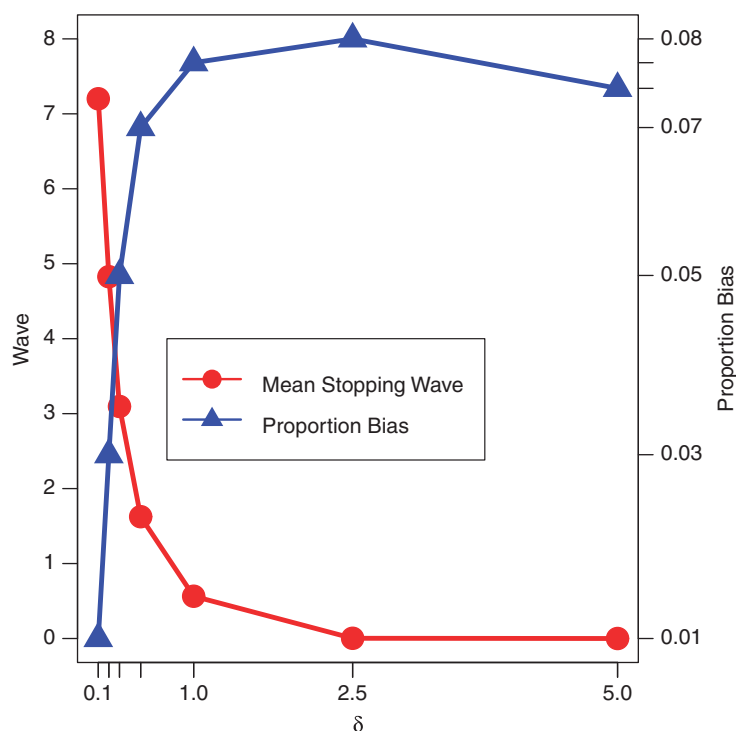


Figure 7. Mean stopping wave and proportion bias by value of δ .

In general, in comparing the ‘stop and impute’ rule to RGG Rules 3.1a and 3.1b, the rule proposed here is more efficient in the situations that are MCAR. In the MCAR situation, the ‘stop and impute’ rule can be very efficient and truncate effort earlier than either of RGG’s rules considered here. When the situation is not MCAR, the ‘stop and impute’ rule appears to be conservative. It may be overly conservative, depending on the relative importance of the bias and variance components of the mean-squared error. The ‘stop and impute’ rule is more protective against potential bias, but generally requires more effort when bias is a risk.

4. Implementation

This approach was implemented with data from the survey of consumers (SCA). The survey collects 300 random-digit dial (RDD) interviews each month. It typically attains an AAPOR response rate 2 of between 42 and 45 per cent. SCA asks respondents for their views on the state of their personal finances and the economy in general. A key statistic produced by this survey is the index of consumer sentiment (ICS). This Index is widely reported and has been found to be highly predictive of economic trends [16].

The ICS is the survey variable (y). The mean ICS for the month analyzed here is 91.46. The r variable in this case is the call number of the completion. The correlation between the ICS and the call number is practically zero (-0.01).

There are several variables on the frame that are related to the ICS. The RDD sample for SCA is generated using a commercially available telephone sampling system. This system associates every telephone exchange with a geographic location using telephone listings. Then Census data of the associated geographic location can then be attached to the telephone number. Of course, with the portability of telephone numbers, this estimate of the geographic location of the number may be wrong. However, only a small proportion of numbers have been ported and these estimates of geography continue to be fairly accurate [17]. In addition to the variables supplied by the sample vendor from Census data, data from the Bureau of Labor Statistics on employment, wages, and prices were added. These data included monthly unemployment rates at the county level, quarterly reports of average weekly wages at the county level, and the monthly reports of the consumer price index (CPI) at the core-based statistical area (CBSA) for larger areas or region and MSA status for smaller areas. As the ICS is a measure of consumers’ views on the economy, these economic measures were expected to be related to the ICS. Curtin [16] presents evidence that this is the case. Curtin suggests that interest rates are also important predictors. These are not included in these models as it is more difficult to obtain these data at a local level. In addition, Curtin notes that the change from one time period to the next in these rates is also predictive of consumer sentiment. Therefore, changes in unemployment rates, average weekly wages, and the CPI from the quarter or month prior to the month of the interview to the current month or quarter were also included.

An important difference between these data and the simulation data has to do with the frame quality. As the SCA uses an RDD frame, there are numerous blanks (non-household, non-working numbers) on the frame. Some of these non-sample cases will never be identified. This is because telephone companies often allow unassigned numbers to ring through as if they were working. These numbers will ring instead of playing a message indicating that the number is not working. Other numbers will

Table VI. Variables used in impute model.	
Variable	
Per cent exchange listed	
Per cent age 18–24	
Per cent age 65+	
Per cent income \$100,000+	
Per cent owner occupied	
Per cent black	
Per cent Hispanic	
Log(median HH income)	
Household density (households per 1000 sq ft)	
Unemployment rate (county)	
CPI (CBSA/MSA status by region)	
Change in CPI (CBSA/MSA status by region)	
Average weekly wages (county)	

Table VII. Stopping rule results using various delta values.					
δ	Stopping call	Est. ICS	VAR(ICS)	Non-response rate	Completes
1	21	91.00	6.3	0.57	266
2	14	89.70	8.0	0.62	241
3	9	91.61	7.9	0.66	215
4	8	91.87	9.3	0.68	206
5	5	93.00	10.8	0.73	173
6	5	93.20	12.5	0.73	173
7	5	93.09	12.2	0.73	173
8	5	93.47	12.2	0.73	173
9	5	92.52	12.6	0.73	173
10	2	91.93	22.5	0.84	112

be identified as non-working or non-household only after repeated attempts. Therefore, the total sample size changes at each call as ineligible telephone numbers are identified.

A subset of the available frame and paradata variables (Table VI) was selected using a stepwise regression modeling approach. Variables were allowed to enter the model if they had a p -value less than 0.3. The final model had an R^2 value of 0.105. A parsimonious model was preferred as it would need to be used with continually updated data sets.

The survey designer has two important choices to make when implementing this stopping rule. The first decision is the probability at which e_1 and e_2 are judged to produce the same estimate. For these analyses, data collection is stopped if the probability that the difference between e_1 and e_2 is less than δ is greater than 0.95. The second decision involves δ . A value needs to be chosen that is small enough such that as long as the difference between e_1 and e_2 is smaller than this value it can be considered to be unimportant.

It may also be useful to consider the number of imputations. In this application, the fraction of missing information is not small—greater than 50 per cent. Graham *et al.* [13] explored the number of imputations needed when the fraction of missing information is large. They suggested as many as 100. Therefore, for this analysis, 100 imputations were used.

The results are reported in Table VII. With $\delta=1$, the rule would stop data collection after 21 calls and collects 266 interviews. The estimate of the ICS would be 91.00—very close to the estimate from all 297 interviews (91.46). The difference between the two estimates is within δ ($91.46-91.00=0.46$, which is less than 1.0). As the rule is not based on the magnitude of change between calls, the small difference between the estimates achieved at 8 and 9 calls does not lead to stopping data collection.

Although these savings are moderate, the important result is that in this application the stopping rule identified a stopping point that was not arbitrary. It was based on the data. In fact, it was based on the data from all the interviews and the complete data from covariates on the sampling frame.

5. Conclusion

Stopping rules have long been used in clinical trials where there is an ethical demand that a trial be stopped when one treatment is clearly better than another. Data monitoring committees are charged with analyzing the accumulating data to determine when a statistically defined stopping rule has been satisfied. Surveys, on the other hand, have relied on data other than the accumulated data on survey variables of interest to decide when to stop. This might include targets for response rates, number of interviews,

budget, timelines, or response rate for important subgroups. Data collection is stopped when the targets are met. In fact, the decision to stop collecting data is complicated by the problem of non-response. This non-response can potentially bias survey estimates.

If we are to build adaptive designs for surveys, we will need some means for determining when to stop collecting data. The current practice does not usually consider the data available on the frame and their relationships to the accumulating survey data. These data may help inform us about the risk of non-response bias. A stopping rule that monitors these frame data and the accumulating survey data is preferred. A rule that attempts to account for risk of bias due to non-response would perform better across a variety of situations. Perhaps such a stopping rule could be an enhancement to responsive designs. The rule could be applied to key subgroups of a survey population (as long as those subgroups are defined by information available on the sampling frame); or we might use the stopping rule to focus on cases that balance the distribution of covariates to a known or desirable population distribution.

This article has proposed such a rule. This rule is based on imputation methods for dealing with non-response bias. In the presence of variables that are correlated with the survey variable, it is possible for this rule to be protective against bias. In contrast to rules proposed by other authors, the 'stop and impute' rule makes use of all the available data to determine whether to stop. This increases the efficiency of the rule. In simulations, this rule proves to be quite robust. In the most favorable situations (i.e. when data are missing completely at random), it also tends to be efficient compared to other rules proposed for surveys. An application to a real survey shows that this rule is feasible if focused on the risk of bias. However, it is a simple modification to accommodate setting targets for sampling error as well.

References

1. de Leeuw E, de Heer W. Trends in household survey nonresponse: a longitudinal and international comparison. In *Survey Nonresponse*, Groves RM, Dillman DA, Eltinge JL, Little RJA (eds). Wiley: New York, 2002; 41–54.
2. Atrostic BK, Bates N, Burt G, Silberstein A. Nonresponse in US government household surveys: consistent measures, recent trends, and new insights. *Journal of Official Statistics* 2001; **17**(2):209–226.
3. Petroni R, Sigman R, Willimack D, Cohen S, Tucker C. Response rates and nonresponse in establishment surveys BLS and Census Bureau. *Presented to the Federal Economic Statistics Advisory Committee* 2004; 1–50.
4. Curtin R, Presser S, Singer E. Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly* 2005; **69**(1):87–98.
5. Keeter S, Miller C, Kohut A, Groves RM, Presser S. Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly* 2000; **64**(2):125–148.
6. Curtin R, Presser S, Singer E. The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly* 2000; **64**(4):413–428.
7. Merkle DM, Edelman M. Nonresponse in exit polls: a comprehensive analysis. In *Survey Nonresponse*, Groves RM, Dillman DA, Eltinge JL, Little RJA (eds). Wiley: New York, 2002; 243–257.
8. Keeter S, Kennedy C, Dimock M, Best J, Craighill P. Gauging the impact of growing nonresponse on estimates from a national rdd telephone survey. *Public Opinion Quarterly* 2006; **70**(5):759–779.
9. Groves RM, Peytcheva E. The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly* 2008; **72**(2):167–189.
10. Rao RS, Glickman ME, Glynn RJ. Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine* 2008; **27**(12):2196–2213.
11. Groves RM, Heeringa SG. Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2006; **169**(3):439–457.
12. Rubin DB. *Multiple imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
13. Graham J, Olchowski A, Gilreath T. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 2007; **8**(3):206–213.
14. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: Hoboken, 2002.
15. Wagner J. Adaptive survey design to reduce nonresponse bias. *Ph.D. Dissertation*, University of Michigan, 2008.
16. Curtin R. Consumer sentiment surveys: worldwide review and assessment. *Journal of Business Cycle Measurement and Analysis* 2007; **3**(1):9–45.
17. Johnson TP, Cho YIK, Campbell RT, Holbrook AL. Using community-level correlates to evaluate nonresponse effects in a telephone survey. *Public Opinion Quarterly* 2006; **70**(5):704–719.