Combining Information from Multiple Modes to Reduce Nonresponse Bias

Mick P. Couper[1], Andy Peytchev[1], Roderick J. A. Little[1], Victor J. Strecher[1], and Kendra Rothert[2]
University of Michigan[1]
Care Management Institute, Kaiser Permanente[2]

Abstract

Over 3,000 subjects were recruited in 3 U.S. regions for a randomized experiment of an online weight management intervention. Participants were sent invitations to web survey reassessments after 3, 6, and 12 months. High and increasing nonresponse to the three follow-up surveys created the potential for nonresponse bias in key program outcomes. A subsample of the nonrespondents at the one-year follow-up was selected for a nonresponse study. This subsample was then randomly assigned to a short telephone or mail survey. This was done in order to evaluate cost efficiency, differential effectiveness of mode combinations in reducing nonresponse bias, and measurement differences by mode. The responses from the nonresponse study were then to be added to the baseline measures and used in an imputation model. Differences between the telephone and mail survey reports posed an added methodological problem, allowing further exploration of sensitivity of the results not just to nonresponse, but also to the mode used in the second stage through comparison of different imputation models. Implications are discussed for cost, nonresponse bias, measurement differences, and post-imputation variance estimates.

Keywords: Survey nonresponse, Web, Telephone, Mail, Imputation, Mode effects.

1. Introduction

To the extent that the probability to respond to a survey request is associated with measures of interest, inferences from the survey will be erroneous. One such example is when the effectiveness of an experimental intervention affects the propensity to respond to a follow-up survey. This can lead to misleading conclusions both about the difference between the experimental and control groups, and the actual direction and magnitude of effect.

Methods have been developed to address nonresponse through study design and through analysis. One is to incorporate a two phase study design (Hansen & Hurwitz, 1946; Deming, 1953), in which the nonrespondents from the initial effort are subsampled and key survey design features altered to increase both contact and response propensities, in order to gain measures on those who did not respond in the first phase. The key design features that are often implemented in the second phase include respondent incentives and change of data collection method. Under the stochastic model for survey nonresponse (Deming, 1953), respondents have response propensities that are conditional on the study's essential survey conditions (Hansen, Hurwitz, & Bershad, 1961). To the extent that changes in the response propensities from the altered essential survey conditions in the second phase are associated with survey variables, nonresponse bias in estimates could be reduced. That is, respondents with very low response propensities in one mode can have much higher propensities in another mode. We also have evidence that incentives increase the response propensities of those not interested in the topic (Baumgartner and Rathbun. 1997; Groves, Singer, & Corning 2000; Groves, Presser, & Dipko, 2004).

These essential survey conditions are comprised of a large number of design possibilities. However, there are three study objectives that have to be simultaneously considered: minimizing cost, nonresponse bias, and measurement error.

Even after these decisions are made, data from the second phase needs to be combined with the data from the first phase. The objectives are to obtain unbiased and efficient estimates, but the various methods for combining the data differ in the assumptions they make about the data and vary in their complexity. Hence they should have different sensitivity to the assumptions that are made.

A common use of data with unit nonresponse is complete case analysis. Under this crude method, nonrespondents are ignored under the assumption of missing completely at random (MCAR) – no association between the outcome of a case and any of the variables of interest.

Another method is the last observation carried forward (LOCF). This involves taking the last observation of the case in studies where multiple measurements are collected over time. This model assumes that there is no change subsequent to the last measure regardless of how many periods are missed or who those nonrespondents are.

A more conservative model is, for those respondents who are missing the last measurement, to assume that it is the same as it had been at baseline (time 1), i.e., no change since the intervention (Implicate in this model is the assumption that participants fail to respond because the desired effect was not achieved)

Weighting models make use of auxiliary data to adjust estimates, requiring fewer assumptions about the missing data, or assuming missingness at random (MAR) within subclasses. One way to use weights is to create subclasses from variables that are associated with the dependent variables and assert within-class homogeneity – nonrespondents are in expectation the same as the respondents within each weighting cell.

One way to incorporate the respondents from a second phase sample is by weighting up the second phase respondents to represent all the nonrespondents. This way the nonrespondents don't have to be similar to the respondents in the first phase, just similar to the respondents in the second phase (the sample of nonrespondents). Like all subclass weighting approaches, it is limited to the number of variables that can be used to define the subclasses, as cross-classification of numerous variables results in small cells, which translates into unstable estimates and sizeable increases in variance from the weights. It is also constrained to the use of variables that are available for (all the cases)

A variant on weighting class adjustment is the use of propensity weights, which weight by the inverse of the predicted response probabilities. This allows the use of more variables. and interactions between variables, than the other weighting methods, but as long as the dependent variable is the response outcome, they are in the same family of adjustment methods – adjustment for the *nonresponse rate*.

All these methods can be seen as special cases of imputation – imputing the mean, imputing zero change, imputing the last observation, or imputing the means of various subclasses. A more statistically sophisticated approach uses information from complete and incomplete variables and their entire covariance structure to impute for the missing cases. One such method is sequential regression multiple imputation, which uses all available variables, and begins the imputation with those with the least data missing, which can then in turn be used in the imputation of the following variable. In addition, multiple imputation allows for incorporation of the level of uncertainty in the imputed values by reflecting it in how much imputed values vary across independent imputations of the missing data (Little & Rubin 1987, Rubin, 1987). For a review of nonresponse adjustment methods, see Kalton & Kasprzyk (1986).

Another important aspect of the multi-stage data collection designs is gaining insight into why the nonresponse occurred in the first phase to inform future studies, not just whether the nonrespondents were different from the respondents.

2. Study Design

Approximately 4,000 people in four separate regions in the U.S. were recruited by Kaiser Permanente to participate in an online weight management program. Our study focuses on 3,260 subjects from 3 regions, as one region did not participate in the nonresponse follow-up survey. All subjects completed a detailed baseline survey online, including measures of motivation to lose weight, self-efficacy, environmental factors, particular threats to diet behavior, etc. Subjects were randomly assigned to one of two conditions, a tailored expert system (treatment) or the standard information-only website (control).

Three months after the baseline survey, all participants were invited to complete a follow-up web survey, hoerl@crd.ge.com, containing many of the measures collected at baseline, such as height and weight, motivation to lose weight, etc. Follow-up surveys were conducted again at 6 and 12 months after baseline, with a decreasing number of respondents.

Although participants could miss an earlier follow-up and still respond to a later one, this did not occur often, and the nonresponse (attrition) was generally monotonic. (The follow-up surveys and response to them are presented in

Figure 1). The initial sample was 3260 respondents in 3 geographic regions in the United States[1], of which two-thirds did not do any of the follow-up surveys, 369 did the 3-month follow-up but not the 6- and 12-month, 291 did the 6-month but not 12-month, and 499 did the 12-month survey.

Last Observed				NRFU	
Baseline (3260)	3 Mo. (3260)	6 Mo. (3260)	12 Mo. (3260)	Mail (398)	Phone (300)
			499	4	4[2]
		291	▨	52	40
	369		▨	53	45
2101			▨	84	66
3260†				193	155

† This row is the total number of respondents to each survey.

Figure 1: Response to the Follow-up Surveys (3-, 6-, and 12-months) and the Nonresponse Follow-up (NRFU) Study.

The main purpose of the follow-up surveys was to track short-term and long-term changes in weight or BMI in order to evaluate the effectiveness of the tailored method of providing weight management information. Looking at those who responded, the treatment seemed to be effective in reducing weight/BMI. However, the relatively high and increasing (with each additional follow-up survey) nonresponse poses a threat to the validity of the results from the study.

A nonresponse follow-up (NRFU) survey was conducted about 18 months after baseline. Since one of the main hypotheses for reasons for nonresponse was contactability via e-mail, the main design change for the nonresponse study was to change the method of data collection. Participants who had provided both telephone numbers and mailing addresses were stratified by whether they participated only in the Baseline, also did the 3-month, 6-month, or 3- and 6-month follow-up surveys. Those with no follow-up were subsampled, while those with at least some follow-up were selected with certainty. In total, 698 nonrespondents were selected.

In order to test which method of data collection is more efficient (cost per sample element and cost per respondent), effective (response rate), and possibly provides better measurement, a mode experiment was conducted. Of the 698 selected respondents, 300 were randomly assigned to telephone interviewing and 398 to mail questionnaires. The mail sample was sent $5 with the questionnaire. The questionnaire was short with 9 questions and 4 subquestions, asking whether they recall receiving the follow-up survey invitations by e-mail, their current e-mail addresses, reasons for not responding, motivation and satisfaction with losing weight and with the program, and their current weight.

There are two main methodological components to this study. The first is to obtain unbiased estimates of the effect of treatment, using information collected on the nonrespondents. The treatment effect can be viewed as Intent to Treat (ITT) rather than Treatment of the Treated (TOT), as information on exposure to treatment in the control group was not recorded. To do so, different types of nonresponse adjustment techniques are compared. As noted earlier, the methods vary in their complexity, with less complex methods making more stringent assumptions about the nonresponse mechanism and a goal in this study is to demonstrate how sensitive results are to these assumptions.

At the very minimum, if the nonresponse mechanism is indeed MCAR, then response propensities in two randomly assigned groups should be the same – if they are not, there is reason to believe that the experimental treatment has altered the nonresponse properties. Response propensities were estimated using available baseline measures and experimental assignment. While the probability of responding to at least one of the three follow-up surveys in the treatment and control groups was not different, the distributions of the response propensities were significantly different as shown in Figure 2 (Wilcoxon-Mann-Whitney-U and Kolmogorov-Smirnov tests, p<.05). There are characteristics that are associated with high likelihood of responding in the treatment but not (as much) in the control condition, while the rest of the respondents in the treatment group are somewhat less likely to respond than their counterparts in the control group. Correlations between the response propensities and reduction in BMI for those who responded to the 12-month follow-up are not significant, but there is an indication that

[1] We excluded one of the regions from the nonresponse follow-up study because of delays getting IRB approval.

[2] Thirteen cases who responded to the 12-month follow-up were sampled by mistake.

they could be different in the two groups (-.133 in the control and -.055 in the treatment).

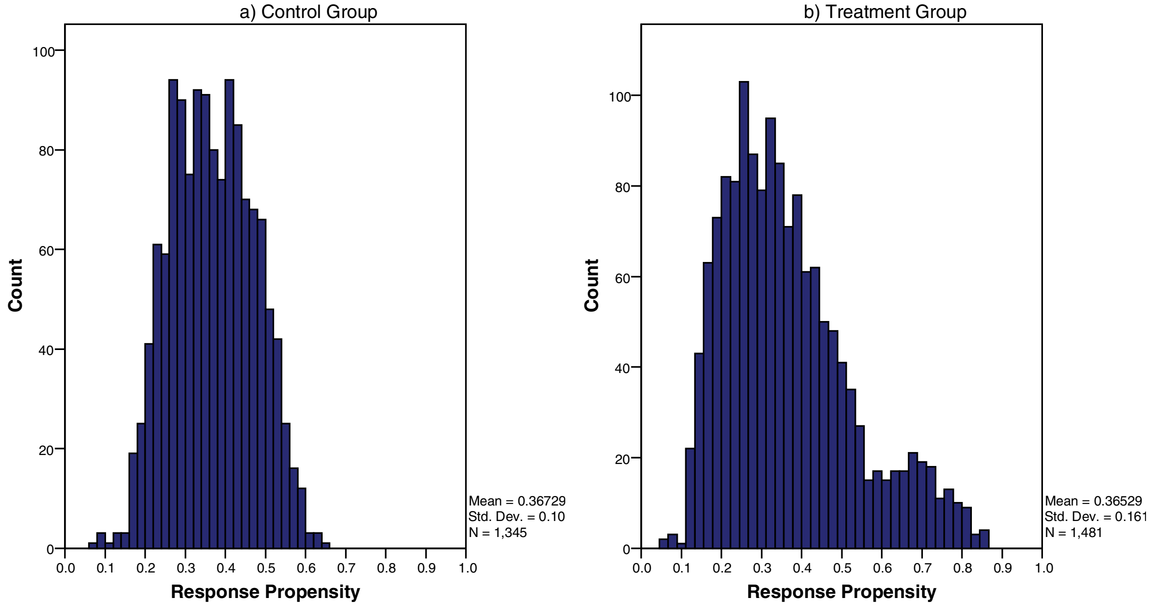


Figure 2: Distribution of Response Propensities by Experimental Condition

The other component of this study is to evaluate differences between the two methods of data collection used in the nonresponse follow-up, looking at both differences for not responding by mode and also at differences in the key variable of interest – weight loss, operationalized as reduction in BMI. This evaluation is especially important as telephone interviewing is typically more successful in gaining cooperation than mail surveys, but has also been shown to elicit more socially desirable responses – a concern here as weight is a potentially sensitive topic.

3. Methods

Nonresponse Adjustment models.

We examined several different adjustment models. The first model is the naïve complete case analysis under the MCAR assumption – only cases that had body weight reported in the 12-month web survey.

Under the model assuming no change, weight loss is imputed to be zero for those who did not respond to the 12-month survey. That is, the cause of nonresponse is absence of a treatment effect. This is another naïve model that at least yields conservative estimates of treatment effects, but can underestimate variances and hence lead to erroneous conclusions of significant differences between groups.

In the LOCF method if a subject did not complete the 12-month survey, the last weight they provided is used for the imputation even if it is the baseline measure.

The first weighting method is to weight the sample by the whole Nonresponse Follow-up (NRFU) respondents, ignoring differences between phone and mail. Weights are the inverse of the selection probabilities in the second phase.

To examine sensitivity of this nonresponse adjustment to the mode of data collection, another weighting scheme is to ignore the mail responses and weight by the phone NRFU respondents as if they were selected with higher probabilities (with no random assignment to mode in the second phase). A similar process is used to create weighted estimates using only the mail NRFU responses.

The last models used sequential regression multiple imputation in IVEware (Raghunathan, Lepkowski, VanHoewyk, and Solenberger 2001). These models were also repeated using only the phone NRFU responses (deleting the mail responses) and then again using the mail NRFU responses (deleting the phone responses). This method (and analytic tool) allows for elaborate multivariate modeling for nonresponse, including continuous and categorical predictors, interactions, and specification of the dependent variable as Gaussian, Poisson, dichotomous, multinomial, and ordinal. Unlike common practice in weighting from practical constraints, imputation models can be different for each variable with missing data, based on which covariates are informative.[3]

4. Results

The two modes used in the nonresponse follow-up (NRFU) achieved similar response rates – 59.3% (of 300) for telephone and 55.8% (of 398) for mail (χ^2 test, p>.05), although mail took longer to collect. Although monetary incentives were given only in mail, mail cases cost less than half of phone ($15 vs. $34), and similarly, per respondent ($28 vs. $57).

We first examined the reasons for which nonresponse occurred. Among the 400 completed surveys in the NRFU, 49% did not recall receiving the e-mail message, but only 6% of those did not have access to e-mail. There was no difference between the mail and the telephone NRFU respondents in recalling the message.

Among those who still had access to e-mail, 53% (208) did not remember reading the invitation, and among them 84% checked e-mail at least once a week, and 86% deleted e-mail messages without reading them.

Mail respondents were more likely to have remembered reading the invitation (50% vs. 39%, p<.05), providing some limited evidence for different reasons for not responding to the web survey invitation. However, this was not driven by attitude towards the center providing the services, as respondents in both modes did not differ in terms of rating of the treatment information and of satisfaction with the center. The proportion responding in the treatment group vs. the control group did not differ in the two modes.

However, we looked at questions that could exhibit social desirability in responses and some key differences arose. Respondents interviewed by phone rated themselves as more confident (t-test, p<.05) and more motivated (t-test, p<.05) in managing their own weight, and more confident in maintaining the recommended levels of activity (t-test, p<.05). These differences did not exist for the same respondents in the web survey at baseline. In terms of the focal goal of the study, the phone and mail groups had the same mean BMI in the web-administered baseline survey, BMI reduction was the same after 3 and 6 months, but respondents reported an average BMI reduction on the phone three times greater than the respondents by mail (1.2 vs. 0.4, p<.05).

Therefore, when computing treatment effects two sets of assumptions need to be tested for sensitivity: those that the models make about the properties of the nonrespondents, functional forms, multivariate associations, etc., and those that are concerned with the method of data collection in the NRFU.

There are two sources of nonresponse bias in estimating treatment effects. One is referred to as exogeneity bias – groups of interest are different prior to the intervention (i.e., respondents and nonrespondents are different at baseline and potentially in different proportions in the control and treatment groups). The other threat is endogeneity bias – the treatment effect is different for the groups (respondents and nonrespondents), and that can also be associated with being assigned to treatment.

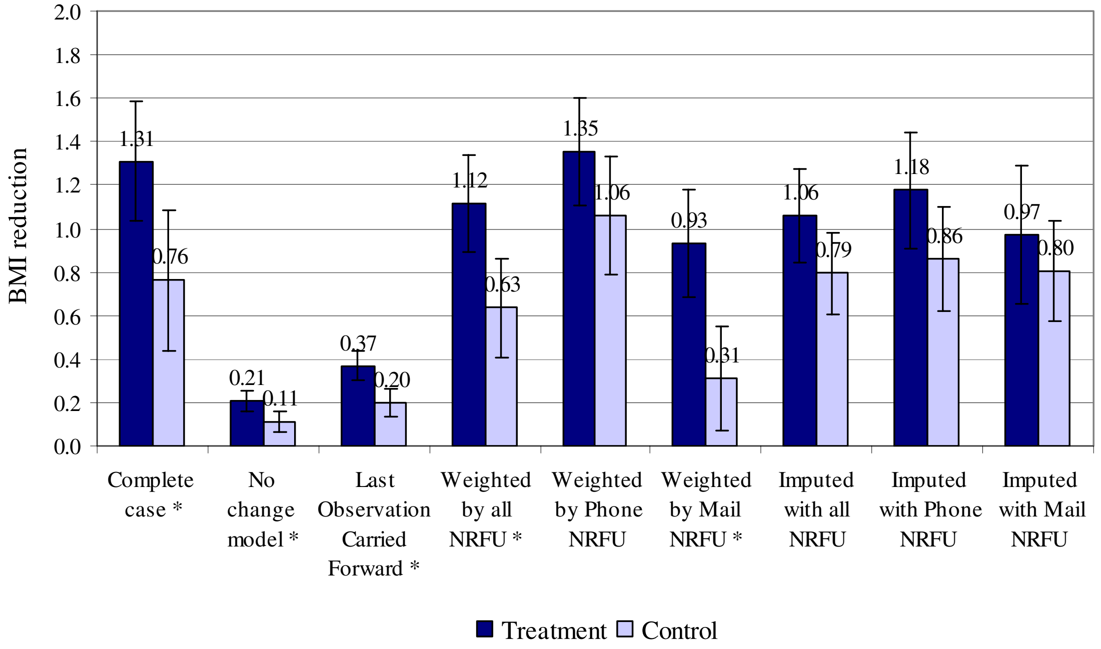
We found no evidence of exogeniety biashoerl@crd.ge.com. The nonrespondents to the NRFU did not differ from the respondents in terms of baseline BMI, and that was true for those assigned to mail and for those assigned to phone.

We also looked at endogeneity bias by looking at reduction in BMI for the NRFU respondents and nonrespondents for those who responded to the 3-month or 6-month follow-up. We found no such bias, as nonresponse in the NRFU was not associated with BMI reduction, and that was consistent for both the mail and phone samples.

The complete case estimates, presented in Figure 3, show that if we completely ignore nonresponse by neither launching a second phase nor adjusting for nonresponse covariates, BMI reduction is 1.3, or 72% larger in the treatment than the 0.76 in the control group (t-test, p<.05). This method assumes that the nonrespondents

[3] We used a criterion of minimum R^2=.02 for a variable to be included in the imputation model of another. For details on this procedure, visit: http://www.isr.umich.edu/src/smp/ive/

have the same weight loss pattern as the respondents.



* p<.05

Figure 3: BMI Reduction in Treatment and Control Groups under Different Treatment of Nonresponse.

As expected for the no change model, the reduction in BMI for the treatment group was much smaller, only 0.2, but still almost twice as large as the control group (t-test, p<.01). While not using any additional observed data than the complete case estimates, the sample size for these estimates was 3,260 rather than the observed 497, due to the imputed zeros.

The LOCF method utilizes the monotonically increasing nonresponse during this study but resembles the no change model in that it imputes zero change for those who did not complete any of the follow-up surveys, hence yielding higher but proportionately similar results.

Incorporating the data from the subsample of nonrespondents (NRFU), with 834 total respondents. and weighting them by the inverse of their selection probabilities, estimates were somewhat similar to the complete case analysis and still significant despite the loss due to weighting. However, initial analysis of the NRFU responses revealed large differences in weight reporting between mail and phone, while they were not different in their baseline measures, nor in follow-up measures for those who completed any of the online surveys. The larger weight loss reports on the phone were found for both the control and treatment groups. This difference was also evident in other questions eliciting self-presentation, which is in line with findings in the literature that respondents exhibit higher social desirability when interviewed by a person on the phone than when completing a self-administered mail questionnaire. Therefore, these adjustments that use the second phase sample were also done separately using either phone or mail.

When weighting by mail NRFU, the difference in BMI reduction is comparable to the complete case analysis, proportionately even twice as large (203%) for the treatment than control group, with an attenuation of the main effect that is unlikely due to time.[4]

When weighting by the phone NRFU, the BMI reduction for both conditions increases in

[4] While the follow-up of interest was done 12 months after baseline, the NRFU was done some 6 months later. Based on results not shown, weight loss did not decrease between the 3-, 6-, and 12-month follow-up surveys for the 364 respondents who completed all three.

magnitude, but the difference diminishes to .3 (or 34%) and is no longer significant.

The first model using multiple imputation was estimated using both mail and phone NRFU samples, adding a main effect of telephone and treating the self-administered web (12-month) and mail (18-month NRFU) modes together.[5] Rather than the commonly used rule-of-thumb of five, twenty full datasets (n=3260 each) were imputed to avoid overestimation of standard errors due to the large proportion of missing data. The difference in BMI between treatment and control was only .3, which was not significant. Repeating the imputations using only phone or mail from the NRFU, yielded differences of .3 and .2, respectively – neither significant at the .05 level.

5. Discussion and Conclusions

Based on just a single phase and assumption of MCAR using methods such as complete case, no change, and last observation carried forward, the conclusion would have been that the treatment was effective a year after the intervention, with a similar proportionate difference between treatment and control outcomes.

A second phase with a change in the data collection procedures, such as a different mode and incentives can be effective in learning about the nonrespondents. Making the less-stringent assumption that those who did not respond are like those who responded among the sample of nonrespondents led to the same conclusions. However, while the two modes used in the NRFU were both effective, they seem to produce different measurement errors, with phone being the most different from web and mail. When using only the phone sample to represent the nonrespondents, estimates of weight loss were higher, but not significantly different for the treatment and control groups.

When using a multiple imputation method that utilizes the data available on all nonrespondents, the difference between the treatment and control groups is not significant. Focusing on the model that uses only the responses from the self-administered modes, the estimated difference was only a third of that under the single-phase complete case analysis, and despite the much smaller standard errors was not significant.

In this case we would have concluded with untested certainty that there is a significantly larger effect of the experimental treatment, both under the single phase design, and under the two-phase design but discarding data on the nonrespondents. Using all the information with the more complex design and analysis did not confirm these results. However, given the small size of our NRFU and the two modes used, our results must be viewed as suggestive.

Our suggestion is that such checks of robustness of results to the assumptions about nonresponse be made, and when possible incorporated both in the study design and in analysis.

Design decisions have to be made not just on the basis of nonresponse error, but on multiple sources of error. Mode-specific measurement properties can alter not just point estimates, but also the variance-covariance structure of the data, and this confound between nonresponse and measurement error can hinder adjustments and analysis.

References

Baumgartner, R., and Rathbun, P. (1997) “Prepaid Monetary Incentives and Mail Survey Response Rates.” Unpublished paper presented at the Annual Conference of the American Association of Public Opinion Research, Norfolk, VA, May 15–18.

Deming, W.E. (1953) “On a probability mechanism to attain an economic balance between the resultant error of response and the bias of nonresponse.” Journal of the American Statistical Association 48(264):743-772.

Groves, R.M., Singer, E., and Corning, A. (2000) “Leverage-Salience Theory of Survey Participation: Description and an Illustration.” Public Opinion Quarterly 64(3):299-308.

Groves, R.M., Presser, S., Dipko, S. (2004) "The Role of Topic Interest in Survey Participation Decisions." Public Opinion Quarterly, 68 (1): 2-31.

Hansen, M. H., and Hurwitz, W. N. (1946). “The problem of nonresponse in sample

[5] Examining differences in reported weight loss over the first year, we have no reason to expect differences between 12 and 18 months due to time alone.

surveys." Journal of the American Statistical Association 41: 517-529.

Hansen, M., Hurwitz, W., and Bershad, M. (1961) "Measurement Errors in Censuses and Surveys." Bulletin of the International Statistical Institute, 38: 359-374.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. Survey Methodology, 12, 1-16.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Raghunathan, T.E., Lepkowski, J.M., J.VanHoewyk, and Solenberger, P. (2001) "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." Survey Methodology, 27:85-95.

Robert, K., et al. (2005), "Web-Based Weight Management Programs in an Integrated Health Care Setting: A Randomized, Controlled Trial." University of Michigan: unpublished paper.

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: Wiley.