Substitution of Nonresponding Units in Probability Sampling

by

Raphael Nishimura

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2015

Doctoral Committee:

Professor James M. Lepkowski, Chair
Professor Roderick J. Little
Professor Keith F. Rust, University of Maryland and Westat
Research Assistant Professor James R. Wagner

“*Samples are not given.*
They must be selected,
assigned or captured.”
- Leslie Kish

“*Our world, our life, our destiny*
are dominated by Uncertainty;
this is perhaps the only statement we
may assert without uncertainty.”
- Bruno de Finetti

To My Parents

ACKNOWLEDGMENTS

First, I would like to acknowledge the fellowship support from the Behner Family through the Harvey G. and Joyce H. Behner Graduate Fellowship, which funded me for the 2013 academic year. I would also like to thank the tremendous funding support from the Michigan Program Survey Methodology (MPSM) for the other four years in the program, going way beyond their initial commitment.

This would also not have been possible without the amazing support of the incredible MPSM staff. I would like to give my heartfelt thanks to the best and most enthusiastic group of ladies that a PhD student coud ask for: Jill Esau, Patsy Gregory, Jodi Holbrook, Elisabeth Schneider, Sumi Raj, Nancy Oeffner and Kellie Buss. Their assistance, support and enthusiasm were crucial from the beginning to the end of this journey.

I am very grateful for all the support I received from all the MPSM and Joint Program in Survey Methodology (JPSM) faculty for the past five years. In particular, Steve Heeringa, who chaired both of my qualifying and comprehensive exams, Trivellore “Raghu” Raghunathan and Sunghee Lee, who served on my comprehensive committee, and gave mevery helpful comments and ideas that I later used in my dissertation research. Mike Elliott, Frauke Kreuter, Rick Valliant, Fred Conrad and Partha Lahiri led the PhD seminars, where they created the perfect environment to develop some of the ideas in this dissertation, always providing great guidance and advice on how to design and conduct research in our field. I also have to thank Brady West for encouranging me to apply for the PhD program in the summer of 2008.

I am in great debt to all the members of my dissertation committee. Rod Little providedme with very important insights, which later become the basis of one of my studies, and his honest and straightforward opinions about my research certainly motivated me go further in improving it. Keith Rust gave me some very important help in understanding the methods I was

investigating and his suggestions certainly improved my research. James Wagner has always been very prompt and open to discussing research ideas with me and provided great feedback on this dissertation and on my research in general. Jim Lepkowski has been most definitely the best mentor I could have asked for. He has been the most patient, understanding and supportive advisor over the past five years. He made some very important and large contributions to all the studies in this dissertation. Moreover, he and his wife, Alicia, have always welcomed me as if I were a member of their own family. I will forever be grateful for everything they have done for me, going way beyond what I could have expected.

Throughout these five years in the doctorate program I made several friends in the MSPM and Ann Arbor. Also, I am fortunate to have some very good friends from high-school and college, who, despite the distance, were very supportive throught all these years away from home. I would love to list each and every one of them here, but I could never forgive myself if I forget to include anyone. I am sure they all know who they are.

I am lucky enough to have such a wonderful family that has always been very supportive in every decision I made. They have also provided me the best education they were able to, which allowed me to get where I am now. Although they were far away and I have not been doing a good job keeping in touch with them as frequent as they would like, they have always been very understanding and encouraged me in my studies and research.

Last, but not least, I am forever indebted to Joy Wilke. Her love and support kept me motivated through these past two and a half years. As if that was not just enough, she also reviewed and edited this entire dissertation more than once. Obviously, every mistake or typo is entirely my responsibility, but if this dissertation is readable, that is due to her. Not only that, she also helped me practice my dissertation defense. I hope one day I am able to return all her love, affection and suppoort.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CR	Complete Response
CSNI	Cluster-Specific Non-Ignorable Nonresponse
ISS	Inflated Sample Size
ISS.W	Inflated Sample Size adjusted by nonresponse propensity Weight
MCAR	Missing Completely at Random
MMM	Matching, Modeling and Multiple imputation
MMM.M	Modified Matching, Modeling and Multiple imputation
MNAR	Missing Not at Random
MSub	Matching Substitution
MSub.C	Calibrated Matching Substitution
MSub.W	Matching Substitution adjusted by nonresponse propensity Weight
NAEP	National Assessment of Educational Progress
PISA	Programme for International Student Assessment
PMM	Pattern-Mixture Model
PPM	Proxy Pattern-Mixture
PPS	Probability Proportional to Size
PSU	Primary Sampling Unit
RB	Relative change of the empirical Bias
RDD	Random Digit Dialing
RMSE	Root Mean Square Error
RSub	Random Substitution
RSub.W	Random Substitution adjusted by nonresponse propensity Weight
RV	Relative change in the empirical sampling Variance
SSU	Secondary Sampling Unit

ABSTRACT

The substitution of a nonresponding unit with one not originally selected in the sample is a commonly used method for dealing with unit nonresponse. Although frequently used in practice, substitution is largely neglected in the survey sampling literature. To date, few studies have attempted to develop a formal framework for describing and evaluating substitution methods, and little research has been done to improve estimates obtained through the use of substitution as a nonresponse adjustment procedure. This dissertation presents the results from three research studies conducted to enhance our understanding of substitution methods and develop new procedures to improve them.

The first study investigates substitution methods in stratified two-stage cluster sampling with nonresponse at the primary sampling unit (PSU) level. A simulation study is presented to evaluate the error properties of substitution procedures compared to other standard nonresponse adjustments. The results show that the use of a matching procedure in the selection of substitutes produces estimates with similar error properties to standard nonresponse-weighted estimates, but the substitution methods have the advantage of producing more accurate standard errors than strata collapsing strategies used in the presence of PSU nonresponse in stratified cluster sampling.

The second study extends an existing multiple imputation method proposed by Rubin and Zanutto (2002) that adjusts differences between nonrespondents and their substitutes on observable covariates to a more economically viable alternative. A new calibration approach is also proposed to perform such adjustments. Simulation results show that the multiple imputation extension performs as well as its predecessor, with the advantage of lower survey costs. Moreover, the proposed calibration procedure produces more precise estimates than the imputation methods with the same level of bias reduction, yielding estimates with smaller mean squared error.

The third study develops a novel procedure to accommodate nonignorable nonresponse in the substitution selection itself. The approach uses pattern-mixture models following Little and Andridge (2011) and Little (1994), and introduces a parameter that can be used in sensitivity analysis to assess assumptions about the nonresponse mechanism. Simulation studies show that the proposed approach can provide practitioners with useful information to evaluate the risk of nonresponse bias.

CHAPTER I

Introduction

Nonresponse occurs when a sampled unit fails to provide either part (item nonresponse) or all (unit nonresponse) of the information requested in a survey. This may be due to noncontact, refusal, inability to understand the request, or other reasons. This source of error has been increasingly studied in statistics and survey methodology, both theoretically and empirically, especially as response rates have fallen dramatically in recent decades (De Leeuw and De Heer, 2002; Rand, 2006, Bethlehem et al., 2011). On the other hand, the relationship between response rates and nonresponse error has been called into question by several studies (Keeter, et al., 2000; Merkle and Edelman, 2002; Curtin, Presser and Singer, 2005; Keeter, et al., 2006; Groves and Peytcheva, 2008), highlighting the importance of a careful exploration of all existing methods of handling nonresponse.

In the survey statistics literature, most of the methods for dealing with nonresponse have focused on post-data collection nonresponse analysis and adjustments, such as weighting, imputation and statistical modeling (Little and Rubin, 2002). Although post-survey adjustments are flexible and relatively inexpensive, methods for dealing with missing data, particularly unit nonresponse, in the survey design and field stages may present unique opportunities to minimize nonresponse error. As Benjamin King once said, “There is only one real cure for nonresponse and that is getting the response” (Frankel and King, 1996). In practice, however, with finite resources and time, nonresponse cannot be entirely eliminated. But some actions and interventions during the data collection stage could potentially mitigate the impact of nonresponse on final estimates.

To that end, a more formal and explicit framework to evaluate and minimize survey errors during the data collection stage of a survey has been proposed: responsive survey designs (Groves and Heeringa, 2006). In this approach, design feature indicators that influence both the survey costs and the errors of estimates are identified and monitored in an initial, pre-data collection, phase. In later phases of data collection, these design features may be modified based on the cost-error trade-offs. Finally, data from the different phases are combined to form a single survey estimate.

One of the most traditional examples of responsive design is the use of two-phase sampling for nonresponse (Hansen and Hurwitz, 1946). After an initial phase of data collection, in which all sampled cases are attempted to be contacted with the initial survey protocol, the second phase (usually called the nonresponse follow-up survey) involves contacting a probability-based subsample of nonrespondents, and subjecting this subsample to a more expensive and (theoretically) more effective data collection protocol. The final estimates are computed by weighting the subsampled cases by the product of the inverse of their second phase selection probability and their first-phase design-weight. If the second phase is completely successful, that is, if the full subsample of nonrespondents selected for the second phase is observed, then these final statistics are unbiased estimates for their population parameters. In practice, however, some level of nonresponse almost always remains. In such cases, there are some instances in which the inclusion of second-phase respondents may actually increase nonresponse bias.

Another approach to dealing with unit nonresponse at the fieldwork stage of a survey is substitution. This method consists of replacing nonresponding sampled units with new units which were not originally selected in the sample. Terms like “reserve” or “replacement” are also used to indicate substituted units. However, these terms are avoided here, especially because the latter, in particular, has another specific meaning in sampling (as in sampling with or without replacement). Most survey methodology and sampling textbooks either ignore (e.g., Cochran, 1977; Särndal et al., 1992; Groves et al., 2009) or present only a brief discussion of substitution (e.g., Kish, 1965; Lessler and Kalsbeek, 1992; Lohr, 1999; Little and Rubin, 2002). In general, the literature tends to criticize substitution and recommends avoiding its use, despite the lack of conclusive evidence suggesting it performs worse than competing alternatives, such as weighting or imputation. For example, Kish (1965, page 558) states:

> “Although substitution is often proposed naively as a solution, it generally is of little help and may actually make matters worse. (…) Entirely distinct from size control is the use of substitutes for reducing the bias of nonresponse. For this purpose substitutes are useless when they merely replace nonresponses with more elements that resemble the responses already in the sample.”

Although Cochran (1977) does not present any discussion of substitution, in its earlier edition (Cochran, 1953, page 302) he presents a similar point-of-view as Kish:

> “The ‘substitution’ method does positive harm if the samplers are deluded into thinking that the non-response problem has been adequately dealt with.”

Another example can be found on Deming (1953, page 744):

> “Substitution does not help: it is only equivalent to building up the size of the initial sample, leaving bias of nonresponse undiminished.”

Among other criticisms, there is an argument that substitution disrupts the selection probabilities of the sample design, making it no longer a probability sample. However, substitution can be seen as a form of imputation for unit nonresponse and, as Little and Rubin (2002, page 60) put:

> “The tendency to treat the resulting sample as complete should be resisted, since the substituted units are respondents and hence may differ systematically from nonrespondents. Hence at the analysis stage, substituted values should be regarded as imputed values of a particular type.”

Though the idea of treating substituted values as a type of imputed values is not further developed in the book, Rubin and Zanutto (2002) propose a method to do just that. It is true, however, that most applications of substitution in surveys do not treat substitutes’ data as imputed values. Related to this notion of substitution as an imputation method, substitution parallels hot-deck imputation (see Andridge and Little, 2010, for a recent review on the topic), with the difference that the latter draws the donors for the nonresponding cases from the respondent pool, while the former selects the substitutes from the unsampled units in the population.

Despite the criticism, substitution has been extensively used in many important probability sample surveys. This is true for survey in academic settings (Sirken, 1985; Vehovar, 1999; Bachman et al., 2011), surveys conducted by private companies, such as Westat (Walksberg, 1985), official statistics in some developing countries, and government surveys in Europe (Vehovar, 1999; Silva et al., 2000; Éltető, 2004)[1]. There are several reasons why substitution is used by many of these studies:

(1) Control of the sample size: When successfully implemented, that is, if most or every nonrespondent is replaced by a responding substitute, then the number of responding units will be the same or nearly the same as the target sample size. This could also be achieved by other means, such as inflating the sample size according to an expected response rate or the use of supplement sample (Kish, 1965). However, these methods in general will not produce an exact sample size for a particular realization of the survey. There is not a strong statistical reason for requiring an exact sample size, other than that estimates may be more precise if compared with an approach that does not take nonresponse into account. Nonetheless, many practitioners and survey clients demand a precise target sample size, sometimes even including this requirement in surveys contracts. Further, there is a certain aesthetic motivation behind this reason, in which laymen may view the observed sample size as an important measure of survey quality.

(2) Reduction of nonresponse bias: Although a main criticism of substitution is that it does not necessarily eliminate nonresponse bias, if compared to the naïve alternative of not using any nonresponse adjustments, substitution may provide some bias reductions under certain conditions. The first study of this dissertation seeks to investigate what such conditions are and the effectiveness of different methods of substitution. Obviously, such bias reduction could also be achieved with alternative nonresponse adjustment methods, such as weighting and imputation,

[1] Recently, however, some European governments have discontinued the use of substitution in their surveys (Vehovar, 1999; Pickery and Carton, 2008).

with less effort and cost. However, an important goal of this study is to assess differences in the effectiveness of a variety of nonresponse bias reduction techniques.

(3) Sample design structure: Related to sample size control, the main idea here is that nonresponse disrupts the design structure of complex samples, such as stratification and clustering, which might cause problems in the analysis, especially for the estimation of sampling variance. This becomes an important problem in designs that select few units per stratum or cluster. For example, deep stratification is a very common design, where two clusters per stratum are selected, maximizing the potential gains of stratification while still enabling sampling variance estimation. If some strata end up with one or no responding clusters, one would have to rely on strata collapsing procedures or other modeling approaches to estimate sampling variance, potentially biasing these estimates. If substitution is successfully implemented, the sample design structure is maintained and standard variance estimation procedure could be employed. However, as Vehovar (1999) pointed out, these two approaches would need to be compared in terms of mean square error of the sampling variance estimate. Thus, this comparison is also one of the objectives of the first study of this dissertation.

(4) Cluster nonresponse: In many applications of substitution, the nonresponding units are clusters, such as schools in a two-stage cluster sample of students, in which schools are selected in the first stage, and students are sampled within selected schools in the second stage. A nonresponding cluster automatically excludes multiple elements of the sample that belong to that cluster (elements that would potentially participate in the survey if requested). Smith (2007) states that many surveys rely on substitution in this case because the clusters are not the units of substantive analysis of those studies, but function only as a technical element of the sampling process, and, therefore, should not be a reason to eliminate the target elements of interest.

(5) Final refusal: Although nonresponse follow-up is considered one of the gold standard approaches to investigate and minimize nonresponse bias, it is often not completely successful; that is, it is not possible to get the full cooperation of all nonrespondents selected for a second phase. There are many reasons for that, but one of the most common is that once a unit selected in the sample (whether it is a person or institution) gives a definitive, final refusal, many survey organ-

izations would not continue attempting to get the cooperation of such cases. This is particularly true for institutions where strategies such as the use of incentives is either not used or not allowed. In those situations, substitution might be seem as an alternative to get responses from cases similar to these nonrespondents that are not in the sample already, although a sub-sample of nonrespondents would be still preferable.

Although extensively used, there are only a handful of studies that have examined substitution from a theoretical perspective (Nathan, 1980; Zanutto, 1998; Vehovar, 1999; Rubin and Zanutto, 2002; Thompson and Wu, 2008) and a few empirical studies, many of which were conducted before the 1990s (Durbin and Stuart, 1954; Cohen, 1955; Sirken, 1975; William and Folsom, 1977; Biemer, Chapman and Roman, 1985; Vives et al., 2009, David et al., 2012; David et al., 2014, Baldissera et al., 2014).

Because of the prevalent use of substitution for handling unit nonresponse in probability samples, but ambiguous evidence of its efficacy and skepticism from researchers' perspective, the primary objectives of the studies in this dissertation are to increase our understanding of this method, and to improve it by relaxing some of its assumptions and extending it to more general cases.

This dissertation continues by reviewing the limited existing literature on substitution in Chapter II. Then, in Chapter III, an investigation of the impact of primary sampling unit nonresponse on estimates of a finite population mean is conducted, followed by a comparison between different substitution methods and nonresponse weighting adjustments in terms of the performance of their point and sampling variance estimates through a large scale simulation study. In some instances, nonrespondents and their corresponding substitutes may differ on some observed auxiliary variables. If these covariates are related to the survey variables, such differences might cause bias in the survey estimates. In Chapter IV, a calibration approach to adjust for these differences is proposed, evaluated and compared to other methods previously developed in the literature through a simulation study. Another important understudied topic in the nonresponse literature, particularly in terms of substitution, is related to methods for dealing with missing not at random (MNAR) mechanisms. In Chapter V, a substitution selection method using pattern-mixture model is proposed to accommodate this missing mechanism, also allowing sensitivity analysis through the use of multiple substitutes. The performance of this method is evaluated and compared to other standard alternatives through a simulation study. Finally, Chapter VI presents a general discussion of the results of these three studies.

References

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40-64.

Bachman, J. G., Johnston, L. D., O'Malley P. M. and Schulenberg, J. E. (2011). *Monitoring the Future Project After Thirty-Seven Years: Design and Procedures.* Ann Arbor, MI. Institute for Social Research, University of Michigan.

Baldissera, S., Ferrante, G., Quarchioni, E., Minardi, V., Possenti, V., Carrozzi, G., Masocco, M., Salmaso, S. (2014). Field substitution of nonresponders can maintain sample size and structure without altering survey estimates - the experience of the Italian behavioral risk factors surveillance system (PASSI). *Annals of Epidemiology*, 24, pp. 241-245.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys.* John Wiley & Sons, Inc., Hoboken, New Jersey

Biemer, P., Chapman, D. W., and Alexander, C. (1985). Some Research Issues in Random-Digit Dialing Sampling and Estimation. *Proceedings First Annual Research Conference,* March 20-23, 1985.Washington DC: Bureau of the Census, 1985.

Cochran, W. G. (1953). *Sampling Techniques*, 1st edition. New York: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons.

Cohen, R. (1955). *An investigation of modified probability sampling procedures in interview surveys*. M.A. thesis submitted for the graduate faculty of The American University, May 26, 1955.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly*, 69, pp. 87-98.

De Leeuw, E. and De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D Dillman, J. Eltinge, and R. Little (eds.) *Survey Nonresponse,* pp. 41-54. New York: Wiley.

David, M. C., Bensink, M., Higashi, H., Donald, M., Alati, R., and Ware, R. S. (2012). Monte Carlo simulation of the cost-effectiveness of sample size maintenance programs revealed the need to consider substitution sampling. *Journal of Clinical Epidemiology*, Vol. 65, Issue 11, pp. 1200-1211.

David, M. C., Ware, R. S., Alati, R., Dower, J. and Donald, M. (2014). Assessing bias in a prospective study of diabetes that implemented substitution sampling as a recruitment strategy. *Journal of Clinical Epidemiology,* Vol 67, Issue 6, pp. 715-721.

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, pp. 743–772.

Durbin, J., and Stuart, A. (1954). Callbacks and clustering in sample surveys: An experimental study. *Journal of the Royal Statistical Society*. Series A, Part IV, pp. 387-428.

Éltető, O. (2004). Substitution in the Hungarian HSB. *The Survey Statistician*. No. 49, pp. 16.

Frankel, M. and King, B. (1996). A conversation with Leslie Kish. *Statistical Science,* Vol. 11, No. 1, pp. 65-87

Groves, R. M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Groves, R. M and Heeringa, S. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society,* 169 (Part 3), pp. 439-457.

Groves, R. M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly,* 72 (2), pp. 167-189.

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

Keeter, S., Miller, C., Kohut, A., Groves, R. M. and Presser, S. (2000). Consequences of Reducing Nonresponse in a Large National Telephone Survey. *Public Opinion Quarterly,* 64, pp. 125-48

Keeter, S., Kennedy, C., Dimock, M., Best, J. and Craighill, P. (2006). Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey. *Public Opinion Quarterly*, 70, pp. 759-779

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Merkle, D. M. and Edelman, M. (2002). Nonresponse in Exit Polls: A Comprehensive Analysis. In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, pp. 243-58. New York: Wiley.

Nathan, G. (1980). Substitution for Non-response as a Means to Control Sample Size. *Sankhyaa*, C42, 1-2, pp. 50-55.

Pickery, J., and Carton, A. (2008). Oversampling in Relation to Differential Regional Response Rates. *Survey Research Methods*, Vol. 2, No. 2, pp. 83-92.

Rand, M. (2006). Telescoping Effects and Survey Nonresponse in the National Crime Victimization Survey. Paper presented at the Joint UNECE-UNODC Meeting on Crime Statistics. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.14/2006/wp.4.e.pdf (accessed on March 21st 2014)

Rubin, D. B., and Zanutto, E. (2002). Using Matched Substitute to Adjust for Nonignorable Non response through Multiple Imputation. In *Survey Nonresponse*, edited by R. Groves, R. J. A. Little, and J. Eltinge. New York: John Wiley, pp. 389-402.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Silva, P. L. N., Bussab, W. O., Andrade, D. F., Freitas, M. P. S. (2000) *Plano Amostral SAEB 99: Avaliação e Substituição de Escolas Perdidas* (nº 3/99). Brasília: INEP 1999. (In Portuguese language)

Sirken, M. (1975). Evaluation and critique of household sample surveys of substance use. In *Alcohol and other drug use in the State of Michigan*. Final report, prepared by the Office of Substance Abuse Service, Michigan Department of Public Health.

Smith, T. W. (2007). *Notes on the Use of Substitution in Surveys*. www.issp.org/member/documents/Substitution_MC_Review.doc. Accessed on September 10th, 2012.

Thompson, M. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34, pp. 3-11.

Vehovar, V. (1999). Field Substitution and Unit Nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, pp. 335-350

Vives, A., Ferreccio, C. and Marshall, G. (2009). A comparison of two methods to adjust for nonresponse bias: field substitution and weighting non-response adjustments based on response propensity (In Spanish with a summary in English). *Gaceta Sanitaria*, 23 (4), pp. 266-271.

Waksberg, J. (1985). Comments on some research issues in random-digit-dialing sampling and estimation. *Proceedings of the Bureau of the Census Annual Research Conference*, vol. 1, 87-92.

Williams, S. R., and Folsom, R. E. Jr. (1977). *Bias resulting from school nonresponse:*

Metodology and findings. Prepared by the Research Triangle Institute for the National Center of Educational Statistics.

Zanutto, E. (1998). Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes. Doctorate thesis submitted for the graduate faculty of Harvard University, May, 1998.

CHAPTER II

Literature Review

Substitution is commonly known as the replacement of a nonresponding sampled unit by a new unit not originally included in the sample during the data collection stage. This strategy for dealing with unit nonresponse has been widely criticized over the years (Deming, 1953; Kish, 1965). However, as stated by Chapman (1983) and Vehovar (1994), no conclusive theoretical or empirical evidence had either rejected or justified this practice. A recent search in the statistics and survey research literature showed that this remains true, with few additional contributions since these earlier accounts. Still, the majority of survey methodology and sampling textbooks either do not recommend this option, especially in probability samples, or fail to mention it as a possible method of dealing with nonresponse. Nonetheless, substitution continues to be used in many studies (Stebe, 1995; Verma, 1992; Vehovar, 1995; Mazzeo et al., 1995; Silva et al., 2000; Éltető, 2004, Demarest et al., 2007; Bachman et al., 2011; Van der Hayden et al., 2013; Baldissera et al., 2014; David et al., 2014).

Like weighting for nonresponse – one the most common methods in surveys for dealing with unit nonresponse – substitution can be seen as a form of imputation. For example, in weighting-class nonresponse-adjustments (Oh and Scheuren, 1983) the nonresponse weights effectively take the average values of the survey variables of respondents in weighting classes to impute those variables of the nonrespondents in the same classes. With unit substitution, the nonrespondents are directly and fully imputed by their respective substitutes. Therefore, as pointed out by Chapman (1983), one of the major criticisms aimed at substitution – namely, that it only replaces nonrespondents by elements whose values resemble responses already in the sample – should also be applied to weighting and imputation-based nonresponse adjustment techniques. Hence, to determine whether substitution is an adequate method of dealing with nonresponse, these methods should be compared in terms of statistical properties (bias and precision) and operational efficiency under the same survey conditions and assumptions.

2.1 Types of Substitution

Chapman (2003) mentioned that substitution may be seen as an inferior method for dealing with unit nonresponse because it has been misused in the past. Much of this is likely due to differences in implementation. In fact, substitution is a vague term, as it may be implemented in various ways. To deal with this issue, Vehovar (2003) suggested a taxonomy for different kinds of substitutions, which was further refined by Lynn (2004).

Formerly, two basic types of substitution procedures had been widely recognized in the literature: the selection of a random substitute and the selection of a specially designated (and, thus, purposively selected) substitute, also denominated as a non-random or purposive substitute. In the former, as the term suggests, a substitute is selected on a probability selection basis to replace each nonrespondent. Although it is usually not explicitly stated, it seems reasonable to assume that the selection mechanism of the substitute follows that of the original sample selection. That is, if the originally sampled unit was selected with probability proportional to size, for example, so is its substitute. This typology on random substitution was further extended by Lynn (2004). More specifically, he distinguished two different methods of random substitution: simple random and stratified random. While in the former the substitute is randomly selected from the entire pool of nonsampled units remaining in the sampling frame, the latter forms or uses predefined strata and randomly replaces a nonrespondent with a unit in the same stratum.

Alternatively, in the non-random substitution method, the case that most closely matches the nonrespondent in terms of auxiliary variables available for all units in the population is selected. These auxiliary variables are usually from the sampling frame, or based on subjective information available from experts on the subject of the study. Selection of substitutes by interviewers in the field is another example of non-random substitution. However, such procedure is strongly discouraged even among proponents of substitution, especially if no constraints are put on when and how interviewers should substitute.

The method of selecting a substitute is, however, only one dimension of the types of substitution procedures considered by Vehovar (2003) and Lynn (2004). Another important aspect is whether the interviewer could or should influence the substitution procedure. Lynn (2004) breaks this down into two further dimensions: whether the *decision* to substitute and whether the actual *selection* is made by the interviewer. When the decision or the selection is not left to the interviewer, it is usually done by the survey management or administrators conducting the survey. In some surveys the substitutes are selected with the original sample, but left as "reserves" to be used if needed, as in Silva et al. (2000), for example.

As noted above, the influence of the interviewer, either on the decision or selection, has an important impact on survey quality. Furthermore, it is one of the main reasons for the criticisms aimed at the method, as pointed out by Vehovar (1994, 1999, 2003), Chapman (2003) and Lynn (2004). Briefly, when interviewers influence substitution, it may reduce the effort made to contact, get cooperation and responses from the original sampled cases. This could mean that, instead of actually solving the nonresponse problem, it can actually make it worse.

Therefore, it is generally recommended that interviewers should not influence the substitution procedure. It is also important to have a strict control over the field work to maintain a high level of effort to obtain the original sample, in order to avoid early respondent/cooperator effects (Vehovar, 1994). As observed by Chapman (1983) and Chapman and Roman (1985a,b), these recommendations are already embedded in RDD telephone surveys. There, the interviewer does not affect the substitution procedure, since there the interviewer typically does not know if the case being contacted is an original sampled unit or a substitute.

Another distinction made by Vehovar (2003) is the use of the substitution procedures in either probability or nonprobability sample surveys. The latter has not received much attention in the survey methodology literature, but it may be an important factor influencing the perception of the method over the years. This is especially true when substitution is confounded with quota sampling, a very popular nonprobability sampling method in market research. Although quota sampling uses substitutions, substitution is also used in non-quota samples.

Finally, as pointed out by Smith (2007), a substitution can be made at different levels of the sampling selection. That is, the substitute may be a primary or secondary sampling unit – such as a school, establishment or area segment – or the unit selected in the last stage of a multi-stage sample, which could be a student, employee, household or adult. Although not generally recommended, substitution at the first or second stage of sample selection is widely used in practice. Smith (2007) raises a point that might explain such discrepancy. Since these units are not generally target for substantive analysis of these studies, but rather serve only as a component for the sample selection, they should not be a reason to eliminate the final target study units.

2.2 Theoretical Research

As mentioned before, few theoretical studies have been conducted to investigate the properties of the substitution procedure. Chapman (1983) states that a possible reason for this is that, to conduct such a study, one would have to resort for a modeling approach and the models required for that would either be too complex to formulate or too simple to be of any usefulness.

This might have been true at that time, when the current missing data theory was still beginning to be developed. However, nowadays, many tools and powerful models are available in that literature. Nonetheless, theoretical work relating to substitution remains extremely scarce, limited and sparse. Actually, only four studies (Nathan, 1980; Vehovar, 1999; Rubin and Zanutto, 2002; Thompson and Wu, 2008) were found that tackle the problem in a more theoretical statistical perspective. Furthermore, two of them (Nathan, 1980; Rubin and Zanutto, 2002) deal with a somewhat different substitution approach that is not common in practice. Vehovar (1999) is a more comprehensive investigation of the substitution method, in which both bias and variance components of an estimator of the mean are studied under some specific settings.

As noted above, substitution, just like weighting for nonresponse, is a particular case of imputation. It shares, therefore, both its strengths and weaknesses. Hence, as pointed out by Chapman (1983), the investigation of the substitution properties should consist of comparing it with these competing alternatives. This has been limitedly done in the theoretical research presented in this section and in some of the empirical studies covered in the next section.

The substitution procedure reviewed by Nathan (1980) is somewhat different from what is usually used in practice, but quite similar to the strict random substitution: units are selected with equal probability until a fixed pre-defined number of respondents is achieved. Therefore, all nonrespondents would be sequentially substituted until the desired sample size is obtained. No further details are given, though, on how this would proceed in a real setting. It is worth noting that in this case the total number of contacts is a random variable. The alternative procedure that is used to make the comparison is more often done in practice: a fixed number of initial contacts is selected as a simple random sample, with no use of substitutions and, hence, the number of respondents is left as a random variable. Although further details are not given about this procedure in practice, it is expected that one would inflate the initial sample size based on the expected response rate in order to ultimately obtain a number of interviews closer to the desired target.

Nathan (1980) compares both proposed methods, conditional on a given level of desired accuracy, in terms of (1) the (expected) number of initial contacts and the (expected) number of respondents, (2) the expected costs, and (3) the variability of the costs. Interestingly, he concluded that both methods lead to approximately the same (expected) number of initial contacts, the same (expected) number of respondents and the same expected costs to attain a given level of accuracy. However, in terms of cost variability, when the ratio of the unit cost per initial contact and the additional cost per respondent is no larger than the response probability, the proposed substitution method is preferred. This is because it would give tighter budgetary control, that is, a smaller cost variation. Otherwise, the alternative method is recommended.

This study also makes very evident a property of using substitution in practice: it is almost guaranteed that the target sample size will be achieved, while only inflating the number of initial contacts will most likely always lead to an observed sample size that is different from the expected one. Although this might not be an important characteristic for a survey design from a statistical point of view, it seems to be a relevant one for practitioners and survey clients.

While most of the applications of substitution simply replace the nonresponding unit by its substitute, Rubin and Zanutto (2002) propose a different approach to the problem: multiply impute the nonresponding unit using its substitute data with an adjustment to take into account possible differences in their auxiliary variables. Assuming that both the nonrespondent's and the substitute's survey variable follows the same model, $Y = \beta_0 + \beta_1 Z + \beta_2 X + \varepsilon$, where X is a matching covariate used for substitution and Z is a modelling covariate used for statistical modelling, Rubin and Zanutto suggest substituting the nonrespondent with another unit with the same value on X (matching substitution) and multiply imputing the survey variable of the nonrespondent unit using the fact that $y_i = y_i^s + \alpha + \beta_1 \left(z_i - z_i^s \right) + \varepsilon_i^{'}$, where y_i and y_i^s are the nonrespondent and its substitute survey variable for the i^{th} case and z_i and z_i^s are the nonrespondent and its substitute auxiliary variable (not used for the matching substitution) for the i^{th} case. The idea is that the variable X , used for the matching substitution, would be difficult to use in a nonresponse adjustment, because it would involve too many parameters, such as address. On the other hand, the variable Z would not be available or used for matching, but it is observed for both substitutes and nonrespondents. This allows for systematic differences between them to be taken into account through modeling. To estimate the coefficients α and β_1 Rubin and Zanutto propose selecting substitutes for some of the respondents. After the survey variables for the nonrespondent are imputed, all substitutes are discarded.

Although Rubin and Zanutto show in a variety of simulation studies that this method, named as MMM (Matching, Modeling and Multiple imputation), performs better than all other competing alternatives (nonresponse weighting adjustments, multiple imputation using only respondent data and traditional substitution), to date, it is not known to have been used in survey practice (Chiu et al., 2005, applied the MMM method in a different context, in which “substitutes” were data aggregates from geographical census units, such as blocks or census tracts). This is probably because of the additional cost associated with collecting substitutes for respondents and not using their data in computing the final estimates.

Vehovar (1999) studies the sample mean under a random substitution procedure in two parts. First, the bias of the estimator is studied in a very general setting, in terms of both sample design and nonresponse mechanism (that is, no particular restriction is given for these two features). Then, the sampling variance property is investigated under a rather restricted scenario: in a two-stage cluster sample where it is assumed that the nonresponse mechanism is missing completely at random (MCAR) at the level where the substitution and the alternative adjustment procedures are performed. Furthermore, the study restricts itself to comparing the substitution procedure to a weighting alternative, but the author states that the extension to an imputation method would be straightforward and lead to very similar results.

Using a deterministic approach, Vehovar (1999) shows that the bias incurred to the sample mean using the substitution method, called *gross substitution bias*, is the sum of two components. First, the well known nonresponse bias expression of the unadjusted respondent mean $\bar{y}_r$: $B\left(\bar{y}_r\right)=\bar{M}\left(\bar{Y}_r-\bar{Y}_n\right)$, where $\bar{M}$ is the population nonresponse rate, $\bar{Y}_r$ and $\bar{Y}_n$ are the population means of respondents and nonrespondents, respectively. The second component, named *net substitution bias*, is given by $\bar{M}\bar{M}_{sn}\left(\bar{Y}_{sr}-\bar{Y}_{sn}\right)$, where $\bar{M}_{sn}$ is the proportion of elements among all initial respondents that would respond if included in the initial sample, but not if they were selected as substitutes (called *secondary nonrespondents*), while $\bar{Y}_{sn}$ and $\bar{Y}_{sr}$ denote the population means of secondary nonrespondents and *secondary respondents* (elements that would respond if selected either originally in the sample or as a substitute).

It is also argued, but without a formal proof, that both bias components generally have the same sign and, therefore, the net substitution bias typically enlarges the initial nonresponse bias. However, because of the product $\bar{M}\bar{M}_{sn}$ this additional bias is generally small. Nonetheless, without a strict control of the field work, not only will this component increase, but so will the initial nonresponse bias component, which can result in a dramatic increase of the gross substitution bias. This reinforces the argument made before on the importance of strict field work supervision when adopting such procedure.

Another very important result given in Vehovar (1999), without proof, is that $Var\left(\bar{y}_{SUB}\right)\approx Var\left(\bar{y}\right)$, when assuming that $E\left(\bar{y}_{SUB}\right)=\bar{Y}$. That is, the sampling variance of the sample mean under the substitution procedure is approximately the same as the variance of the sample mean when there is no nonresponse, as long as the sample mean is an unbiased estimator for both cases.

Under the restricted setting mentioned above, Vehovar (1999) shows that the increase in the variance of the sample mean due to using the alternative approach of attaching a nonresponse weighting adjustment, $w_i = b_i / b_{ri}^*$, over the substitution approach is directly related to $VIF = b_i E\left(1/b_{ri}^*\right)$. Here, b_i is the actual sample size in the i^{th} cluster – fixed and equal to the final size for the substitution method, and b_{ri}^* is the actual number of respondents in the i^{th} cluster, which is a random variable for the weighting approach. The overall increase in the sampling variance is also a function of the proportion of the within variance component.

Therefore, the gains in precision associated with using the substitution method depend on the number of units taken by cluster, the response rate, and the intracluster correlation. Through the investigation of several possible values for each of these factors, Vehovar concludes that the gains in precision by using substitution instead of weighting adjustments are usually small, but there are rare scenarios in which they can be relatively large, reaching as high as a 20% increase in precision. This would happen when all the aforementioned factors are small. Hence, his conclusion is that, other than those extreme cases, the gains in using the substitution method are too small to justify the possible extension of the field period – and the cost increase associated with it – that would be necessary for implementation.

Despite the distinction between random and non-random substitution, it is not clear how substitutes would be selected in the former case. It is reasonable to assume that they would be selected with the same methods that the original cases were. However, this can still create problems in the selection, especially when the substitution is used at the primary sampling unit level, which is usually selected with probability proportional to size. This problem has been studied only by Thompson and Wu (2008), using a Monte-Carlo simulation approach. Hence, the method of substitute selection also deserves further investigation to guide practitioners that rely on this alternative in their studies.

2.3 Empirical Studies

Most of the research of substitution’s impact on survey estimates comes from empirical studies. Chapman (1983) gave a comprehensive review of four important investigations: Durbin and Stuart (1954), Cohen (1955), Sirken (1975) and William and Folsom (1977). Each study conducted research using a different type of substitution and only one (Durbin and Stuart, 1954) compared the procedure with a competing alternative. Furthermore, as Chapman (1983) pointed out, none of these studies were carried out under ideal conditions, where efforts to get cooperation of initial nonrespondents would continue even after substitutions have been completed and data collection would stop only when all or most originally sampled cases are persuaded to participate (or, alternatively, administrative records would be used).

Although they used different approaches and methodologies, all four studies indicated that nonresponse bias was not completely eliminated with the substitution procedure. But, as Chapman (1983) noted, this does not invalidate the use of the method, since none of the competing methods are able to completely remove nonresponse bias either. At most, one can say that substitution does not underperform alternative procedures dealing with nonresponse error.

Smith (2007) also reported the results of eleven empirical studies on substitution, but without giving any specific references. The author stated that their reports were either only short summaries or rather limited and that none of them used an ‘optimal substitution’, a term used with a similar meaning of what Chapman (1983) referred as ‘ideal conditions’, mentioned above. Smith (2007) also noted that the substitution literature is dated: all the eleven empirical studies were conducted before the 1990s, and half before the 1980s.

Both Chapman (1983) and Smith (2007) concluded that more empirical research is needed; particularly that which comparing the alternative competitors with substitution under optimal circumstances, that is, with a strict field supervision, full-efforts to obtain both original and substitute cases and using random – preferably stratified random – substitution. Also, both authors agree that an important question to be answered is which of the two approaches performs better in reducing nonresponse bias: stratified random substitution or imputing nonrespondents by a mix of respondents in the same weighting cell.

Another empirical study conducted by Chapman and Roman (1985), under more ideal conditions, in a RDD telephone survey, compared the performance of a stratified random substitution procedure to an equal-cost sample relying solely on nonresponse weighting adjustment. They concluded that both methods perform almost equally well in terms of bias and sampling variance, with the substitution approach proving slightly more advantageous in terms of the latter. However, they strongly recommended only using the substitution procedure for nonresponse in RDD surveys, where there is a very strict control over the field operations when dealing with the substitutes, to avoid potential early cooperator biases.

Three recent studies compared the performance of substitution to weighting adjustments methods in health surveys and found mixed results. Although observing a large difference for one of the study's variables (smoking), Vives et al. (2009) concluded that the substitution procedure produces comparable results to a weighting adjustment based on response propensity. Baldissera et al. (2014) reached the same conclusion on a surveillance system survey. On the other hand, Van der Hayden et al. (2013) found that using only post-stratification had a larger impact on most of the estimates than using only substitution, under the same set of covariates. However, they concluded that using both strategies together might be more efficient than using only one of them alone (although substitution did not seem to make much of a difference once stratification was used).

David et al. (2012) looked at the cost of using substitution compared to an "usual practice" alternative that used repeated contacts, reminder letters and financial incentives (five prizes of $1,000, in a raffle) to gain the cooperation of nonrespondents and to minimize attrition in a longitudinal survey. In their study, they used random substitution on the second wave of the survey with the goal of achieving a desired sample size. By running cost-effectiveness analysis using the costs (including postage, telephone, courier and interview costs) and response propensities as input, and Monte Carlo simulations to evaluate the sensitivity of their results, they concluded that substitution was a more cost-effective strategy, with a smaller average cost per completed interview. Later, David et al. (2014) analyzed if there were any differences in the outcomes of the same survey between the substitution and the "usual practice" procedures. Comparing early respondents of both procedures (as a proxy for the substitutes) to early and late respondents of the "usual practice" strategy, they did not find many differences between the two groups, leading to the conclusion that, for their study, substitution would produce comparable results to the usual practice, but with a larger sample size and at a lower relative cost.

Dorsett (2010) also attempted to use substitution in a longitudinal study. However, instead of using random substitution as in David et al. (2012), he tried using propensity score matching to find substitutes that match the scores of dropouts to adjust for nonignorable attrition. Despite trying several approaches, he was not completely successful in solving the nonignorability issue.

Although there is some limited research examining the issue of substitution of nonresponding units in probability samples, the empirical results are mostly inconclusive, providing no definite guidelines for survey practitioners on when and how to properly use substitution. Moreover, problems like the use of substitution for cluster nonresponse or how to adapt the procedure to a missing not at random mechanism have not been fully addressed in the literature. The work presented in this dissertation takes a step towards filling some of these gaps.

References

Bachman, J. G., Johnston, L. D., O'Malley P. M. and Schulenberg, J. E. (2011). *Monitoring the Future Project After Thirty-Seven Years: Design and Procedures.* Ann Arbor, MI. Institute for Social Research, University of Michigan.

Baldissera, S., Ferrante, G., Quarchioni, E., Minardi, V., Possenti, V., Carrozzi, G., Masocco, M., Salmaso, S. (2014). Field substitution of nonresponders can maintain sample size and structure without altering survey estimates - the experience of the Italian behavioral risk factors surveillance system (PASSI). *Annals of Epidemiology*, 24, pp. 241-245.

Chapman, D. W. (1983). The Impact of Substitutions on Survey Estimates. *Incomplete Data in Sample Surveys*, Vol. II, Theory and Bibliographies, eds. W. Madow, I. Olkin, and D. Rubin, New York: National Academy of Sciences, Academic Press, pp. 45-61.

Chapman, D. W. and Roman, A. M. (1985a). Appendix 6 (Substitution). In *Results of the 1984 NHIS/RDD Feasibility Study: Final Report*, internal U.S. Bureau of Census report, February.

Chapman, D. W. and Roman, A. M. (1985b). An investigation of substitution for an RDD survey. *Proceedings of the Survey Research Methodology Section*, ASA, pp. 269-274.

Chapman, D. W. (2003). To Substitute or Not to Substitute – That is the question. *The Survey Statistician*. No. 48, pp. 32-34.

Chiu, W. F., Yucel, R. M., Zanutto, E. and Zaslavsky, A. M. (2005). Using Matched Substitutes to Improve Geographically Linked Databases. *Survey Methodology*, Vol. 31, No. 1, pp. 65-72.

Cohen, R. (1955). *An investigation of modified probability sampling procedures in interview surveys*. M.A. thesis submitted for the graduate faculty of The American University, May 26, 1955.

David, M. C., Bensink, M., Higashi, H., Donald, M., Alati, R., and Ware, R. S. (2012). Monte Carlo simulation of the cost-effectiveness of sample size maintenance programs revealed the need to consider substitution sampling. *Journal of Clinical Epidemiology*, Vol. 65, Issue 11, pp. 1200-1211.

David, M. C., Ware, R. S., Alati, R., Dower, J. and Donald, M. (2014). Assessing bias in a prospective study of diabetes that implemented substitution sampling as a recruitment strategy. *Journal of Clinical Epidemiology*, Vol 67, Issue 6, pp. 715-721.

Demarest, S., Gisle, L. and Van der Heyden, J. (2007). Playing hard to get: field substitutions in health surveys. *Internation Journal of Public Health*, 52, pp. 188-189.

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, pp. 743–772.

Dorsett, R. (2010). Adjusting for Nonignorable Sample Attrition Using Survey Substitutes Identified by Propensity Score Matching: An Empirical Investigation Using Labour Market Data. *Journal of Official Statistics*. Vol. 26, No. 1, 2010, pp. 105-125.

Durbin, J., and Stuart, A. (1954). Callbacks and clustering in sample surveys: An experimental study. *Journal of the Royal Statistical Society*. Series A, Part IV, pp. 387-428.

Éltető, O. (2004). Substitution in the Hungarian HSB. *The Survey Statistician*. No. 49, pp. 16.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Lynn, P. (2004). The Use of Substitution in Surveys. *The Survey Statistician*. No. 49, pp. 14-16.

Mazzeo, J., Allen, N.L., and Kline, D.L. (1995). *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading*. Washington, DC: National Center for Education Statistics.

Nathan, G. (1980). Substitution for Non-response as a Means to Control Sample Size. *Sankhyaa*, C42, 1-2, pp. 50-55.

Oh, H. L., and Scheuren, F. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies*, edited by W. G. Madow, I. Okin, and D. Rubin), pp. 143-184. New York: Academic Press.

Rubin, D. B., and Zanutto, E. (2002). Using Matched Substitute to Adjust for Nonignorable Non response through Multiple Imputation. In *Survey Nonresponse*, edited by R. Groves, R. J. A. Little, and J. Eltinge. New York: John Wiley, pp. 389-402.

Silva, P. L. N., Bussab, W. O., Andrade, D. F., Freitas, M. P. S. (2000) *Plano Amostral SAEB 99: Avaliação e Substituição de Escolas Perdidas* (nº 3/99). Brasília: INEP 1999. (In Portuguese language)

Sirken, M. (1975). Evaluation and critique of household sample surveys of substance use. In *Al cohol and other drug use in the State of Michigan*. Final report, prepared by the Office of Substance Abuse Service, Michigan Department of Public Health.

Stebe, J. (1995). Non-response in the Slovene Public Opinion Survey. *Contributions to Methodology and Statistics*, eds. A. Ferligoj and A. Kramberger, Ljubljana: Faculty of Social Sciences, pp. 21-37.

Smith, T. W. (2007). *Notes on the Use of Substitution in Surveys*. www.issp.org/member/documents/Substitution_MC_Review.doc. Accessed on September 10th, 2012.

Thompson, M. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34, pp. 3-11.

Van der Hayden, J., Demarest, S., Van Herck, K., De Barcquer, D., Tafforeau, J., Van Oyen, H. (2014). Association between variables used in the field substitution and post-stratification adjustment in the Belgian health interview survey and non-response. *International Journal of Public Health*, Vol 59, Issue 1, pp. 197-206.

Vehovar, V. (1994). Field substitution – a neglected option? *Proceedings of the Survey Methods Section,* ASA, pp. 589–94.

Vehovar, V. (1995). The Field Substitution in the Slovene Public Opinion Survey. *Contributions to Methodology and Statistics,* eds. A. Ferligoj and A. Kramberger, Ljubljana:Faculty of Social Sciences, pp. 38-66.

Vehovar, V. (1999). Field Substitution and Unit Nonresponse, *Journal of Official Statistics,* Vol. 15, No. 2, pp. 335-350

Vehovar, V. (2003). Field Substitution redefined*. The Survey Statistician*. No. 48, pp. 35-37.

Verma, V. (1992). Household Surveys in Europe: Some Issues in Comparative Methodologies. Paper presented at the Seminar: International Comparisons of Survey Methodologies, Eurostat, Athens,April 1992.

Vives, A., Ferreccio, C. and Marshall, G. (2009). A comparison of two methods to adjust for nonresponse bias: field substitution and weighting non-response adjustments based on response propensity (In Spanish with a summary in English). *Gaceta Sanitaria*, 23 (4), pp. 266-271.

Williams, S. R., and Folsom, R. E. Jr. (1977). *Bias resulting from school nonresponse: Metodology and findings*. Prepared by the Research Triangle Institute for the National Center of Educational Statistics.

CHAPTER III

Substitution of Nonresponding Primary Sampling Units in Probability Samples

Summary

Nonresponse occurs when a sampled unit fails to provide either part (item nonresponse) or all (unit nonresponse) of the information requested in a survey. The nonresponse literature has emphasized study of nonresponse arising at an element level, that is, the nonrespondent is the ultimate unit in the sampling process. However, in some multi-stage samples nonresponse occurs at earlier stages of the sampling process, such as in surveys of institutions like schools or establishments. In stratified multi-stage samples with few primary sampling units (PSUs) per stratum the risk is increased that if PSUs do not respond some strata will have only one or no responding PSUs, a problem for sampling variance estimation. A common strategy is to form pseudo strata with at least two PSUs each by collapsing strata with one or no PSUs, but sampling variability may be over-estimated. An alternative approach is to select substitute PSUs from units not originally selected in the sample. Vehovar (1999) observes that substitution for PSU-level nonresponse maintains the sample design structure allowing sampling variance estimation using the original stratification and cluster sampling design.

There are many different ways PSU-level substitution for nonresponse can be implemented. This study evaluates the impact on the survey estimates when various forms of substitution are used to compensate for nonresponse at the PSU-level of a two-stage cluster sample. Twelve methods are examined and compared in a simulation study to evaluate under which scenarios these substitution procedures are justified and compares substitution to alternative strategies such as sample size inflation, weighting, and strata collapsing. The bias and sampling variances are compared across substitution and non-substitution methods for handling PSU-level nonresponse.

3.1 Introduction

Nonresponse occurs when a sampled unit fails to provide either part (item nonresponse) or all (unit nonresponse) of the information requested in a survey. Nonresponse may be due to noncontact, refusal, an inability to understand a survey request for information, or other reasons. This source of potential error in survey estimates has been increasingly studied in statistics and survey methodology, both theoretically and empirically, especially as response rates have fallen dramatically in recent decades (De Leeuw and De Heer, 2002; Rand, 2006, Bethlehem et al., 2011). On the other hand, the relationship between response rates and nonresponse error has been called into question by several studies (Keeter, et al., 2000; Merkle and Edelman, 2002; Curtin, Presser and Singer, 2005; Keeter, et al., 2006; Groves and Peytcheva, 2008), highlighting the importance of a careful exploration of all existing methods for dealing with nonresponse.

In the survey statistics literature, most of the methods for dealing with nonresponse have focused on post-data collection adjustments such as weighting, imputation, and statistical modeling (Little and Rubin, 2002). Although post-survey adjustments are flexible and relatively inexpensive methods for dealing with missing data, survey data collection presents unique opportunities to minimize nonresponse error. As Benjamin King once said, “There is only one real cure for nonresponse and that is getting the response” (Frankel and King, 1996). In practice, however, with finite resources and time, nonresponse cannot be eliminated entirely. But some actions and interventions during the data collection stage could potentially mitigate the impact of nonresponse on final estimates.

An approach to dealing with unit nonresponse during survey data collection is substitution. This method consists of replacing nonresponding sampled units with new units which were not originally selected in the sample. Terms like “reserve” or “replacement” are also used to indicate substituted units. As indicated by Vehovar (1999), most survey methodology and sampling textbooks either ignore (e.g., Cochran, 1977; Särndal et al., 1992; Groves et al., 2009) or present only a brief discussion of substitution (e.g., Kish, 1965; Lessler and Kalsbeek, 1992; Lohr, 1999; Little and Rubin, 2002).

As pointed out by Lynn (2004), the literature, in general, tends to criticize substitution on two grounds. First, some forms of substitution involve interviewer decision making about when a substitute is to be used. Interviewers are given the flexibility to decide that a substitute is needed for a nonresponding unit. Second, some forms of substitution also allow the interviewer to choose the substituting unit, or a convenient unit is chosen as a substitute. There is compelling evidence that interviewer decision making about substitution is faulty and can lead to substantial bias in survey estimates (Chapman, 1983; Chapman, 2003; Chapman and Roman, 1985; Lessler and Kalsbeek, 1992; Lohr, 1999; Moser and Kalton, 1972; Vehovar, 1993). Much of the critical literature recommends avoiding the use of interviewer controlled or implemented substitution. These interviewer choice methods are not considered in this study.

Instead, the focus here is on forms of substitution in which the determination of when to substitute and which units to use as substitutes is controlled by survey investigators. The survey investigators reserve the right to decide or specify when a substitute is needed, and they select substitutes carefully, and not conveniently, to have similar characteristics to the nonresponding units. The choice of substitute units often involves matching on observable characteristics or a stochastic selection.

It is in these latter forms of substitution that there is a lack of conclusive evidence suggesting it performs worse than competing alternatives such as weighting or imputation. There have been a handful of theoretical studies on substitution (Nathan, 1980; Zanutto, 1998; Vehovar, 1999; Rubin and Zanutto, 2002; Thompson and Wu, 2008) and some empirical investigation of actual implementation (Durbin and Stuart, 1954; Cohen, 1955; Sirken, 1975; William and Folsom, 1977; Biemer, Chapman and Roman, 1985; Vives et al., 2009, David et al., 2012; David et al., 2014, Baldissera et al., 2014). There is still concern about this kind of more deliberate and controlled substitution as a procedure for dealing with nonresponse, in part because these prior studies have not generated the kind of conclusive results that have been sought.

The limited existing research on substitution focuses mainly on its use at the element level. For example, Vehovar (1999) examines in a two-stage cluster sample nonresponse and substitution occurring only at the second stage of the sampling process.

But many surveys use substitution as a remedy for nonresponse of entire clusters, as primary sampling units (PSUs) in two-stage sampling. This is particularly true in school-based surveys that sample schools as first stage units and students in the second stage within selected responding schools. For instance, the sample design guideline of the Programme for International Student Assessment (PISA) suggests substituting non-cooperating schools if the initial school response rate falls between 65% to 85% (PISA, 2012). The National Assessment of Educational Progress (NAEP) resorts to substitution for nonresponding schools, particularly for private schools that are not obliged to comply with study requests for testing students. The University of Michigan's Monitoring the Future survey substitutes for non-cooperating schools (Bachman et al, 2011). This usage of substitutes at the PSU–level has not yet been examined in the literature.

Further, nonresponse at the cluster level is another area for which there is a paucity of research. Studies that look at nonresponse in cluster sampling usually assume that there is at least one respondent in every cluster in the sample (Vehovar, 1999; Yuan and Little, 2006; Skinner and D'Arrigo, 2011). Such an assumption might be reasonable in some household surveys, where the clusters are typically cities, counties, census tracts or city blocks. When it occurs, the rate of nonresponding clusters is typically not high. However, the fact that the nonresponding PSUs are often either in high-income neighborhoods, such as gated communities, or dangerous areas, such as in slums or drug trafficking zones, might raise concerns about nonresponse bias. Nonresponding PSUs are even more common in surveys that use institutions to get access to the target population, such as school-based surveys targeting students. The number of nonresponding schools can be moderate to high, compromising the participation of all students in those schools and, thus, resulting in many clusters with no respondents. The nonresponse in these cases is usually the result of a lack of cooperation by school authorities.

This study focus on two-stage cluster sampling. It is assumed that some of the PSUs are nonrespondents and none of the corresponding secondary units respond, but in responding PSUs all secondary units respond. Although this may seem to be a strong assumption, in school-based surveys, for example, student response rates tend to be very high, particularly compared to household or individual response rates in household surveys.

Two sets of findings are presented. First, to demonstrate the importance of PSU nonresponse and to evaluate which parameters of the population and sample design have an impact on the nonresponse bias, theoretical results for the unadjusted respondent mean are given. Then, the results of a simulation study are presented to assess the performance of different substitution procedures compared to alternative nonresponse weighting-adjustment methods.

3.2 Bias of Unadjusted Respondent Mean Under PSU Nonresponse

3.2.1 Equal-sized Clusters

For the sake of simplicity, it is first analyzed the case in which the population consists of A clusters of equal size, B, so that the overall population size is $N=\sum_{\alpha=1}^{A} B = AB$. Let $Y_{\alpha\beta}$ be the value of a survey variables Y for the β^{th} element in α^{th} the cluster, for $\alpha = 1,...,A; \beta = 1,...,B$. The objective is to estimate the finite population mean:

$$\overline{Y} = \frac{\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B} Y_{\alpha\beta}}{N}.$$

For that purpose, a two-stage cluster sample is selected. At the first stage, a sample of a PSUs of the A clusters is selected with equal probability, but in only a_r is it possible to obtain a subsample of elements, due to nonresponse. At the second stage, b secondary sampling units (SSUs) of the B elements are selected in the α^{th} responding cluster. It is assumed that all selected SSUs respond to the survey. Because this design is a fixed size equal probability sample, if there were no nonresponse, the usual estimator for the population mean would be the sample mean:

$$\overline{y} = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b} y_{\alpha\beta}}{ab}$$

With nonresponse, a naïve approach would discard the nonresponding PSUs and use this same estimator using only the respondent data, that is, an unadjusted respondent mean:

$$\overline{y}_r = \frac{\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} y_{\alpha\beta}}{a_r b}$$

Denoting r_α and $I_{\beta|\alpha}$ the PSU cluster response indicator and the SSU sample indicator for the β^{th} sampled element in the α^{th} selected cluster, respectively:

$$r_\alpha = \begin{cases} 1, \text{ if the } \alpha^{th} \text{cluster is included in the sample and responds} \\ 0, \text{ otherwise} \end{cases}$$

and

$$I_{\beta|\alpha} = \begin{cases} 1, \text{ if the } \beta^{th} \text{ element in the } \alpha^{th} \text{cluster is included in the sample} \\ 0, \text{ otherwise} \end{cases}$$

Then the estimator can be re-written as

$$\overline{y}_r = \frac{1}{a_r b}\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} y_{\alpha\beta} = \frac{\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B}\frac{r_\alpha I_{\beta|\alpha} y_{\alpha\beta}}{ab}}{\frac{\sum_{\alpha=1}^{A} r_\alpha}{a}}$$

Under the given sample design, assuming that the cluster selection and response mechanisms are independent, the expected values of r_α and $I_{\beta|\alpha}$ are, respectively, $E\left(r_\alpha\right) = \frac{a}{A} p_\alpha$ and $E\left(I_{\beta|\alpha}\right) = \frac{b}{B}$, where p_α is the response propensity for cluster α . In order to derive the bias of this respondent mean, first, notice that this is a ratio estimator and hence, the Taylor Series expansion can be used to find its approximated expected value (Wolter, 2007). Let

$$\overline{y}_r = g\left(\hat{Y}_1, \hat{Y}_2\right) = \frac{\hat{Y}_1}{\hat{Y}_2} = \sum_{\alpha=1}^{A}\sum_{\beta=1}^{B}\frac{r_\alpha I_{\beta|\alpha} y_{\alpha\beta}}{ab} \Bigg/ \frac{\sum_{\alpha=1}^{A} r_\alpha}{a}$$

to approximate

$$g\left(Y_1,Y_2\right)=\frac{Y_1}{Y_2}=\frac{\overline{Y}}{\overline{p}}=\frac{\sum_{\alpha=1}^{A}\overline{Y}_{\alpha}}{A}\Bigg/\frac{\sum_{\alpha=1}^{A}p_{\alpha}}{A} \quad .$$

Then, the respondent mean can be approximated by

$$\overline{y}_r = g\left(\hat{Y}_1,\hat{Y}_2\right) \doteq g\left(Y_1,Y_2\right)+\frac{1}{Y_2}\left(\hat{Y}_1-Y_1\right)-\frac{Y_1}{Y_2^2}\left(\hat{Y}_2-Y_2\right)=\frac{\overline{Y}}{\overline{p}}+\frac{1}{\overline{p}}\left(\hat{Y}_1-\overline{Y}\right)-\frac{\overline{Y}}{\overline{p}^2}\left(\hat{Y}_2-\overline{p}\right)$$

Now, assuming that the first and second stage sample selections and the cluster nonresponse are independent,

$$E\left(\hat{Y}_1\right)=\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B}\frac{E\left(r_{\alpha}\right)E\left(I_{\beta|\alpha}\right)y_{\alpha\beta}}{ab}=\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B}\frac{\frac{a}{A}p_{\alpha}\frac{b}{B}y_{\alpha\beta}}{ab}=\frac{\sum_{\alpha=1}^{A}p_{\alpha}\overline{Y}_{\alpha}}{A} \quad \text{and}$$

$$E\left(\hat{Y}_2\right)=\frac{\sum_{\alpha=1}^{A}E\left(r_{\alpha}\right)}{a}=\frac{\sum_{\alpha=1}^{A}\frac{a}{A}p_{\alpha}}{a}=\frac{\sum_{\alpha=1}^{A}p_{\alpha}}{A}=\overline{p}\,,$$

the expected value of the respondent mean is approximately

$$E\left(\overline{y}_r\right)\doteq\frac{\overline{Y}}{\overline{p}}+\frac{1}{\overline{p}}\left(\frac{\sum_{\alpha=1}^{A}p_{\alpha}\overline{Y}_{\alpha}}{A}-\overline{Y}\right)-\frac{\overline{Y}}{\overline{p}^2}\left(\overline{p}-\overline{p}\right)=\frac{\overline{Y}}{\overline{p}}+\frac{1}{A}\frac{\sum_{\alpha=1}^{A}p_{\alpha}\overline{Y}_{\alpha}}{\overline{p}}-\frac{\overline{Y}}{\overline{p}}=\frac{1}{A}\frac{\sum_{\alpha=1}^{A}p_{\alpha}\overline{Y}_{\alpha}}{\overline{p}}$$

Therefore, the bias of the respondent mean in this case is given by

$$
\begin{aligned}
&Bias\left(\overline{y}_r\right) = E\left(\overline{y}_r\right) - \overline{Y} \doteq \frac{1}{A}\frac{\sum_{\alpha=1}^{A} p_\alpha \overline{Y}_\alpha}{\overline{p}} - \frac{1}{A}\sum_{\alpha=1}^{A}\overline{Y}_\alpha = \\
&= \frac{1}{A}\frac{1}{\overline{p}}\sum_{\alpha=1}^{A}\overline{Y}_\alpha\left(p_\alpha - \overline{p}\right) = \frac{1}{A}\frac{1}{\overline{p}}\sum_{\alpha=1}^{A}\left(\overline{Y}_\alpha - \overline{Y}\right)\left(p_\alpha - \overline{p}\right) = \\
&= \frac{1}{\overline{p}} Cov_a\left(Y, p\right)
\end{aligned}
$$

where $Cov_a\left(Y,p\right)$ is the covariance of the survey variable, Y, and the response propensity, p, with the subscript a to denote that this covariance is being evaluated at the cluster level.

This is similar to the bias expression for nonresponding *elements* in Bethlehem (1988). The expression for the bias of $\overline{y}_r$ can be further expanded as

$$
\begin{aligned}
&Bias\left(\overline{y}_r\right) \doteq \frac{1}{\overline{p}} Cov_a\left(Y, p\right) = \\
&= \frac{1}{\overline{p}} Corr_a\left(Y, p\right)\sigma_a\left(Y\right)\sigma_a\left(p\right) = \\
&= \frac{1}{\overline{p}} Corr_a\left(Y, p\right)\sigma_a\left(p\right)\sqrt{\frac{\sigma^2\left(Y\right)}{B}\left[1 + \rho\left(B - 1\right)\right]}
\end{aligned}
$$

where $Corr_a\left(Y,p\right)$ is the correlation of the survey variable, Y, and the response propensity, p; $\sigma_a\left(Y\right)$ and $\sigma_a\left(p\right)$ are the standard deviations of the survey variable and response propensity, respectively; and ρ is the intra-cluster correlation of the survey variable. Again, the subscript a denotes that these statistics are being evaluated at the cluster level.

The nonresponse bias in this case also depends upon the degree of homogeneity due to clustering. This is an intuitive result, since here all elements in a nonresponding cluster are missing, even though some, if not all, of these elements would respond to the survey, if requested. Hence, survey outcomes with high intra-cluster correlation will tend to have a higher bias compared to outcomes with lower within-cluster homogeneity.

3.2.2 Unequal-sized Clusters

Consider the case in which the population consists of A unequal-sized clusters, with B_α elements in the α^{th} cluster, so that the population size is $N=\sum_{\alpha=1}^{A} B_\alpha$. In this case, the finite population mean is given by

$$\overline{Y}=\frac{\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B_\alpha} y_{\alpha\beta}}{\sum_{\alpha=1}^{A} B_\alpha}.$$

Once again it is assumed that a two-stage cluster sample is selected. At the first stage, a sample of a PSUs of the A clusters is selected with probability proportional to size (PPS), B_α , but only a_r of them comply. At the second stage, b SSUs of the B_α elements are selected in the α^{th} responding cluster. Just as the previous case, it is assumed that all selected SSUs respond to the survey. Because this particular design (PPS two-stage cluster sample) is a fixed size equal probability sample, if there were no nonresponse, the usual estimator for the population mean would also be the sample mean:

$$\overline{y}=\frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b} y_{\alpha\beta}}{ab}$$

As previously, under the presence of nonresponse, a naïve approach would be to discard the nonresponding PSUs and use an unadjusted respondent mean:

$$\overline{y}_r=\frac{\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} y_{\alpha\beta}}{a_r b}=\frac{\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B_\alpha}\frac{r_\alpha I_{\beta|\alpha} y_{\alpha\beta}}{ab}}{\frac{\sum_{\alpha=1}^{A} r_\alpha}{a}}$$

where r_α and $I_{\beta|\alpha}$ are the same as before.

Under a PPS selection, and assuming that the cluster selection and response mechanisms are independent and that the SSUs are selected with equal probability within the PSUs, the expected values of r_α and $I_{\beta|\alpha}$ are, respectively, $E\left(r_\alpha\right)=a\frac{B_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}p_\alpha$ and $E\left(I_{\beta|\alpha}\right)=\frac{b}{B_\alpha}$.

Let

$$\overline{y}_r = g\left(\hat{Y}_1,\hat{Y}_2\right)=\frac{\hat{Y}_1}{\hat{Y}_2}=\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B_\alpha}\frac{r_\alpha I_{\beta|\alpha}y_{\alpha\beta}}{ab}\Bigg/\frac{\sum_{\alpha=1}^{A}r_\alpha}{a}$$

to approximate

$$g\left(Y_1,Y_2\right)=\frac{Y_1}{Y_2}=\frac{\overline{Y}}{\overline{p}}=\frac{\sum_{\alpha=1}^{A}B_\alpha\overline{Y}_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}\Bigg/\frac{\sum_{\alpha=1}^{A}B_\alpha p_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}\ .$$

Assuming that the first and second stage sample selections and the cluster nonresponse are independent,

$$E\left(\hat{Y}_1\right)=\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B_\alpha}\frac{E\left(r_\alpha\right)E\left(I_{\beta|\alpha}\right)y_{\alpha\beta}}{ab}=\sum_{\alpha=1}^{A}\sum_{\beta=1}^{B_\alpha}\frac{a\frac{B_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}p_\alpha\frac{b}{B_\alpha}y_{\alpha\beta}}{ab}=\frac{\sum_{\alpha=1}^{A}B_\alpha p_\alpha\overline{Y}_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}=\frac{\sum_{\alpha=1}^{A}B_\alpha p_\alpha\overline{Y}_\alpha}{N}$$

and

$$E\left(\hat{Y}_2\right)=\frac{\sum_{\alpha=1}^{A}E\left(r_\alpha\right)}{a}=\frac{\sum_{\alpha=1}^{A}a\frac{B_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}p_\alpha}{a}=\frac{\sum_{\alpha=1}^{A}B_\alpha p_\alpha}{\sum_{\alpha=1}^{A}B_\alpha}=\overline{p}\ ,$$

using Taylor Series approximation, the expected value of the respondent mean in this case is approximately

$$E\left(\overline{y}_r\right) \doteq \frac{\overline{Y}}{\overline{p}} + \frac{1}{\overline{p}}\left(\frac{\sum_{\alpha=1}^{A} B_\alpha p_\alpha \overline{Y}_\alpha}{N} - \overline{Y}\right) - \frac{\overline{Y}}{\overline{p}^2}\left(\overline{p} - \overline{p}\right) = \frac{\overline{Y}}{\overline{p}} + \frac{1}{N}\frac{\sum_{\alpha=1}^{A} B_\alpha p_\alpha \overline{Y}_\alpha}{\overline{p}} - \frac{\overline{Y}}{\overline{p}} = \frac{1}{N}\frac{\sum_{\alpha=1}^{A} B_\alpha p_\alpha \overline{Y}_\alpha}{\overline{p}}$$

Therefore, the bias of the respondent mean in this case is given by

$$\begin{aligned}
&Bias\left(\overline{y}_r\right) = E\left(\overline{y}_r\right) - \overline{Y} \doteq \\
&\doteq \frac{1}{N}\sum_{\alpha=1}^{A}\frac{B_\alpha p_\alpha \overline{Y}_\alpha}{\overline{p}} - \frac{1}{N}\sum_{\alpha=1}^{A} B_\alpha \overline{Y}_\alpha = \frac{1}{N\overline{p}}\sum_{\alpha=1}^{A} B_\alpha \overline{Y}_\alpha\left(p_\alpha - \overline{p}\right) = \\
&= \frac{1}{N\overline{p}}\sum_{\alpha=1}^{A} B_\alpha\left(\overline{Y}_\alpha - \overline{Y}\right)\left(p_\alpha - \overline{p}\right)
\end{aligned}$$

This is of a similar form of the bias expression of the equal-sized clusters case, but with the covariance weighted by the cluster sizes, which implies that larger clusters might have a larger impact on the nonresponse bias.

3.3 Simulation Study

Despite being extensively used in practice, the statistical properties of the various types of substitution methods are still not well understood, particularly the substitution of clusters, such as PSUs, in complex survey design settings, involving stratification, clustering, multiple stages, and unequal selection probabilities. Furthermore, a comparison of the performance of these substitution procedures to other nonresponse adjustment methods is needed to guide practitioners about the implications of using each one of these methods in different populations and contexts.

More specifically, it would be helpful to know: (1) which methods lead to unbiased estimates; (2) what methods produce the most precise estimates; and (3) which methods lead to the smallest mean squared error for the estimates of the population parameters. It would be important to evaluate these properties of the different method for dealing with nonresponse under a range of values for important population and survey design features that can impact nonresponse. Moreover, the sampling variance estimates of these methods should also be evaluated for a more complete portrait of their statistical inference properties.

For these purposes, a series of simulations were carried out, each selecting 5,000 stratified two-stage cluster samples of size *n* = 1,500 (*a* = 100 clusters and *b* = 15 elements per cluster) from populations of approximately *N* = 400,000 elements composed of *A* = 2,000 clusters of unequal size.

The simulation process involved the generation of a population of clusters, the generation of a population of elements within each cluster, the selection of a sample of PSUs and elements within sample PSUs, the application of a missing data mechanism to the sample to obtain the responding unit sample, the selection of substitute PSUs for nonresponding PSUs, and the calculation of various estimates from each sample, including bias and variance.

In these simulations, the objective was to estimate the finite population mean of a survey variable *Y*. An auxiliary variable, *X*, at the cluster level was assumed to be observed for all clusters, respondents or nonrespondents. The simulations were conducted with:

- Three levels of correlation between *Y* and *X*: low ($Corr_a\left(Y, X\right) = 0.01$), medium ($Corr_a\left(Y, X\right) = 0.30$) and high ($Corr_a\left(Y, X\right) = 0.70$) (as before, the subscript *a* denotes that the correlations are at the cluster-level);
- Three levels of intra-cluster correlation for the *Y* survey variable: low ($\rho = 0.01$), medium ($\rho = 0.20$) and high ($\rho = 0.50$);
- Three cluster-level response propensity means (cluster response rate): low ($\overline{p} = 0.50$), medium ($\overline{p} = 0.75$) and high ($\overline{p} = 0.90$), and;
- Two missing mechanisms: missing at random conditional on the variable *X* (MAR) and missing not at random (MNAR).

Thus, there were 3 x 3 x 3 x 2 = 54 different simulation settings, derived from the combinations of correlation, intra-cluster correlation, response rate, and missing data mechanisms examined. First, nine populations were generated corresponding to the combinations of correlation and intraclass correlation given above. Then, for each of these nine populations, six nonresponse scenarios were considered, the combinations of the three response rate levels with the MAR and MNAR nonresponse mechanisms.

Finite Populations Generation

The parameters of the nine finite stratified and clustered populations, derived from the combination of the three (*X*, *Y*) correlation and the three intra-cluster correlations, are summarized in Table 3.1.

Because of the stratified, clustered nature of these populations, the values of the survey variable *Y* were hierarchically generated in two steps. First, the cluster means $\bar{Y}_\alpha$ were generated from a multivariate normal distribution together with other three cluster variables:

- X_α, which denotes a cluster variables to be used in matching substitutions and nonresponse adjustments;
- W_α, which was used to stratify the clusters, and;
- U_α, which assisted in the generation of the cluster sizes, B_α.

Once the clusters characteristics were generated, the survey outcome values of the B_α elements in each of the clusters were drawn from a normal distribution with mean $\bar{Y}_\alpha$. The two-step algorithm that implemented this population generation is given below with more details of this process:

1. At the cluster level, *A* = 2,000 vectors of cluster characteristics were generated independently under the following multivariate normal distribution:

$$\begin{pmatrix} \overline{Y}_\alpha \\ X_\alpha \\ W_\alpha \\ U_\alpha \end{pmatrix} \sim N_4\left(\begin{pmatrix} 100 \\ 100 \\ 100 \\ 5 \end{pmatrix}, \begin{pmatrix} \sigma^2_{Y_B} & \sigma_{YX} & \sigma_{YW} & \sigma_{YU} \\ \sigma_{XY} & 400 & 0 & 0 \\ \sigma_{WY} & 0 & 400 & 0 \\ \sigma_{UY} & 0 & 0 & 1 \end{pmatrix} \right), \quad \alpha = 1,\ldots,2000$$

To simulate cluster sizes similar to ones that might be found in school-based surveys, the size of each cluster was generated from the variable U as $B_\alpha = \lceil \exp(U_\alpha) \rceil + b,\ \alpha = 1,\ldots,2000$. To avoid undersized clusters that would complicate sample selection, an additional *b* units were added to the cluster sizes. Some cluster sizes were trimmed to prevent oversized units (Kish, 1965) with sizes so large they would be selected with certainty, or multiple times, in the subsequent probability proportionate to size selection of clusters.

Stratification of clusters was based on the variable W. Clusters were sorted by the value of W and divided into $H = 50$ strata of approximately equal size. The subscript h is added in the notation hereafter to denote cluster stratum.

The covariances σ_{YW} and σ_{YU} were set so that the correlations between Y and W and between Y and U were both 0.2 (the correlation between Y and the cluster sizes, B, was approximately 0.1) in all populations. The covariance σ_{YX} was set accordingly to the variance $\sigma^2_{Y_B}$ so that the correlation between Y and X at the cluster level assumes the three different levels mentioned before: low ($Corr_a(Y,X) = 0.01$), medium ($Corr_a(Y,X) = 0.30$) and high ($Corr_a(Y,X) = 0.70$). Below, the way in which the values of $\sigma^2_{Y_B}$ were set is discussed.

2. The survey variable for the $B_{h\alpha}$ elements within the α^{th} cluster in the h^{th} stratum was generated independently following

$$Y_{h\alpha\beta} \sim N\left(\overline{Y}_{h\alpha}, \sigma^2_{Y_W}\right), \quad h = 1,\ldots,50;\ \alpha = 1,\ldots,1000;\ \beta = 1,\ldots,B_{h\alpha}$$

The between- and within-cluster variability of the Y variable, $\left(\sigma^2_{Y_B};\sigma^2_{Y_W}\right)$, were set to $(4;396)$, $(80;320)$ and $(200;200)$ so that $\sigma^2_Y \approx 400$ and the intra-cluster correlation, computed as $\rho = \sigma^2_{Y_W} \Big/ \left(\sigma^2_{Y_W} + \sigma^2_{Y_B}\right)$, takes approximately the three different levels: low ($\rho = 0.01$), medium ($\rho = 0.20$) and high ($\rho = 0.50$), respectively.

Table 3.1: Population parameters used in simulations

Population	ρ	$Corr_a(Y,X)$	$\sigma^2_{Y_B}$	$\sigma^2_{Y_W}$	σ_{YX}
1	0.01	0.01	4	396	0.40
2	0.01	0.30	4	396	12.00
3	0.01	0.70	4	396	28.00
4	0.20	0.01	80	320	1.79
5	0.20	0.30	80	320	53.67
6	0.20	0.70	80	320	125.22
7	0.50	0.01	200	200	2.83
8	0.50	0.30	200	200	84.85
9	0.50	0.70	200	200	197.99

Sample Design

Each simulation selected a stratified two-stage cluster sample from each of the cluster populations. A first-stage sample of $a_h = 2$ PSUs were selected with a random systematic-PPS procedure from each stratum *h* = 1, …, 50, using the cluster sizes, B_α, as measures of size. A second-stage sample of $b = 15$ SSUs were selected with simple random sampling without replacement within each sampled PSU. This is a design that maximizes the benefits of stratification to the point where design-based sampling variance estimation is still possible. Because the strata were approximately the same size, design weights were used to adjust for some minor differences in the selection probabilities across elements in different strata.

Some of the methods evaluated in this study inflated the selected sample size to compensate for nonresponse to obtain the target sample size. Because the missing mechanism operated at the cluster level, only the first-stage sample size were inflated in each stratum according to the overall cluster response propensity mean. In particular, instead of selecting a_h clusters in the h^{th} stratum, $a_h^{'} = a_h / \bar{p}$, where $\bar{p} = \sum_{\alpha=1}^{A} p_\alpha / A$ were selected to compensate for PSU level nonresponse. For example, for an overall cluster response propensity $\overline{p} = 0.75$, $a_h^{'} = 2/0.75 = 2.67 \approx 3$ PSUs were selected in each stratum. This approach assumes a prior estimate of the response propensity. In practice, response rates of previous studies are used as estimates for this quantity.

Nonresponse mechanism

The nonresponse mechanism was then applied to the selected clusters. Because the interest here is in nonresponse at the cluster-level, all the elements within respondent clusters were considered to be respondents. Two nonresponse mechanisms were used in this study: missing at random (MAR), in which the nonresponse depends only on the variable X, and missing not at random (MNAR), in which the nonresponse also depends on the survey variable Y. For each mechanism, three cluster response propensity levels were used: low ($\overline{p} = 0.50$), medium ($\overline{p} = 0.75$) and high ($\overline{p} = 0.90$).

The response propensities, $p_{h\alpha}$, were computed for each cluster in the population according to the following model:

$$\text{logit}\left(p_{h\alpha}\right) = \beta_0 + \beta_1 X_{h\alpha} + \beta_2 \overline{Y}_{h\alpha}, h = 1,...,50; \alpha = 1,...,2000$$

The model's coefficients were set according to the missing mechanism and cluster response propensity means as summarized in Table 3.2.

Table 3.2: Nonresponse mechanism model coefficients

	$\overline{p}$	β0	β1	β2
MAR	0.50	-1.25	0.0125	0
	0.75	-0.15	0.0125	0
	0.90	0.95	0.0125	0
MNAR	0.50	-2.5	0.0125	0.0125
	0.75	-1.4	0.0125	0.0125
	0.90	-0.3	0.0125	0.0125

For each sampled cluster a random variable $u_{h\alpha} \sim \text{Uniform}(0,1)$ was independently generated. The cluster was treated as a nonrespondent, that is, none of its elements were observed, if $u_{h\alpha} < p_{h\alpha}$; otherwise the cluster was considered a responding PSU and a sub-sample of b = 15 elements was selected. This same missing mechanism was applied to units selected to substitute nonresponding PSUs. That is, there was also nonresponse among substitutes and its mechanism was assumed to be the same that operates on the original sample. In practice, this latter assumption is only true if the substitutes receive the same survey protocol as the originally selected units.

Nonresponse adjustment methods evaluated in this simulation study

The methods used to compensate for nonresponding PSUs compared in this simulation study, are summarized in Table 3.3. The objective is to estimate the finite population mean $\overline{Y} = \sum_{i=1}^{N} Y_i / N$. Design-weights were used in all estimates. With the sample design used in the simulations, the design-weights are given by[2]

$$d_{h\alpha\beta} = \pi_{h\alpha\beta}^{-1} = \left(\pi_{h\alpha} \times \pi_{\beta|h\alpha}\right)^{-1} =$$

$$= \left(a_h \frac{B_{h\alpha}}{\sum_{\alpha=1}^{A_h} B_{h\alpha}} \times \frac{b}{B_{h\alpha}} \right)^{-1} = \left(\frac{a_h b}{\sum_{\alpha=1}^{A_h} B_{h\alpha}} \right)^{-1}, h = 1,...,50; \alpha = 1,...,a_h; \beta = 1,...,15$$

As previously mentioned, most of the proposed methods in the literature to adjust and compensate for nonresponse assume that the missingness occurs at the element level. However, many of the element level compensation methods can also be applied to clusters. For the purposes of this study, the substitution methods are compared to two very common alternative solutions: the unadjusted respondent mean and the nonresponse-adjusted weighted mean. In order to achieve, on average, the target sample size and have comparable sample sizes for methods relying on substitution, both of these approaches used a number of PSUs inflated by the average cluster response propensity.

[2] For the methods that used an inflated sample size, a_h is substituted by $a_h^{'} = a_h / \overline{p}$

Table 3.3: Methods to compensate for missing PSUs investigated in the simulation study

Abbreviation	**Method**	**Substitution iterations**	**Nonresponse weighting?**
ISS	Inflated Sample Size	-	No
ISS.W	Inflated Sample Size	-	Yes
MSub1	Matching Substitution	1	No
MSub3	Matching Substitution	3	No
MSub7	Matching Substitution	7	No
RSub1	Random Substitution	1	No
RSub3	Random Substitution	3	No
RSub7	Random Substitution	7	No
MSub1.W	Matching Substitution	1	Yes
MSub3.W	Matching Substitution	3	Yes
MSub7.W	Matching Substitution	7	Yes
RSub1.W	Random Substitution	1	Yes
RSub3.W	Random Substitution	3	Yes
RSub7.W	Random Substitution	7	Yes

Unadjusted respondent mean with inflated sample size (ISS)

The unadjusted respondent mean, or inflated sample size estimate (ISS), is the naïve complete-case analysis estimator without any adjustment for nonresponse. Although this method is usually not recommended in practice, it is used here as a baseline point to evaluate the magnitude of nonresponse bias reduction when the other methods are used. In this case, the estimate for the population mean of the survey variable Y uses only the data of the elements selected within the responding PSUs. Weights in the ISS method include only the design weights, with no further weighting or imputation-adjustments:

$$\overline{y}_{ISS} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} d_{h\alpha\beta}\, y_{h\alpha\beta}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} d_{h\alpha\beta}}$$

Nonresponse adjusted weighted mean with inflated sample size (ISS.W)

The nonresponse adjusted weighted mean, or the inflated sample size weighted mean (ISS.W) is used as a benchmark for the substitution methods, as it is the most commonly used alternative, particularly with regard to element-level nonresponse weighted approach (here adapted for cluster-level nonresponse). In this approach, the responding PSUs are weighted by the inverse of their predicted response propensity, $\hat{p}_{h\alpha}$, which is estimated using the selected sample with the following logistic regression model:

$$\text{logit}\left(\hat{p}_{h\alpha}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{h\alpha} + \hat{\beta}_2 B_{h\alpha}$$

This is the correct response model under MAR, but it is misspecified under MNAR. Both the cluster sizes and stratification indicator were initially included in the model to take into account all the design variables, as suggested by Little and Vartivarian (2003). The stratification indicator was dropped from the model because, in preliminary simulations, it led to extreme weights in some situations without any further reduction in bias. However, this may be due to the simulation design; if the stratification effect was stronger, the bias reduction could depend on the inclusion or exclusion of the stratification indicator in the model. Future studies should investigate this further.

With these estimated response propensities for the responding PSUs, the population mean was then estimated by:

$$\overline{y}_{ISS.W} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} d_{h\alpha\beta}\,\hat{p}_{h\alpha}^{-1} y_{h\alpha\beta}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r}\sum_{\beta=1}^{b} d_{h\alpha\beta}\,\hat{p}_{h\alpha}^{-1}}$$

Substitution methods

Substitution can be implemented in a variety of forms. This study investigates two features of substitution: the selection method of the substitute and the number of substitution iterations to be used. Each one of them is described next.

First, there are different ways a substitute can be selected to replace a nonresponding unit. In this study two of these forms were evaluated:

- **Matching Substitution (MSub)**: the substituted unit is the closest match to the nonresponding unit in terms of the auxiliary variables observed for all units in the population, such as frame variables. There are several matching procedures studied in the literature (Rosenbaum, 1995). However, a very simple matching rule based on the auxiliary variable X and stratification was used: the substitute was the most similar non-selected unit (and not used in the substitution of any other case) to the nonresponding PSU in terms of X within the same stratum. This is equivalent to selecting the unit within the same stratum with the smallest Euclidean distance from the nonrespondent in terms of the auxiliary variable X. This study does not focus on evaluating other matching methods.

- **Random Substitution (RSub)**: a randomly selected unit among those that were not originally sampled (nor used in the substitution of any other unit) within the same stratum was selected as the substitute the nonresponding PSU. There are different ways of randomly selecting a substitute. Since the original selection of PSUs was done with random systematic-PPS, the substitutes were also selected using this selection procedure. As done in many school-based surveys that use substitution, the selection of the substitutes was conducted during the selection of the original sample. This was implemented by breaking the sampling interval of the systematic selection as many times as needed, according to the number of substitution iterations (see below). Then the PSU that would be the next unit to be selected in the same "zone" (Kish, 1965) of the nonresponding PSU using the new sampling interval is chosen as a substitute.

Another feature of substitution investigated in this study is the number of substitution iterations to be used. As mentioned before, it is possible that the substitute selected to replace a nonresponding unit may be a nonrespondent as well. Then, one can stop with no observation for the designated cluster, keep substituting until a responding unit is found, or keep substituting until a predetermined number of substitution iterations is reached. To study this aspect of substitution, up to one, three, and seven substitution iterations were examined for both MSub and RSub methods. These levels of substitution iteration were chosen to yield response rates approximately equivalent to inflated sample sizes for 90%, 75% and 50% response rates, respectively.

The combination of the substitution selection method and the number of iterations results in a total of six substitution methods (see Table 3.3 above). No further nonresponse adjustments were used to compensate for nonresponse in these substitution methods. The estimate of the population mean under the six substitution methods is similar to the unadjusted respondent mean:

$$\overline{y}_{Sub} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r^*}\sum_{\beta=1}^{b} d_{h\alpha\beta}\, y_{h\alpha\beta}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r^*}\sum_{\beta=1}^{b} d_{h\alpha\beta}}$$

where a_r^* is the number of responding PSUs after all substitution iterations.

Because some nonresponse may remain after all iterations are completed, every substitution method had a corresponding nonresponse weighting-adjusted estimator similar to that used in the nonresponse-adjusted weighted mean ISS.W -- MSub1.W, MSub3.W, MSub7.W, RSub1.W, RSub3.W, and RSub7.W. The cluster response propensities were estimated according to the same logistic regression model for ISS.W, but estimates were computed using data from the responding and all nonresponding units including nonresponding substitutes. The estimator for the population mean of these methods is

$$\overline{y}_{Sub.W} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r^*}\sum_{\beta=1}^{b} d_{h\alpha\beta}\, \hat{p}_{h\alpha}^{-1}\, y_{h\alpha\beta}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_r^*}\sum_{\beta=1}^{b} d_{h\alpha\beta}\, \hat{p}_{h\alpha}^{-1}}$$

Properties evaluated

Three empirical measures were used to compare the different methods. Using $\overline{y}_m$ to denote the general estimator of the population mean for method *m*, the empirical measures computed for each method were

1. The empirical relative bias, $\text{RelBias}\left(\bar{y}_m\right)=\frac{\text{Bias}\left(\bar{y}_m\right)}{\bar{Y}}=\frac{\text{E}\left(\bar{y}_m\right)-\bar{Y}}{\bar{Y}}=\frac{\sum_{k=1}^{5000}\frac{\bar{y}_{m,k}}{5000}-\bar{Y}}{\bar{Y}}$;
2. The empirical sampling variance, $\text{Var}\left(\bar{y}_m\right)=\sum_{k=1}^{5000}\frac{\left[\bar{y}_{m,k}-\text{E}\left(\bar{y}_m\right)\right]^2}{5000}$ where

   $\text{E}\left(\bar{y}_m\right)=\sum_{k=1}^{5000}\frac{\bar{y}_{m,k}}{5000}$; and
3. The empirical root mean square error, $\text{RMSE}\left(\bar{y}_m\right)=\sqrt{\sum_{k=1}^{5000}\frac{\left(\bar{y}_{m,k}-\bar{Y}\right)^2}{5000}}$

with $m=$ ISS, ISS.W, MSub1, MSub3, MSub7, RSub1, RSub3, RSub7, MSub1.W, MSub3.W, MSub7.W, RSub1.W, RSub1.W, RSub3.W, RSub7.W

A potential advantage of substitution over nonresponse weighting adjustment with inflated sample size is that substitution tends to preserve the original sample structure (Vehovar, 1999). This is especially important in designs with few units selected per stratum, such as the sample design used in this simulation that selected two PSUs per stratum. If substitution is not used in cases like this, some strata could potentially end up with no, or only one, observed units due to nonresponse, even when an inflated sample size is used. In this situation, strata may be collapsed to allow for design-based sampling variance estimation (Hansen et al., 1953; Wolter, 2007), which leads to overestimates of sampling variability.

To evaluate the potential benefit of substitution over collapsing strata, the following empirical measures of the sampling variance estimator $\text{var}\left(\bar{y}_m\right)$ were compared across the methods:

1. The empirical relative bias, $\text{RelBias}\left[\text{var}\left(\bar{y}_m\right)\right]=\frac{\text{Bias}\left(\text{var}\left(\bar{y}_m\right)\right)}{\text{Var}\left(\bar{y}_m\right)}=$

$$= \frac{\mathrm{E}\left[\mathrm{var}\left(\bar{y}_m\right)\right]-\mathrm{Var}\left(\bar{y}_m\right)}{\mathrm{Var}\left(\bar{y}_m\right)} = \frac{\sum_{k=1}^{5000}\frac{\mathrm{var}\left(\bar{y}_m\right)}{5000} - \sum_{k=1}^{5000}\frac{\left[\bar{y}_{m,k}-\mathrm{E}\left(\bar{y}_m\right)\right]^2}{5000}}{\sum_{k=1}^{5000}\frac{\left[\bar{y}_{m,k}-\mathrm{E}\left(\bar{y}_m\right)\right]^2}{5000}};$$

2. The empirical sampling variance, $\mathrm{Var}\left[\mathrm{var}\left(\bar{y}_m\right)\right] = \sum_{k=1}^{5000}\frac{\left[\mathrm{var}\left(\bar{y}_{m,k}\right)-\mathrm{E}\left[\mathrm{var}\left(\bar{y}_m\right)\right]\right]^2}{5000}$,

   where $\mathrm{E}\left[\mathrm{var}\left(\bar{y}_m\right)\right] = \sum_{k=1}^{5000}\frac{\mathrm{var}\left(\bar{y}_m\right)}{5000}$; and
3. The empirical root mean square error,

$$\mathrm{RMSE}\left[\mathrm{var}\left(\bar{y}_m\right)\right] = \sqrt{\sum_{k=1}^{5000}\frac{\left[\mathrm{var}\left(\bar{y}_m\right)-\mathrm{Var}\left(\bar{y}_m\right)\right]^2}{5000}} = \sqrt{\sum_{k=1}^{5000}\frac{\left[\mathrm{var}\left(\bar{y}_m\right)-\sum_{k=1}^{5000}\frac{\left[\bar{y}_{m,k}-\mathrm{E}\left(\bar{y}_m\right)\right]^2}{5000}\right]^2}{5000}}$$

with $m =$ ISS, ISS.W, MSub1, MSub3, MSub7, RSub1, RSub3, RSub7, MSub1.W, MSub3.W, MSub7.W, RSub1.W, RSub1.W, RSub3.W, RSub7.W

The variances of all methods were estimated using Taylor Series approximation. Strata with no or one responding PSUs were collapsed with the closest strata until there are at least two PSUs in the collapsed stratum. In some instances of the substitution method, particularly methods which rely on only one iteration of substitution, some strata may be empty or may contain only a single PSU. In these cases, strata were also collapsed to estimate sampling variability.

3.4 Results

In this section the results of the estimators evaluated on this simulation study are shown. First, the properties of the point estimates for the population mean in terms of their empirical relative bias, sampling variance and RMSE are presented. Then, the results of the sampling variance estimates of these estimators are shown with respect of these same measures.

3.4.1 Properties of the Population Mean Estimates

Empirical Relative Bias

Tables 3.4 and 3.5 present the empirical relative bias for the estimates of the population mean under the different methods for each simulation on this study under the MAR and MNAR missing mechanisms, respectively.

As would be expected, as the response rate decreases, the nonresponse bias of the ISS increases, and as the correlation between the auxiliary variable X (which drives the nonresponse mechanism) and the survey variable Y increases, this bias increases. These simulation results also confirm what was found in the bias expression of the unadjusted respondent mean under PSU nonresponse: as the intra-cluster correlation increases, the nonresponse bias also increases. In general, the nonresponse weighting adjustment on ISS.W almost completely eliminates bias under the MAR mechanism, since the model is correctly specified. But there is still substantial remaining bias under the MNAR mechanism, especially for moderate to high intra-cluster correlations and low to moderate response rates.

There is an interesting difference between the matching and random selection substitution methods relative to the number of iterations. As the number of substitution iterations increases, the empirical (absolute) relative bias of the matching substitution methods (MSub) decreases and approaches that of the ISS.W. This is evident for the MNAR mechanism, in which this pattern is identified in most of the cases and the relative bias of the MSub7 is very close to that of the ISS.W. Although this same general pattern is not as clear for MAR, the average empirical absolute relative bias of the MSub across all the simulation settings tends to get closer to that of the ISS.W. Also, for MAR, the bias of the MSub7 is, on average, larger than of the ISS.W.

In contrast, for the random substitution methods (RSub), not only does its empirical relative bias remain almost constant across iterations, but it also is very close to the relative bias of the ISS. This difference between MSub and RSub is explained by the fact that while the former seeks to select substitutes that resemble as closely as possible the nonresponding PSUs in terms of the auxiliary variable *X*, the latter only draws a random PSU that, despite being in the same

Table 3.4: Empirical Relative Bias (x 10^5) - MAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	13	8	-28	-54	-50	14	18	18	-36	-62	-57	11	15	14
		0.30	49	-9	8	-25	-33	55	54	54	-51	-80	-86	-5	-5	-5
		0.70	156	-7	106	27	-7	171	181	181	-58	-132	-164	10	18	18
	0.20	0.01	-18	-4	-136	-212	-216	-10	-10	-8	-138	-212	-215	5	5	7
		0.30	296	6	69	-71	-142	279	272	276	-205	-339	-407	-15	-18	-16
		0.70	712	15	402	68	-65	726	725	721	-291	-597	-714	24	14	10
	0.50	0.01	39	-27	-143	-174	-190	59	60	62	-178	-220	-230	-4	-4	-1
		0.30	412	-34	256	-7	-111	456	462	464	-180	-424	-524	19	22	26
		0.70	1,070	5	553	120	-70	1,078	1,084	1,080	-503	-896	-1,061	10	4	5
0.75	0.01	0.01	12	9	-16	-26	-26	13	11	12	-19	-29	-29	11	9	9
		0.30	39	10	-8	-23	-25	24	25	25	-39	-54	-55	-6	-4	-4
		0.70	81	-3	42	5	0	98	100	101	-43	-79	-83	15	17	18
	0.20	0.01	-8	-3	-94	-114	-116	14	17	16	-92	-111	-112	19	22	21
		0.30	143	-13	-17	-80	-88	143	146	146	-163	-223	-230	-7	-4	-3
		0.70	348	-9	120	-20	-37	382	379	378	-236	-373	-388	25	22	21
	0.50	0.01	52	20	-86	-104	-106	36	22	21	-112	-131	-132	2	-12	-13
		0.30	193	-29	51	-53	-61	215	225	226	-173	-273	-280	-9	2	2
		0.70	530	-14	111	-78	-99	550	553	553	-435	-613	-632	0	2	2
0.90	0.01	0.01	-8	-8	-3	-4	-4	11	10	10	-4	-5	-5	10	9	9
		0.30	26	15	-11	-15	-15	5	6	6	-22	-26	-26	-6	-5	-5
		0.70	31	-3	18	10	10	50	51	51	-17	-24	-24	15	16	16
	0.20	0.01	8	10	-24	-28	-28	18	21	21	-22	-26	-26	20	23	23
		0.30	49	-12	-15	-29	-30	65	66	66	-74	-87	-88	3	5	5
		0.70	146	-9	12	-20	-21	167	167	167	-141	-171	-172	13	14	14
	0.50	0.01	12	9	-34	-43	-43	9	11	11	-37	-47	-47	8	10	10
		0.30	62	-39	-58	-85	-86	98	97	97	-155	-183	-184	1	1	1
		0.70	243	12	-5	-46	-46	215	216	217	-231	-269	-269	-9	-7	-6
Mean			174	-4	40	-40	-63	183	184	184	-135	-211	-231	6	6	7
Mean of absolute bias			176	13	90	57	64	184	185	185	135	211	231	10	11	11

Table 3.5: Empirical Relative Bias (x 10^5) - MNAR mechanism

$\bar{p}$	ρ	$Corr_s(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	37	32	-3	-29	-24	38	42	40	-11	-37	-31	35	39	37
		0.30	74	15	33	1	-8	79	78	78	-29	-57	-63	18	17	17
		0.70	181	8	124	42	6	193	203	202	-50	-128	-162	21	30	28
	0.20	0.01	484	500	371	289	283	490	485	483	373	293	287	508	502	500
		0.30	793	468	563	403	320	772	754	749	261	109	29	445	429	424
		0.70	1,170	270	794	386	208	1,171	1,165	1,159	-103	-480	-639	272	256	247
	0.50	0.01	1,286	1,229	1,091	1,041	1,015	1,281	1,242	1,233	1,065	1,000	979	1,229	1,186	1,178
		0.30	1,618	1,110	1,408	1,113	993	1,641	1,632	1,624	903	628	513	1,135	1,128	1,124
		0.70	2,160	636	1,492	892	609	2,151	2,157	2,140	-36	-594	-844	616	610	597
0.75	0.01	0.01	24	21	-3	-13	-13	25	23	24	-6	-16	-16	23	21	22
		0.30	52	22	5	-11	-12	35	37	37	-27	-42	-43	5	7	7
		0.70	94	4	51	12	7	110	112	113	-40	-78	-83	22	24	24
	0.20	0.01	247	253	165	139	137	267	266	265	168	144	142	273	272	271
		0.30	395	220	225	155	144	393	391	391	62	-5	-15	224	222	222
		0.70	590	122	295	116	94	611	606	603	-174	-349	-370	143	137	136
	0.50	0.01	680	650	533	499	495	642	621	619	510	473	470	612	590	588
		0.30	813	552	638	517	505	827	830	832	372	255	244	563	569	571
		0.70	1,241	309	594	269	224	1,257	1,252	1,249	-334	-650	-692	345	341	339
0.90	0.01	0.01	-3	-3	1	-1	0	16	15	15	0	-2	-1	15	14	14
		0.30	33	21	-6	-10	-10	10	11	11	-17	-22	-22	-1	0	0
		0.70	36	-1	20	12	12	54	55	55	-17	-24	-25	17	18	18
	0.20	0.01	114	115	78	73	73	120	122	122	80	75	75	122	123	123
		0.30	155	86	85	69	69	167	168	168	19	4	3	97	98	98
		0.70	267	68	95	55	54	261	261	261	-102	-142	-143	64	64	64
	0.50	0.01	229	222	222	222	221	251	250	251	218	218	217	246	246	246
		0.30	323	204	204	181	181	323	324	324	85	63	63	201	202	202
		0.70	485	139	149	88	87	453	457	457	-194	-252	-253	114	118	119
Mean			503	269	342	241	210	505	502	500	110	14	-14	273	269	267
Mean of absolute bias			503	270	342	246	215	505	502	500	195	227	238	273	269	267

stratum as the nonrespondent, may still be quite different from it. The matching selection of the substitution performs a similar role to the nonresponse weighting-adjustments in reducing bias. Hence, if X is associated with Y, it would be expected that the matching substitutes would be, on average, more similar to the nonrespondents than the random substitutes. This is confirmed by the fact that as this correlation between X and Y increases, the absolute relative bias of the MSub *decreases*, whereas the relative bias of the RSub *increases*.

The results from the nonresponse weighting-adjusted substitution methods reveal another difference between the two substitution selection approaches. While for RSub.W the nonresponse weighting adjustment substantially diminishes the relative bias to levels comparable that of the ISS.W, this same adjustment procedure tends to increase the absolute relative bias in the MSub.W. This latter phenomenon seems to be particularly true for a MAR mechanism, where it occurs consistently across the levels of response rate, intra-cluster correlation, and correlation between X and Y. This happens because the matching substitution already is a nonresponse adjustment; imposing another adjustment on top of that, through weighting, disrupts the adjustments made in the first place. This finding is corroborated by the fact that the increase in bias from an unadjusted to a weighting-adjusted approach tends to be larger for RSub7.W and to increase as the correlation between X and Y increases. On the other hand, for MNAR this phenomenon is less evident, and in some cases the nonresponse adjustments seems to actually reduce the bias in MSub.W.

Empirical Variance

The results for the empirical variance, presented in Tables 3.6 and 3.7 show that for stronger correlations between X and Y, such as $Corr_a\left(Y,X\right)=0.70$, the sampling variance of the ISS.W tends to be smaller than that of ISS, as predicted from the results of Little and Vartivarian (2005). This is particularly true for moderate to higher levels of intra-cluster correlation ($\rho=0.20$ and $\rho=0.50$), under both missing mechanisms across the different response rate levels. A similar pattern is observed for the MSub7, RSub7.W and MSub7.W, but not for corresponding methods using a smaller number of iterations. Also, since RSub7 does not use the auxiliary variable X for the selection of substitutes nor for weighting adjustments, it does not have the same gains in precision found in the other three methods.

Table 3.6: Empirical Variance ($x10^2$) - MAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	30	31	41	33	31	41	32	30	42	33	31	42	33	30
		0.30	31	32	41	33	30	40	32	31	42	34	31	41	33	31
		0.70	30	30	39	31	29	39	32	30	40	32	30	39	31	30
	0.20	0.01	100	103	134	105	96	130	104	97	138	107	98	133	106	99
		0.30	95	94	128	101	96	125	99	92	131	107	102	125	99	92
		0.70	94	81	126	97	90	125	97	91	121	104	102	104	80	75
	0.50	0.01	197	201	281	216	201	268	206	192	287	220	205	276	213	199
		0.30	194	195	285	222	206	267	210	194	291	231	218	275	213	196
		0.70	182	152	247	190	176	245	193	181	237	208	206	202	156	146
0.75	0.01	0.01	27	27	32	31	31	33	30	30	33	31	31	33	30	30
		0.30	27	27	32	30	30	32	30	30	32	31	30	32	30	30
		0.70	27	27	31	29	29	31	29	29	32	30	30	31	29	29
	0.20	0.01	82	83	104	98	97	104	98	98	105	99	98	105	99	99
		0.30	81	81	101	95	95	101	93	93	101	97	96	101	93	93
		0.70	78	70	99	91	91	100	92	92	101	100	100	91	85	84
	0.50	0.01	172	174	213	197	196	214	199	198	214	198	197	216	201	200
		0.30	169	169	216	201	201	213	198	197	221	208	208	215	200	199
		0.70	158	145	193	181	181	190	178	178	203	204	206	172	160	160
0.90	0.01	0.01	33	33	31	31	31	31	30	30	31	31	31	31	30	30
		0.30	33	33	31	30	30	31	30	30	31	30	30	30	30	30
		0.70	33	33	30	29	29	30	30	30	30	30	30	30	30	30
	0.20	0.01	108	109	100	98	98	98	97	97	100	99	99	99	98	98
		0.30	101	101	94	93	93	95	94	94	94	93	93	95	94	94
		0.70	100	96	89	87	87	87	86	86	92	92	92	84	83	83
	0.50	0.01	199	199	180	177	177	178	177	177	179	177	177	177	176	176
		0.30	197	196	177	174	174	178	175	175	178	176	177	176	174	174
		0.70	206	197	182	180	180	180	178	178	191	192	192	174	172	172
Mean			103	101	121	107	104	119	105	103	122	111	109	116	103	100

Table 3.7: Empirical Variance (x10^2) - MNAR mechanism

p	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	30	31	41	33	31	41	32	30	42	34	32	42	33	31
		0.30	31	32	41	33	30	40	32	31	42	34	31	41	33	31
		0.70	30	30	39	31	29	39	32	30	40	32	30	40	32	30
	0.20	0.01	100	103	135	105	98	130	104	97	139	108	100	133	107	99
		0.30	93	93	128	101	96	127	99	93	132	107	102	128	100	93
		0.70	91	82	126	97	90	123	95	89	122	105	103	104	81	75
	0.50	0.01	191	195	275	216	200	266	208	195	283	221	204	274	216	203
		0.30	191	194	280	221	208	259	206	191	288	231	219	267	210	194
		0.70	174	154	244	190	173	236	184	173	244	214	209	205	159	148
0.75	0.01	0.01	27	27	32	31	31	32	30	30	33	31	31	33	30	30
		0.30	26	27	32	30	30	32	30	30	32	31	30	33	30	30
		0.70	27	27	31	29	29	31	29	29	32	30	30	31	29	29
	0.20	0.01	82	83	103	97	96	104	98	97	104	97	97	105	99	98
		0.30	82	82	100	94	94	100	92	92	101	96	96	100	92	92
		0.70	79	70	99	92	91	100	93	92	102	100	101	92	86	85
	0.50	0.01	171	173	215	198	197	215	200	199	217	200	199	217	203	202
		0.30	168	169	213	198	197	213	197	196	218	206	205	215	198	198
		0.70	170	157	208	194	193	207	192	191	220	224	226	184	170	169
0.90	0.01	0.01	33	33	31	31	31	31	30	30	31	31	31	31	30	30
		0.30	33	33	31	30	30	31	30	30	31	30	30	30	30	30
		0.70	33	33	30	29	29	30	30	30	30	30	30	30	30	30
	0.20	0.01	107	108	99	97	98	98	97	97	99	98	98	99	98	98
		0.30	102	102	93	92	92	95	94	93	94	93	93	95	94	93
		0.70	100	98	93	91	91	94	92	92	97	97	97	92	90	90
	0.50	0.01	209	209	178	176	176	181	178	178	178	177	177	181	179	179
		0.30	220	219	191	188	188	190	188	188	192	191	191	190	188	188
		0.70	204	198	180	178	178	178	175	175	192	193	193	174	172	172
Mean			104	102	121	108	105	119	106	104	124	113	111	117	104	102

For both nonresponse mechanisms, across the levels of the three parameters manipulated in this simulation study, the variance of both RSub1 and MSub1 tend to be higher than the ISS and ISS.W. This is expected, since their sample size is smaller, due to remaining nonresponse after the substitution. That is, if there were nonrespondents among the substitutes, no further attempt was made to continue substituting the nonrespondents on these methods. This difference in the sampling variance tends to decrease as the number of iterations (and therefore, the observed sample size) increases to the point that the variances of both RSub7 and MSub7 are very close that of the ISS and ISS.W.

In general, the variances of RSub and MSub are very close to the corresponding sampling variances for RSub.W and MSub.W. The only exceptions occur for high levels of intra-cluster correlation ($\rho = 0.50$) and high correlations between Y and X ($Corr_a(Y, X) = 0.70$) when RSub.W tends to produce smaller sampling variances compared to the corresponding RSub. This occurs since these are cases in which gains in precision with nonresponse weighting-adjustments are expected. For these same situations, the opposite effect is observed for MSub.W. The MSub.W sampling variances are larger than the corresponding MSub sampling variances, similar to what happens for bias. This confirms that making further adjustments on estimates relying on matching substitution using the same set of covariates might have a negative effect on the properties of the final estimates.

Empirical RMSE

When both bias and variance are taken into account through the RMSE, the patterns identified for the bias and variance are repeated (see Tables 3.8 and 3.9). The RMSE of ISS.W is, in general, smaller than that of ISS. This occurs particularly as the correlation between Y and X and the intra-cluster correlation increase, and as the response rate decreases, under both missing mechanisms.

For the MNAR mechanism, it is clear that MSub1 has a larger RMSE than ISS.W, but as the number of iterations increase up to seven in MSub7, for example, the RMSE approaches and becomes very close that of ISS.W. On the other hand, the RMSE of RSub tends to be slightly larger than that of ISS, approaching it as the number of substitution iterations increases. The only

Table 3.8: Empirical RMSE (x 10^2) - MAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	55	56	64	57	56	64	57	55	65	58	56	65	57	55
		0.30	56	56	64	57	55	63	57	56	65	59	57	64	57	56
		0.70	57	55	64	56	54	65	59	58	64	58	57	63	56	54
	0.20	0.01	100	102	117	105	101	114	102	98	118	106	102	115	103	99
		0.30	102	97	113	101	99	115	103	100	117	109	109	112	100	96
		0.70	120	90	119	99	95	133	122	120	114	118	124	102	89	86
	0.50	0.01	140	142	168	148	143	164	144	139	170	150	145	166	146	141
		0.30	145	140	171	149	144	170	152	147	172	158	157	166	146	140
		0.70	172	123	167	138	133	190	176	173	162	170	179	142	125	121
0.75	0.01	0.01	52	52	57	55	55	57	55	55	57	56	56	57	55	55
		0.30	52	52	57	55	55	57	55	55	57	55	55	57	55	55
		0.70	53	52	56	54	54	56	55	55	57	55	55	55	54	54
	0.20	0.01	91	91	103	100	99	102	99	99	103	100	100	102	100	99
		0.30	91	90	101	98	98	102	98	97	102	101	101	100	96	96
		0.70	95	83	100	96	95	107	103	103	103	107	107	96	92	92
	0.50	0.01	131	132	146	141	140	146	141	141	147	141	141	147	142	141
		0.30	131	130	147	142	142	148	143	142	150	147	147	147	141	141
		0.70	137	120	139	135	135	148	144	144	149	156	157	131	127	126
0.90	0.01	0.01	58	58	56	55	55	55	55	55	56	55	55	55	55	55
		0.30	58	58	55	55	55	55	55	55	56	55	55	55	55	55
		0.70	58	58	54	54	54	55	55	55	55	54	54	55	55	55
	0.20	0.01	104	104	100	99	99	99	99	99	100	99	99	99	99	99
		0.30	101	101	97	96	96	98	97	97	97	97	97	98	97	97
		0.70	101	98	94	93	93	95	94	94	97	97	97	92	91	91
	0.50	0.01	141	141	134	133	133	133	133	133	134	133	133	133	133	133
		0.30	141	140	133	132	132	134	133	133	134	134	134	133	132	132
		0.70	145	140	135	134	134	136	135	135	140	141	141	132	131	131
Mean			100	95	104	98	97	106	101	100	105	103	103	101	96	95

Table 3.9: Empirical RMSE (x 10^2) - MNAR mechanism

$\mathcal{P}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	55	56	64	57	56	64	57	55	65	58	56	65	57	56
		0.30	56	56	64	57	55	64	57	56	65	58	56	64	58	56
		0.70	58	55	64	56	54	66	60	58	64	58	57	63	56	55
	0.20	0.01	111	113	122	107	103	124	113	110	124	108	104	126	115	112
		0.30	125	107	126	108	103	137	125	122	118	104	101	122	109	105
		0.70	151	95	138	106	97	161	152	150	111	113	120	106	94	90
	0.50	0.01	189	186	199	181	174	208	191	187	199	180	174	207	189	185
		0.30	214	179	219	186	175	231	218	214	193	165	157	199	184	180
		0.70	254	139	216	164	145	265	256	252	156	158	168	156	140	136
0.75	0.01	0.01	52	52	57	55	55	57	55	55	57	56	55	57	55	55
		0.30	52	52	57	55	55	57	55	55	57	55	55	57	55	55
		0.70	53	52	56	54	54	57	55	55	57	55	55	56	54	54
	0.20	0.01	94	95	103	99	99	105	102	102	103	100	99	106	103	103
		0.30	99	93	103	98	98	108	104	104	101	98	98	103	99	98
		0.70	107	85	104	96	96	117	114	114	102	106	107	97	94	93
	0.50	0.01	147	147	156	150	149	160	155	154	156	149	149	160	154	154
		0.30	153	141	159	150	149	168	163	163	152	146	145	157	152	152
		0.70	181	129	156	142	141	192	187	187	152	163	166	140	135	134
0.90	0.01	0.01	58	58	56	55	55	55	55	55	56	55	55	55	55	55
		0.30	58	58	55	55	55	55	55	55	55	55	55	55	55	55
		0.70	58	58	54	54	54	55	55	55	55	54	54	55	55	55
	0.20	0.01	104	105	100	99	99	100	99	99	100	99	99	100	100	100
		0.30	102	101	97	96	96	99	98	98	97	96	96	98	97	97
		0.70	104	99	97	96	96	100	100	100	99	99	100	96	95	95
	0.50	0.01	146	146	135	135	135	137	136	136	135	135	135	137	136	136
		0.30	152	150	140	138	138	142	141	141	139	138	138	139	138	138
		0.70	151	141	135	134	134	141	140	140	140	141	141	132	132	132
Mean			114	102	112	103	101	119	115	114	108	104	104	108	102	101

exception is for a high response rate ($\bar{p} = 0.90$), when the RMSE of the RSub is very close to, and sometimes slightly smaller than, that of ISS. For the MAR mechanism, the pattern for RSub is similar, but for MSub is not as clear. First, for a high response rate ($\bar{p} = 0.90$) MSub seems to have a smaller RMSE than that of ISS.W, especially as the number of substitution iterations increases. However, for low and moderate response rates ($\bar{p} = 0.50$ and $\bar{p} = 0.75$), the ISS.W performs better than the MSub, even when the number of substitution iterations is high (MSub7), although the gap on the RMSE between these two approaches gets smaller as the number of substitution iterations increases.

As a result of the disruption on the adjustment of the matching substitution caused by the nonresponse weighting-adjustment on the same variables used to match the substitutes to the nonrespondents, as found in the previous sections, the accuracy of the MSub.W tends to be worse than that of the MSub, in particular for substitutions up to seven iterations. On the other hand, the RSub.W estimates have a smaller RMSE than their corresponding estimates without the nonresponse weighting adjustment and the improvements tend to be the same, regardless of the number of substitution iterations.

3.4.2 Properties of the Sampling Variance Estimates

Empirical Relative Bias

Tables 3.10 and 3.11 show the empirical relative bias of the estimates of the sampling variance of the different estimators for the population mean under MAR and MNAR mechanisms, respectively.

Most of the estimates tend to overestimate the sampling variability due to the fact that strata collapsing was used in every alternative when necessary, that is, when strata had no or one responding cluster. Bias in sampling variance increases with the intra-cluster correlation of the survey variable at the same degree across all methods and parameters investigated in this study. On the other hand, there seems to be no relationship between the bias of the sampling variance estimates and the other parameters used in these simulations, such as the correlation between the auxiliary covariate and the survey variable, the response rate, or the missing data mechanism.

Table 3.10: Empirical Relative Bias (x 10^4) of Sampling Variance Estimates - MAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	73	-13	30	-11	-9	-85	163	215	78	10	1	13	219	287
		0.30	-215	-86	95	29	84	275	89	-58	109	-23	24	332	127	-24
		0.70	6	412	378	479	585	250	168	145	508	426	422	633	570	541
	0.20	0.01	99	125	429	594	758	353	266	371	482	631	814	412	340	462
		0.30	745	1,166	1,004	913	757	880	967	1,077	1,007	651	406	1,202	1,248	1,408
		0.70	28	2,341	66	315	404	116	484	469	1,167	118	-464	3,122	3,597	3,625
	0.50	0.01	843	957	341	666	678	698	983	1,087	453	746	759	744	937	1,046
		0.30	881	1,160	-297	-190	-97	416	577	780	-131	-244	-306	541	813	1,056
		0.70	520	3,467	423	848	946	448	565	602	1,651	407	-244	3,641	4,035	4,095
0.75	0.01	0.01	269	293	4	-65	-64	-6	100	62	49	-33	-31	43	129	93
		0.30	258	244	169	155	136	106	162	168	158	119	97	132	174	176
		0.70	-126	29	471	433	463	503	509	487	448	307	313	698	715	690
	0.20	0.01	1,201	1,188	691	704	697	413	337	324	706	716	712	441	360	346
		0.30	1,048	1,253	1,006	930	910	817	1,006	991	1,063	841	790	973	1,150	1,132
		0.70	1,020	2,811	175	316	320	108	303	316	235	-262	-341	1,413	1,590	1,611
	0.50	0.01	1,200	1,207	929	1,052	1,079	800	914	933	988	1,087	1,108	854	918	938
		0.30	930	1,093	42	93	86	327	440	436	11	-63	-82	442	553	552
		0.70	797	2,251	754	760	723	873	886	848	614	-148	-262	2,451	2,561	2,544
0.90	0.01	0.01	184	160	-202	-131	-131	-4	33	35	-199	-122	-121	18	52	53
		0.30	236	243	17	42	43	44	54	56	33	51	52	98	108	110
		0.70	159	201	457	431	430	167	163	159	438	387	384	253	248	243
	0.20	0.01	593	593	485	541	532	504	491	492	455	512	503	462	448	449
		0.30	1,149	1,203	1,152	1,146	1,144	867	897	905	1,162	1,110	1,106	937	960	968
		0.70	381	1,033	589	652	655	757	782	773	438	333	330	1,389	1,414	1,407
	0.50	0.01	610	602	733	744	747	872	828	826	772	785	789	920	868	867
		0.30	980	1,106	1,043	1,083	1,082	1,025	1,067	1,065	996	992	987	1,149	1,184	1,183
		0.70	526	1,263	838	865	865	907	954	956	611	445	442	1,563	1,614	1,616
Mean			533	974	438	496	512	460	525	538	530	362	303	921	997	1,018
Mean of absolute bias			558	981	475	526	534	467	525	542	554	428	440	921	997	1,019

Table 3.11: Empirical Relative Bias (x 10^4) of Sampling Variance Estimates - MNAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	99	12	27	-98	-79	-87	113	154	63	-94	-80	16	161	208
		0.30	-208	-89	47	-15	75	164	-3	-133	77	-70	17	238	61	-71
		0.70	47	471	404	474	579	217	114	92	533	417	416	607	527	512
	0.20	0.01	65	116	360	557	622	362	304	382	409	557	642	407	320	425
		0.30	909	1,273	959	886	754	707	854	974	930	616	384	972	1,144	1,311
		0.70	155	2,408	-63	272	341	207	532	635	1,195	167	-424	3,199	3,571	3,732
	0.50	0.01	1,080	1,195	461	543	644	676	817	852	515	583	694	734	740	764
		0.30	881	1,134	-322	-308	-317	570	609	769	-152	-322	-472	710	870	1,019
		0.70	777	3,714	425	729	934	665	838	923	1,708	409	-205	3,818	4,172	4,277
0.75	0.01	0.01	289	311	-1	-49	-50	-7	93	69	40	-13	-14	29	112	89
		0.30	270	254	173	133	107	62	95	97	160	96	67	90	106	106
		0.70	-148	12	453	420	442	454	418	408	425	291	290	642	620	608
	0.20	0.01	1,184	1,198	762	795	813	434	370	360	799	833	855	445	382	371
		0.30	908	1,116	1,098	1,010	992	855	1,068	1,061	1,101	870	819	994	1,203	1,200
		0.70	904	2,764	125	275	284	-29	118	141	289	-237	-336	1,387	1,512	1,542
	0.50	0.01	1,262	1,213	756	904	909	735	811	826	786	927	933	758	792	811
		0.30	937	1,041	135	154	160	250	426	414	115	-1	-11	361	541	526
		0.70	483	2,131	545	547	531	381	533	549	561	-357	-530	2,460	2,718	2,749
0.90	0.01	0.01	190	165	-205	-126	-127	-10	21	21	-202	-119	-119	12	40	41
		0.30	220	226	24	41	42	43	56	58	39	48	49	95	110	111
		0.70	149	194	437	410	409	201	205	198	422	366	364	281	285	278
	0.20	0.01	660	651	576	626	619	509	485	490	556	606	598	484	461	465
		0.30	1,096	1,142	1,175	1,169	1,167	913	931	944	1,171	1,118	1,113	978	991	1,005
		0.70	390	914	219	252	250	114	146	144	39	-106	-111	609	631	629
	0.50	0.01	909	900	1,600	1,587	1,590	1,469	1,508	1,501	1,597	1,578	1,579	1,450	1,486	1,479
		0.30	209	316	642	645	644	663	695	692	629	579	579	759	789	786
		0.70	530	1,298	930	928	928	992	1,021	1,020	631	408	401	1,637	1,680	1,679
Mean			528	966	435	473	491	426	488	505	535	339	278	895	964	987
Mean of absolute bias			554	972	479	517	534	436	488	515	561	437	448	895	964	992

Compared to the sampling variance estimates obtained by the nonresponse weighting adjustment method (ISS.W) with collapsing, those obtained by the substitution methods tend to be less biased. Curiously, the sampling variance estimates obtained by the unadjusted respondent mean (ISS) were slightly less biased than the ones provided by the nonresponse weighted-adjusted mean. Although increasing the number of substitution iterations would tend to produce a sample with a smaller number of strata with none or one responding cluster, this does not seem to be a factor on the bias of the sampling variance estimates. Moreover, there was not much difference in the sampling variance estimates produced by the substitution method with or without a further nonresponse weighting adjustment.

Empirical Variance

The empirical variances of the sampling variance estimates for the MAR and MNAR mechanisms are shown in Tables 3.12 and 3.13, respectively. In general, as the intra-cluster correlation of the survey variable increases the sampling variance of these estimates is smaller, but the variances do not seem to be related with the correlation between the auxiliary and survey variables, the response rate, nor the missing data mechanism.

There was also not much difference between the different methods evaluated in this study in terms of the variances of their sampling variance estimates across the different simulation parameters. However, as the number of substitution iterations increases, the variability of these sampling variance estimates decreases. This is to be expected, since the sample size increases as a function of the number of substitution iterations. Further, nonresponse weighting adjustment on these substitution methods slightly increases the variability of their sampling variance estimates, which is also an expected result, since weighting can increase sampling variability, and presumably the variance of the sampling variance.

Empirical RMSE

Since there are not many differences between the methods in terms of the empirical variances of the sampling variance estimates, the difference in the results of the empirical RMSE of these sampling variance estimates are mostly due to what was found in their empirical bias. Therefore, as can be seen in Tables 3.14 and 3.15 for MAR and MNAR nonresponse mecha-

Table 3.12: Empirical Variance of Sampling Variance Estimates (x10^4) - MAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	36	41	85	43	36	85	45	38	108	50	42	104	52	44
		0.30	36	43	84	43	37	84	43	37	103	51	42	101	50	41
		0.70	35	45	82	43	37	80	42	36	99	51	43	101	52	43
	0.20	0.01	368	433	950	491	409	892	453	382	1,129	567	463	1,055	516	435
		0.30	389	508	1,001	503	422	954	474	402	1,403	657	510	1,177	581	494
		0.70	320	573	802	406	339	793	409	343	1,357	593	469	1,542	757	603
	0.50	0.01	1,639	1,968	4,239	2,164	1,759	4,027	2,078	1,754	5,117	2,451	1,950	4,973	2,466	2,049
		0.30	1,636	2,112	3,698	1,831	1,550	3,896	1,967	1,683	4,955	2,321	1,892	5,301	2,727	2,318
		0.70	1,295	2,674	3,181	1,637	1,385	3,189	1,647	1,379	6,463	2,644	2,081	6,485	3,387	2,533
0.75	0.01	0.01	24	25	45	38	37	45	39	38	48	40	39	48	41	40
		0.30	24	26	43	36	36	44	37	36	46	38	38	46	38	38
		0.70	22	25	44	37	37	44	37	37	48	40	39	47	39	39
	0.20	0.01	253	271	494	415	410	457	384	380	522	436	430	490	410	405
		0.30	251	282	513	427	422	478	401	396	572	466	458	538	443	438
		0.70	231	288	403	344	340	396	341	341	552	434	418	497	422	419
	0.50	0.01	1,124	1,172	2,199	1,822	1,795	2,155	1,820	1,797	2,337	1,904	1,870	2,295	1,929	1,903
		0.30	1,053	1,181	1,877	1,574	1,563	1,993	1,670	1,645	2,146	1,744	1,724	2,228	1,839	1,807
		0.70	943	1,393	1,784	1,481	1,473	1,692	1,432	1,419	2,430	1,934	1,900	2,307	1,908	1,886
0.90	0.01	0.01	50	51	38	37	37	38	37	37	39	38	38	39	37	37
		0.30	49	51	36	36	36	37	36	36	37	36	36	38	36	36
		0.70	50	52	37	37	37	38	37	37	38	37	37	39	38	38
	0.20	0.01	562	575	416	404	404	394	384	384	422	410	410	402	391	391
		0.30	526	554	440	429	429	424	412	412	458	447	446	446	431	431
		0.70	453	495	335	327	328	333	324	324	362	351	352	364	353	353
	0.50	0.01	1,792	1,837	1,458	1,409	1,409	1,483	1,439	1,438	1,486	1,434	1,433	1,518	1,470	1,470
		0.30	1,768	1,803	1,310	1,271	1,269	1,370	1,331	1,331	1,335	1,291	1,289	1,392	1,350	1,350
		0.70	1,964	2,511	1,559	1,510	1,510	1,496	1,457	1,457	1,807	1,717	1,715	2,408	2,357	2,357
Mean			626	777	1,006	696	650	997	695	652	1,312	822	747	1,333	893	815

Table 3.13: Empirical Variance of Sampling Variance Estimates (x10^4) - MNAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	35	41	85	43	37	85	45	38	107	50	42	104	52	44
		0.30	36	43	84	43	37	84	43	36	103	52	42	102	51	41
		0.70	34	45	82	43	37	80	42	36	101	51	43	103	52	44
	0.20	0.01	358	424	930	481	407	890	453	380	1,096	558	462	1,067	518	430
		0.30	380	499	971	496	418	909	457	386	1,446	648	505	1,199	593	508
		0.70	311	722	776	399	332	770	404	340	1,719	752	568	1,841	844	704
	0.50	0.01	1,598	1,929	4,092	2,067	1,699	3,876	2,017	1,727	4,955	2,353	1,886	4,782	2,407	2,023
		0.30	1,624	2,248	3,568	1,793	1,525	3,729	1,865	1,621	5,087	2,392	1,934	5,104	2,645	2,288
		0.70	1,228	4,005	3,135	1,601	1,352	3,047	1,567	1,317	10,493	4,125	3,019	10,126	5,102	3,893
0.75	0.01	0.01	24	25	45	38	37	45	39	38	48	40	39	48	40	40
		0.30	24	26	43	36	36	44	37	36	46	38	38	47	38	38
		0.70	22	25	44	37	37	44	37	37	48	40	39	47	40	40
	0.20	0.01	254	274	492	410	405	454	384	380	519	432	426	486	408	403
		0.30	248	284	513	428	422	475	399	395	577	471	462	554	451	443
		0.70	226	311	401	345	341	392	335	334	581	497	482	562	468	463
	0.50	0.01	1,099	1,148	2,133	1,797	1,767	2,133	1,806	1,780	2,277	1,888	1,853	2,279	1,922	1,893
		0.30	1,064	1,229	1,862	1,545	1,536	1,978	1,655	1,632	2,163	1,756	1,732	2,273	1,870	1,847
		0.70	996	1,475	1,917	1,589	1,573	1,824	1,552	1,537	2,829	2,230	2,162	2,842	2,276	2,251
0.90	0.01	0.01	50	51	38	37	37	38	37	37	39	38	38	39	37	37
		0.30	49	51	37	36	36	37	36	36	37	36	36	38	36	36
		0.70	50	52	37	37	37	38	37	37	38	37	37	39	38	38
	0.20	0.01	564	575	418	407	407	395	387	386	425	413	413	403	395	394
		0.30	522	556	438	426	426	423	410	411	452	439	439	441	428	428
		0.70	439	529	346	336	336	331	323	323	398	384	384	386	375	375
	0.50	0.01	1,944	1,972	1,563	1,525	1,525	1,568	1,523	1,522	1,576	1,537	1,537	1,579	1,533	1,532
		0.30	2,030	2,097	1,518	1,474	1,472	1,549	1,490	1,490	1,556	1,503	1,502	1,583	1,521	1,521
		0.70	1,948	3,288	1,550	1,498	1,497	1,472	1,431	1,432	1,974	1,854	1,849	3,189	3,136	3,135
Mean			636	886	1,004	702	658	989	697	656	1,507	912	814	1,528	1,010	922

Table 3.14: Empirical RMSE Variance Estimates (x10^3) - MAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	60	64	92	65	60	92	67	62	104	71	65	102	73	67
		0.30	60	65	92	66	61	92	66	60	101	72	65	101	71	64
		0.70	59	68	92	67	63	90	65	60	102	73	66	104	74	68
	0.20	0.01	192	209	314	230	215	302	215	199	343	248	230	329	230	214
		0.30	210	251	341	242	218	328	238	224	397	266	230	375	271	257
		0.70	179	306	283	204	188	282	207	190	394	244	222	509	397	365
	0.50	0.01	438	483	658	487	441	662	499	468	727	522	468	735	535	498
		0.30	439	513	614	430	394	634	460	437	705	485	440	743	550	524
		0.70	372	739	574	436	408	575	420	387	894	521	459	1,090	856	781
0.75	0.01	0.01	49	51	67	61	61	67	62	62	69	63	63	69	64	63
		0.30	49	51	66	61	60	66	61	60	68	62	61	68	62	62
		0.70	47	50	68	62	62	68	62	62	71	64	63	72	66	66
	0.20	0.01	187	192	234	215	214	218	199	198	240	220	219	226	206	204
		0.30	180	196	248	225	223	234	221	219	262	231	227	252	236	234
		0.70	172	259	202	188	187	199	187	187	236	210	207	257	246	246
	0.50	0.01	394	401	509	474	473	495	464	462	528	487	484	513	476	475
		0.30	361	390	433	397	396	452	418	415	463	418	416	482	443	439
		0.70	332	496	447	409	405	443	410	406	508	441	439	639	599	595
0.90	0.01	0.01	71	71	62	61	61	62	61	61	63	62	62	62	61	61
		0.30	70	72	60	60	60	61	60	60	61	60	60	61	60	60
		0.70	71	72	63	62	62	62	61	61	63	62	62	63	62	62
	0.20	0.01	246	248	210	208	208	205	202	202	210	209	208	206	203	203
		0.30	257	265	236	233	233	222	220	220	240	235	235	229	226	227
		0.70	216	244	190	190	190	194	192	192	194	190	190	224	222	221
	0.50	0.01	440	445	404	398	398	415	407	406	410	403	404	422	413	413
		0.30	463	477	406	403	403	413	410	410	406	400	399	425	421	421
		0.70	456	560	423	419	419	420	418	418	441	423	423	561	559	559
Mean			225	268	274	235	228	272	235	229	307	250	240	330	285	276

Table 3.15: Empirical RMSE Variance Estimates (x10^3) - MNAR mechanism

$\bar{p}$	ρ	$Corr_a(X,Y)$	Inflated Sample Size		Matching Substitution Unadjusted			Random Substitution Unadjusted			Matching Substitution Nonresponse Weighted			Random Substitution Nonresponse Weighted		
			ISS	ISS.W	MSub1	MSub3	MSub7	RSub1	RSub3	RSub7	MSub1.W	MSub3.W	MSub7.W	RSub1.W	RSub3.W	RSub7.W
0.50	0.01	0.01	59	64	92	65	61	92	67	62	104	71	65	102	72	67
		0.30	60	66	92	66	61	92	66	60	102	72	65	102	71	64
		0.70	59	68	92	67	63	90	65	60	103	73	67	105	74	68
	0.20	0.01	189	206	309	227	211	302	215	198	336	244	224	331	230	212
		0.30	212	253	335	240	217	315	230	216	400	263	228	368	269	256
		0.70	177	334	279	201	185	279	207	193	440	275	242	543	410	387
	0.50	0.01	450	497	652	470	432	648	480	448	719	502	457	720	516	476
		0.30	437	523	604	429	396	628	450	428	715	495	452	739	546	518
		0.70	376	852	569	423	402	574	425	396	1,106	648	551	1,275	975	890
0.75	0.01	0.01	50	51	67	61	61	67	62	62	69	63	63	69	64	63
		0.30	49	51	66	60	60	66	61	60	68	62	61	68	62	62
		0.70	47	50	68	62	62	68	62	62	71	64	63	72	66	65
	0.20	0.01	187	193	235	217	216	218	199	198	243	223	222	225	206	204
		0.30	174	192	252	228	226	234	223	221	265	232	229	256	240	238
		0.70	166	262	201	187	186	198	183	183	243	224	222	269	252	252
	0.50	0.01	395	399	490	460	457	488	455	453	507	472	469	505	467	465
		0.30	362	392	432	394	393	448	415	412	466	419	416	483	446	442
		0.70	326	509	452	413	410	434	407	406	546	479	480	700	664	664
0.90	0.01	0.01	71	72	62	61	61	62	61	61	63	62	62	62	61	61
		0.30	70	72	61	60	60	61	60	60	61	60	60	61	60	60
		0.70	71	72	63	62	62	62	61	61	63	62	62	63	62	62
	0.20	0.01	248	250	212	211	211	205	202	202	213	212	212	206	204	204
		0.30	254	263	236	233	233	223	221	221	239	234	234	230	227	227
		0.70	213	247	187	185	185	182	180	180	200	196	196	204	202	202
	0.50	0.01	480	482	487	481	481	477	474	473	488	481	481	476	473	472
		0.30	453	463	408	403	402	413	407	407	413	403	403	423	417	417
		0.70	455	628	428	421	421	422	419	419	460	438	437	633	630	630
Mean			226	278	275	237	230	272	235	230	322	260	249	344	295	286

-nisms, respectively, the RMSE of the sampling variance estimates is an increasing function of the intra-cluster correlation of the survey variable, but there is no relationship with the correlation between the survey and auxiliary variables, response rate or the missing mechanism. The use of nonresponse weighting tends to create sampling variance estimates that are less accurate than the corresponding estimates of the unadjusted respondent mean. Overall, substituting nonrespondents lead to more accurate sampling variance estimates than using a strata collapsing technique. Additionally, increasing the number of substitution iterations also improves the accuracy of the sampling variance estimates, as expected. Finally, there are no differences between random and matching substitution in terms of the RMSE of their sampling variance estimates. However, the use of nonresponse weighting on top of the substitution tends to create slightly less accurate sampling variance estimates.

3.5 Discussion

The results of the simulations presented here show some interesting patterns that can provide some useful guidelines to survey practitioners about the use of substitution procedures for nonresponding PSUs. It is important to emphasize, however, that the results of this limited set of simulations should not be readily generalized to broader settings that were not covered here. Such limitations are further discussed below. The conclusions presented here are limited to the conditions stipulated by the simulation design.

No substitution procedure outperforms the nonresponse weighting- adjustment method with an inflated sample size, both in terms of bias reduction or gains in precision. On the other hand, matching substitution seems to perform just as well as a nonresponse-weighting adjustment, when both are compared under the same conditions, which includes having the same set of variables available before and after the data collection. Therefore, while there is no evidence to discredit the use of the kind of substitution examined here when a matching procedure is implemented, there is no reason to prefer this method over a post-data collection adjustment, in terms of nonresponse bias reductions, if the same set of covariates is used for matching or nonresponse adjustment. Moreover, using an inflated sample size approach instead of substitution heavily depends on having a good estimate of the response rate – in this case, the PSU response rate. If this estimate is severely under or over-estimated, the desired number of units will not be obtained.

The same conclusion does not apply to random substitution. In fact, the unadjusted estimates produced by this procedure have properties that resemble the unadjusted respondent mean with no substitution, which, as shown before, can be subject to substantial bias. However, random substitution can be greatly improved with a nonresponse weighting adjustment. Even if random substitution is fully successful, that is, a substitute is found for every nonrespondent, it is important to employ additional nonresponse adjustments to reduce the bias of the survey estimates.

The standard nonresponse weighting adjustment had the opposite effect on the matching substitution estimates, at least when the variables used for the nonresponse weighting adjustment are the same as those used to match the nonrespondents to their substitutes. Both bias and sampling variance tend to increase when a weighting-adjustment is added to matching substitution. When relying on matched substitutions, even when some nonresponse remains in the sample, applying a post-data collection nonresponse adjustment using the same variables used for the matching procedure is not recommended. If a nonresponse adjustment is applied, different auxiliary variables or different methods, such as calibration, should be used to compensate for nonresponse.

If a nonresponse weighting adjustment with the same set of covariates used in the matching procedure is employed, it is important to consider which units to include in the nonresponse model estimation procedures and how to treat them. In this instance, simulation results (not shown here) suggest that including only the originally selected units (respondents and nonrespondents) and the final respondent substitutes, leaving aside substitute units that turned out to be nonrespondents, can bring bias reductions and gains in precision. An alternative approach is to use only the originally selected units on the nonresponse adjustment model estimation treating the response status of the nonresponding units that were substituted as of their corresponding substitutes and then apply the weights of the nonrespondents to the responding substitutes. Although such procedure was not tested in the simulations presented on this study, it is expected to perform similarly to the method described above.

Another important component in the comparison of the methods evaluated in this study is sampling variance estimation. In designs with deep stratification, in which there are as many strata as possible and with the minimum number of units per stratum to still allow for design-based sampling variance estimation, some of the strata might end up with zero or one responding unit(s) after nonresponse. There different techniques to deal with this problem. Among those, collapsing strata is most commonly used in practice, where strata with one or no responding units are collapsed with other strata so to form new strata with a minimum number of units for standard design-based variance estimation. A disadvantage of such technique is that it tends to overestimate the sampling variance.

Vehovar (1999) pointed out that one of the strengths of using substitution of nonrespondents is its potential to preserve the original sample design after nonresponse occurs. Substitution can fill in the gaps due to nonresponse making it no longer necessary to collapse strata or use other model-based techniques for sampling variance estimation. Moreover, Vehovar (1999) emphasized the need for a comparison between strata collapsing and substitution for sampling variance estimation. The simulation results showed that substitution leads to less biased and more accurate estimates for the sampling variance than strata collapsing.

The general results presented in this study show that substitution can be a valid alternative for the nonresponse problem. Its performance is similar to other standard nonresponse approaches, especially when a matching procedure is used or when further adjustments are made. Substitution can be a particularly valuable tool in samples with few units per stratum or cluster, in which nonresponse might disrupt the design to a point that standard design-based sampling variance estimation is not possible. In such situations substitution actually proved to be slightly more superior to strata collapsing in samples that do not use substitution, for instance.

As mentioned before, these conclusions are limited to the assumptions and conditions set in this simulation study. Some of these limitations are unlikely to change the conclusions in more general cases, while others may differ if the conditions are different. Next, these limitations are described in details.

First, the sample design used in these simulations was very specific, focusing on few clusters per stratum selected with probability proportional to size. In fact, the minimum number of units per stratum were selected while still enabling design-based sampling variance estimation. While this deep stratification is commonly used in many surveys, many others select more than a few units per stratum. This condition was set in this simulation to test the performance of the substitution method compared to strata collapsing when some strata have one or no responding units. When a larger number of units are selected per stratum, the likelihood of this situation and, therefore, the differences between these two methods are expected to diminish, although this must be confirmed in future studies. Moreover, with a large sampling fraction it may be more difficult to find good matched substitutes, since there could be fewer units with similar characteristics from the nonrespondents to be selected from the unasampled population.

Also, the simulations used nearly equal probability sample design and, despite base weighting, it was an approximately self-weighting design. In practice, most surveys used an unequal probability sample design, to oversample certain sub-populations and, thus, has a larger departure from a self-weighting design, relying more on the design-weights. The conclusions of this study are unlikely to change in these types of designs, but more investigations are needed to confirm this hypothesis. Finally, the stratification effects and correlation between the survey variable and the cluster sizes were not varied in this study. Some of the results in this study could differ under other size stratification effects or associations of the cluster sizes with the survey variables, particularly given some of the limitations associated with the nonresponse mechanism described below.

Second, the nonresponse mechanism used in this simulation was limited in terms of the variables and the degree of association with the survey and auxiliary variable. For simplicity, the two variables used in the design, the stratification indicators and the cluster sizes, were not included in the model that generated the missing mechanism. If these variables were included, some of the results could potentially change, particularly the estimation of the response propensity in which the stratification indicators were not included. Moreover, the matching substitution method may have different results based on whether the cluster sizes are included or not among the matching covariates.

Third, and also related to the nonresponse mechanism, the results of these simulations are based on an assumption that the mechanism for the original selected units is the same as for substitute units. This assumption might be reasonable when substitution is conducted in a more rigorous environment, with substitution selections carefully made by the office rather than the interviewers. It probably also would require a strict survey protocol that emphasizes obtaining as high a level as possible of cooperation of the originally selected units. When such conditions do not apply, the nonresponse mechanism of the substitutes may differ from the original sample and, depending on how different these mechanisms are from those that apply to originally selected units, nonresponse bias could actually increase rather than decrease. Therefore, the results obtained here should not be readily extended to all situations in which substitution is used, and more research is needed to address situations in which the nonresponse mechanism may differ between original sampled units and substitutes.

Lastly, this simulation only examined the situation in which there is only one covariate available for matching the substitutes to the nonrespondents. In most applications, multiple covariates are usually available. An evaluation of different methods for the matching procedure should be explored, such as using a multidimensional distance measure or propensity score matching. The results may differ depending on the method used and on the association of the matching covariates with the survey variables. Moreover, only a normally distributed variable was considered in these simulations. Future studies should investigate how these methods perform with both continuous variables with other stochastic distributions and categorical variables.

Despite extensive use in practice, substitution of nonrespondents has been a neglected and understudied problem in survey sampling. Further research on this topic may yield bias and sampling variance reductions that merit wider use of substitution to compensate for survey nonresponse. For example, a matching procedure might leave differences between nonrespondents and their substitutes on observed variables that could not be used in the matching. Rubin and Zanutto (2002) proposed an imputation method that adjusts for such differences, although their technique might produce increased survey cost and sampling variability in settings like simulated in this study. Alternative methods to calibrate respondent and substitute differences suitable for the two-stage sampling examined here may keep survey costs or sampling variability at levels equivalent to other nonresponse adjustment methods.

Survey cost was largely ignored in this study, for the sake of simplicity. But it is an important factor to be considered in future research. While substitution attempts to fill in the gaps left in the sample due to nonresponse, this procedure increases costs as additional units are selected and targeted to substitute the nonrespondents. Original sample size must be smaller to have resources to select and recruit substitute units. Inflating the sample size, on the other hand, might suffer from an incorrect estimate of the response rate, particularly if it is under-estimated, in which case more units may be selected than necessary leading to larger costs or lower response rates. A more complete portrait of the nonresponse problem should consider trade-offs between survey errors and survey costs in the use of substitution and inflated sample size methods.

Nonresponse also occurs at other stages of sample selection in multistage sampling. The use of substitution for other sampling units needs research attention. This study looked specifically at the problem of nonresponding PSUs, a fairly common problem establishment and school surveys, in particular. The approach of this study could be extended to nonresponse among second stage units and other similar situations readily.

Finally, the approaches examined here ignored the contributions that might be made using imputation methods as an alternative or a complement to substitution. While weighting might be a preferred procedure for many survey practitioners for adjusting unit nonresponse, imputation has proven valuable at an element-level, and could be extended to dealing with PSU level nonresponse as well.

References

Bachman, J. G., Johnston, L. D., O'Malley P. M. and Schulenberg, J. E. (2011). *Monitoring the Future Project After Thirty-Seven Years: Design and Procedures.* Ann Arbor, MI. Institute for Social Research, University of Michigan.

Baldissera, S., Ferrante, G., Quarchioni, E., Minardi, V., Possenti, V., Carrozzi, G., Masocco, M., Salmaso, S. (2014). Field substitution of nonresponders can maintain sample size and structure without altering survey estimatesdthe experience of the Italian behavioral risk factors surveillance system (PASSI). *Annals of Epidemiology,* 24, pp. 241-245.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3), 251-260.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons, Inc., Hoboken, New Jersey

Biemer, P., Chapman, D. W., and Alexander, C. (1985). Some Research Issues in Random-Digit Dialing Sampling and Estimation. *Proceedings First Annual Research Conference*, March 20-23, 1985.Washington DC: Bureau of the Census, 1985.

Chapman, D. W. (1983). The Impact of Substitutions on Survey Estimates. *Incomplete Data in Sample Surveys*, Vol. II, Theory and Bibliographies, eds. W. Madow, I. Olkin, and D. Rubin, New York: National Academy of Sciences, Academic Press, pp. 45-61.

Chapman, D. W. (2003). To Substitute or Not to Substitute – That is the question. *The Survey Statistician*. No. 48, pp. 32-34.

Chapman, D. W. and Roman, A. M. (1985). An investigation of substitution for an RDD survey. *Proceedings of the Survey Research Methodology Section*, ASA, pp. 269-274.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons.

Cohen, R. (1955). *An investigation of modified probability sampling procedures in interview surveys*. M.A. thesis submitted for the graduate faculty of The American University, May 26, 1955.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly*, 69, pp. 87-98.

David, M. C., Bensink, M., Higashi, H., Donald, M., Alati, R., and Ware, R. S. (2012). Monte Carlo simulation of the cost-effectiveness of sample size maintenance programs revealed the need to consider substitution sampling. *Journal of Clinical Epidemiology*, Vol. 65, Issue 11, pp. 1200-1211.

David, M. C., Ware, R. S., Alati, R., Dower, J. and Donald, M. (2014). Assessing bias in a prospective study of diabetes that implemented substitution sampling as a recruitment strategy. *Journal of Clinical Epidemiology*, Vol 67, Issue 6, pp. 715-721.

De Leeuw, E. and De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D Dillman, J. Eltinge, and R. Little (eds.) *Survey Nonresponse*, pp. 41-54. New York: Wiley.

Durbin, J., and Stuart, A. (1954). Callbacks and clustering in sample surveys: An experimental study. *Journal of the Royal Statistical Society*. Series A, Part IV, pp. 387-428.

Frankel, M. and King, B. (1996). A conversation with Leslie Kish. *Statistical Science*, Vol. 11, No. 1, pp. 65-87

Groves, R. M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72 (2), pp. 167-189.

Groves, R. M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

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

Keeter, S., Miller, C., Kohut, A., Groves, R. M. and Presser, S. (2000). Consequences of Reducing Nonresponse in a Large National Telephone Survey. *Public Opinion Quarterly*, 64, pp. 125-48

Keeter, S., Kennedy, C., Dimock, M., Best, J. and Craighill, P. (2006). Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey. *Public Opinion Quarterly*, 70, pp. 759-779

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.

Little, R. J., and Vartivarian, S. L. (2003). On Weighting the rates in non-response weights. *Statistics in Medicine*. 22, pp. 1589-1599.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Lynn, P. (2004). The Use of Substitution in Surveys. *The Survey Statistician*. No. 49, pp. 14-16.

Merkle, D. M. and Edelman, M. (2002). Nonresponse in Exit Polls: A Comprehensive Analysis.

In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, pp. 243-58. New York: Wiley.

Moser, C.A., and Kalton, G., (1972) *Survey Methods in Social Investigation*. New York: Basic Books,.

Nathan, G. (1980). Substitution for Non-response as a Means to Control Sample Size. *Sankhyaa*, C42, 1-2, pp. 50-55.

PISA (2012). Technical Report. OECD. http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf (accessed on May 28th 2015)

Rand, M. (2006). Telescoping Effects and Survey Nonresponse in the National Crime Victimization Survey. Paper presented at the Joint UNECE-UNODC Meeting on Crime Statistics. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.14/2006/wp.4.e.pdf (accessed on March 21st 2014)

Rosenbaum, P. R. (1995). *Observational Studies*. New York: Springer-Verlag.

Rubin, D. B., and Zanutto, E. (2002). Using Matched Substitute to Adjust for Nonignorable Nonresponse through Multiple Imputation. In *Survey Nonresponse*, edited by R. Groves, R. J. A. Little, and J. Eltinge. New York: John Wiley, pp. 389-402.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Skinner, C. J., and D'Árrigo, J. D. (2011). Inverse probability weighting for clustered nonresponse. *Biometrika*, 98, 4, pp. 953-966.

Sirken, M. (1975). Evaluation and critique of household sample surveys of substance use. In *Alcohol and other drug use in the State of Michigan*. Final report, prepared by the Office of Substance Abuse Service, Michigan Department of Public Health.

Thompson, M. and Wu, C. (2008). Simulation-based randomized systematic pps sampling under substitution of units. *Survey Methodology*, 34, pp. 3-11.

Vehovar, V. (1999). Field Substitution and Unit Nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, pp. 335-350

Vives, A., Ferreccio, C. and Marshall, G. (2009). A comparison of two methods to adjust for nonresponse bias: field substitution and weighting non-response adjustments based on response propensity (In Spanish with a summary in English). *Gaceta Sanitaria*, 23 (4), pp. 266-271.

Williams, S. R., and Folsom, R. E. Jr. (1977). *Bias resulting from school nonresponse:*

Metodology and findings. Prepared by the Research Triangle Institute for the National Center of Educational Statistics.

Wolter, K. M. (2007). *Introduction to Variance Estimation, Second Edition*. Springer-Verlag.

Yuan, Y., and Little, R. J. A. (2007). Model-based estimates of the finite population mean for two-stage cluster sample with unit non-response. *Applied Statistics*, 56, Part 1, pp. 79-97.

Zanutto, E. (1998). Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes. Doctorate thesis submitted for the graduate faculty of Harvard University, May, 1998.

CHAPTER IV

Imputation and Calibration Adjustment Methods to Improve Substitution

Summary

Substitution, in which a nonresponding unit in a survey is replaced by another unit not originally selected in the sample, is a widely used strategy to deal with nonresponse in many surveys in practice. However, little research has been conducted about this often criticized or neglected procedure in the survey statistics literature. Rubin and Zanutto (2002) proposed the only method to date that attempts to improve this methodology by reducing nonresponse bias caused by differences between nonrespondents and their corresponding substitutes. However, their method requires the selection of substitutes for a sub-sample of the respondents for estimation purposes, which leads to additional costs to the survey operation. This paper presents two new methods to enhance substitution for nonrespondents. First, a modification is suggested that eliminates the selection of the additional sample of substitutes from non-sampled units by selecting substitutes from among responding units, therefore making the method more cost-effective. Second, differences between nonrespondents and substitutes are adjusted using a calibration procedure. This latter methodology eliminates the need to collect additional substitutes from non-sampled units for some of the respondents, and increases the precision of the estimates through calibration. These methods are evaluated and compared through two simulation studies under a variety of settings.

4.1 Introduction

Substitution is a survey procedure for compensating for nonresponse among sample units by selecting a replacement unit from the population for each nonresponding unit. Nonresponse bias reduction is a main driver for using substitution to replace nonresponding units. Such reductions will only be achieved if nonrespondents and substitutes are similar on survey variables.

However, substitutes are respondents and may differ systematically from the nonresponding units they replace. Such differences might be related to one or more of three different types of variables.

First, there are variables that are observed for both nonrespondents and their substitutes. These variables could be used in survey estimation in statistical models to adjust estimates to account for nonrespondent-substitute differences. For example, in household surveys there may be demographic variables available for households or persons that could be used to adjust the values of other variables to account for nonrespondent-substitute differences. In an establishment survey, the size of an organization might be used to increase or decrease the contribution of a substitute if the substitute size is smaller or larger than the unit it replaces. The differences between nonrespondents and their substitutes with respect to this type of variables may be adjusted through a variety of statistical methods (Little and Rubin, 2002) that assume nonresponse is missing at random (MAR).

A second type of variable is the set of variables that are directly or indirectly observed for all units in the population, but their inclusion in a statistical model would be difficult or impractical. These are what might be called higher dimensional attributes, such as geographic location (an address or a zip code) which is difficult to use in models because of its high dimensionality. Alternatively, there may be a large number of categorical variables for which all or most of their interactions are needed to explain the outcome variables. Typically, this higher dimensional relationship between nonrespondent and substitute could be taken into account through matching (Rubin, 1973). That is, the selection of a substitute for a nonrespondent could be based on a measure of distance between that nonrespondent and the unsampled units on these variables.

Finally, differences between nonrespondents and their substitutes might be explained by unobserved variables or even, in the worst-case scenario, by the survey variables themselves. In this case the nonresponse is nonignorable and statistical adjustments must rely on untestable assumptions. Although it is an important problem in survey statistics, this non-ignorable missing data situation is outside the scope of this paper and will not be further considered here.

Rubin and Zanutto (2002) propose a method they call “matching, modeling, and multiple imputation” (MMM) to adjust for differences between nonrespondents and their substitutes. MMM assumes MAR, and thus uses the first two types of variables described above. Rubin and Zanutto show their proposed method reduces nonresponse bias, even though it requires substitution for all nonrespondents and for a sub-sample of the *respondents*. This additional sample selection imposes additional costs that many practitioners are not willing or able to incur. Further, the substitutes for respondents are discarded from the dataset after having been used in the adjustment process (a multiple imputation procedure). Another potential disadvantage of the MMM method is that it introduces added variability to the estimates through the imputation process. Rubin and Zanutto (2002) demonstrate that under a variety of circumstances, the estimates obtained from the MMM have much larger variability than existing alternatives, such as weighting or other standard forms of substitution. Survey designers have been reluctant to use such a method without clear evidence that the added costs leads to substantial reduction in bias or increases in precision[3].

In this paper, modifications to the MMM method that eliminate the need for selecting and collecting data for a sample of substitutes for respondents but still successfully reduces nonresponse bias are proposed. One selects substitutes from among existing sample respondents, and uses those in a method similar to the MMM method. Another uses calibration to adjust for differences between substitutes and nonrespondents on auxiliary variables not used in the matching substitution. This latter method could lead to estimates with a smaller sampling variances compared to those obtained under the MMM method. The performance of these methods is evaluated through the use of simulations.

In the next section, the MMM method is presented before the alternative methods (substitution by sample respondents and calibration) are examined. The paper concludes with a description of the design and the results from two simulation studies and summary remarks.

[3] Chiu et al. (2005) illustrates an application of the MMM method in a different context, in which the “substitutes” were data aggregates from geographical census units, such as blocks or census tracts.

4.2 Matching, Modeling, and Multiple Imputation

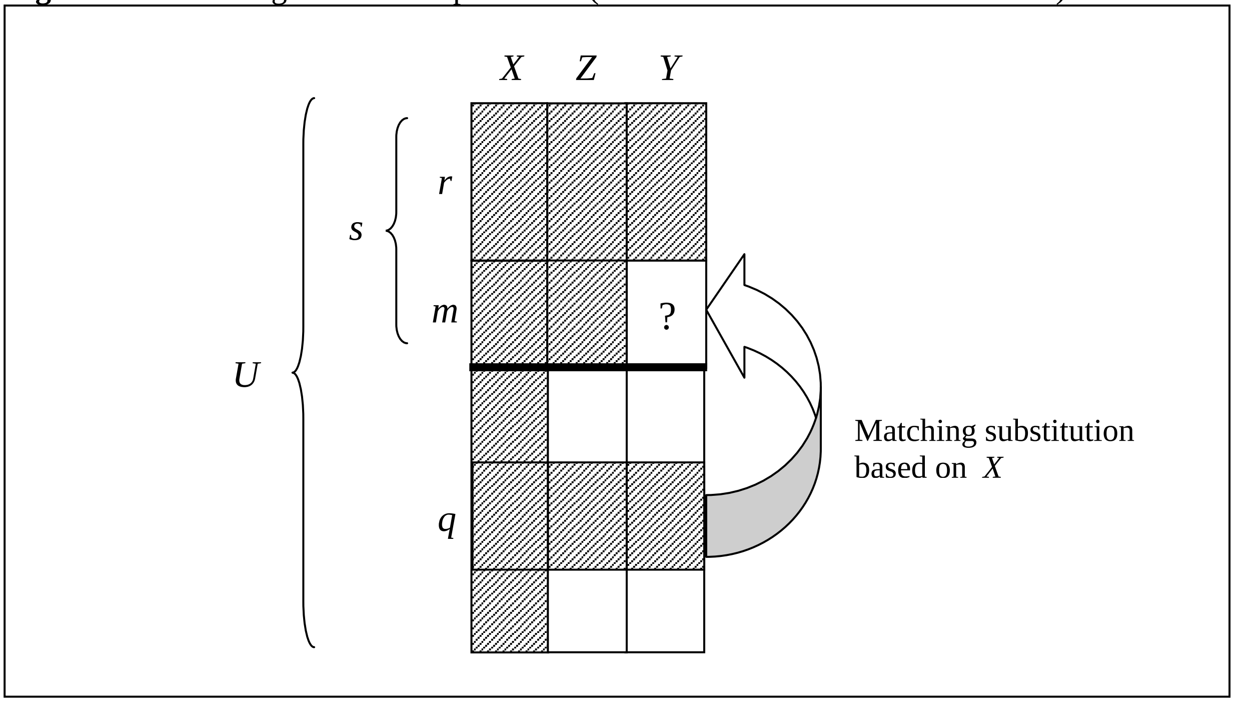
Rubin and Zanutto (2002) distinguish between two types of variables that can potentially explain differences between nonrespondents and corresponding substitutes, matching and modeling covariates.

Matching covariates, denoted by X, are available for every unit in the population, typically variables available from the sampling frame materials. These covariates are used to match nonrespondents to unsampled units to serve as their substitutes. Such variables would typically not be used in models for nonresponse adjustments, because their use would require too many parameters or an arbitrary categorization to a smaller number of classes. Address or geographic location might be considered as the basis for matching nonresponding and substitute cases where, for example, a substitute is selected for a nonresponding unit from the same block, or school, or other unit from the same geographic location. Geographic location though is difficult to be used in a statistical model, such as including a sequential identification as a predictor. Because nonrespondents and substitutes are matched on these variables, they will potentially share the same values of other variables that cannot be directly observed, and therefore, are not available for analysis. Chiu et al. (2005) call these variables "contextual variables".

Modeling covariates, denoted by Z, are variables typically available only for nonrespondents and their substitutes collected during data collection, such as paradata (Couper, 1998) in a cross-sectional survey or, in longitudinal surveys, data from previous waves for current-wave nonrespondents. Because nonrespondents and their corresponding substitutes usually cannot be matched by these variables, there might be some differences between them with respect to these covariates. For this reason, Rubin and Zanutto suggest modeling these differences and using the results of these models for multiple imputation of the nonrespondents using the substitutes' data.

The interest is in estimating a population parameter associated with a survey variable Y, such as a mean, median, or an association with other variables. For this purpose, a probability sample *s* of size *n* from a finite population $U = \{1,...,i,...N\}$ is selected. The set of respondents and nonrespondents are denoted by r and m, respectively (see Figure 4.1). In a matching substitution procedure, substitutes for each nonrespondent in m are selected according to a matching variable X by finding the non-sampled case with the closest proximity to the nonrespondent. Denote the set of matched substitutes by q, where each one of its units has a one-to-one correspondence to a unit in the nonrespondent set m. For simplicity, it is assumed throughout this study that the matching substitution procedure is fully successful, that is, a responding substitute is successfully obtained for every nonrespondent. This assumption can be relaxed for some of the methods presented below.

Figure 4.1: Matching substitution procedure (shaded area indicates available data)



To adjust for potential differences between nonrespondents and their corresponding substitutes Rubin and Zanutto proposed the following model for the survey variable Y:

$$y_i = \alpha + \beta_1 z_i + \beta_2 x_i + \varepsilon_i \qquad (1)$$

with $\varepsilon_i \sim N\left(0, \sigma^2\right), i \in s$. Further, they argue that since substitutes are also part of the same population of the originally selected units, their survey variable should also follow the same distribution:

$$y_i^s = \alpha + \beta_1 z_i^s + \beta_2 x_i^s + \varepsilon_i^s \qquad (2)$$

where y_i^s, z_i^s, and x_i^s denote, respectively, the y_i, z_i, and x_i values for the substitute of the i^{th} nonrespondent and $\varepsilon_i^s \sim N\left(0, \sigma^2\right), i \in m$. Since substitutes are matched to the nonrespondents on the matching variables, that is, $x_i^s = x_i$, the difference between them in terms of the survey variable is

$$y_i - y_i^s = \beta_1\left(z_i - z_i^s\right) + \varepsilon_i^{'}, \quad i \in m \qquad (3)$$

However, this model cannot be fit because the survey variable is unobserved for nonrespondents. Rubin and Zanutto suggested selecting substitutes from the non-sampled units for a sub-sample of the *respondents*, say r^s, to then fit the following model:

$$y_i - y_i^s = \beta_0 + \beta_1\left(z_i - z_i^s\right) + \varepsilon_i^{'}, \quad i \in r^s \qquad (4)$$

where the intercept β_0 is included to minimize possible misspecification bias. An important assumption is that the same relationship of the difference in the survey variables between nonrespondents and their substitutes in *m* also holds for respondents and their substitutes in r^s. To weaken this assumption, the respondents selected for this sub-sample should be similar to the nonrespondents in terms of the modeling covariates, Z, and, if possible the matching covariates, X.

Rubin and Zanutto then propose multiply imputing the nonrespondent's survey variable values based on draws from

$$y_i \sim N\left(y_i^s + \beta_0 + \beta_1\left(z_i - z_i^s\right), \sigma^2\right), i \in m \qquad (5)$$

using flat prior distributions on the model parameters $\left(\beta_0, \beta_1, \sigma^2\right)$. After imputation, the substitute's data from nonrespondents in m and the sub-sample of respondents in r^s are discarded. An estimate for the population mean of the survey variable and its standard error are computed using Rubin's combining rule (Rubin, 1987) across the multiply imputed data sets. For the detailed algorithm on how to implement this method, see Zanutto (1998, pages 131-132).

4.3 A Modification to the MMM Method

A clear disadvantage of MMM is that it requires the selection of substitutes for a sub-sample of the respondents which is then discarded after model estimation and multiple imputation. Zanutto (1998) tests the performance of the MMM method using different sub-sample sizes of respondents $n^* = kn_m$, where n_m is the number of nonrespondents in the sample. She evaluates the performance of MMM in a simulation study using $k = 0.1, 0.3, 0.5, 0.8$ and $n^* = n_r$, where the n_r is number of respondents. She finds that for every sub-sample size n^*, the amount of bias reduction on the estimates of the population mean of the survey variable, Y, is roughly the same. But as n^* increases, the sampling variance of the estimate decreases. Zanutto concludes that future research should investigate the trade-offs between bias and precision of the survey estimates and the cost associated with selecting substitutes for respondents, with possible guidelines for the choice of sub-sample size.

A modification to the MMM method would eliminate additional costs associated with selecting substitutes for a sub-sample of the respondents. Because substitutes of the sub-sample of the respondents are also respondents, instead of drawing these substitutes from the non-sampled population, the modified procedure proposes to select them from the pool of the remaining respondents on the sample, a procedure similar to hot-deck imputation or flexible matching procedures (Kalton and Kasprzyk, 1986). The difference between a hot-deck imputation and such a substitution procedure is that the former uses the donor values to replace the missing data of the nonrespondents, where the latter only uses the substitutes for the sub-set of respondents to allow for the estimation of model (4).

This modified method first selects substitutes matched on X for the nonrespondents from the non-sampled population. A sub-sample of respondents who are similar to the nonrespondents in terms of the modeling covariates Z are selected, and matched substitutes for these cases found among *the remaining respondents*. Then the following model would be estimated:

$$y_i - y_i^s = \beta_0 + \beta_1 \left(z_i - z_i^s \right) + \varepsilon_i^{'}, \quad i \in r^s \tag{6}$$

The subsequent imputation of missing values for nonrespondents follows Rubin and Zanutto (2002).

This modification would produce estimates of the population of the survey variable with a larger sampling variance compared to Rubin and Zanutto's MMM, since the substitutes for the sub-sample of respondents in this modified approach are drawn from existing pool of respondents. Thus, no additional information is being added to the sample. Also, depending on the sample size, the nature of the matching covariates, and the response rate, it may not be possible to find different substitutes for every unit in the sub-sample of respondents. For example, if the matching covariate is a cluster indicator, the cluster sample size is small and the response rate is low, the size of the sub-sample of respondents to be substituted might be larger than the number of remaining respondents in the cluster. In such situations, selecting these substitutes *with replacement* from the pool of remaining respondents can be used without consequences to the nonresponse bias reduction. The sampling variance, however, might be higher than it would be if the selection was made without replacement.

An important advantage of this modification over Rubin and Zanutto's original method is that under MAR the same bias reduction can be achieved at a lower cost, since it eliminates the need to collect extra data from substitutes for the respondent sub-sample out of the pool of non-sampled units.

4.4 A New Approach Using Calibration

Rubin and Zanutto (2002) propose modeling to adjust for differences between the nonrespondents and their substitutes on the modeling variables and then multiply imputing the nonrespondent's survey variables. As mentioned before, this method has two disadvantages: (1) it requires the selection of substitutes for a sub-sample of respondents (which the modification proposed above is designed to overcome), and (2) it increases the sampling variability through the imputation procedure. Further, after the imputation of the nonrespondents' data, the substitutes' data are discarded and not used in the estimation of the population mean of the survey variable. On one hand, this might be justifiable on the basis that, if included in the inference, such substitutes would modify the probability sample design by adding extra cases in the unobserved data "blocks" (Rubin and Zanutto, 2002). On the other, it seems like a waste of data to discard these substitutes. A new approach to adjust for differences between nonrespondents and their matched substitutes with a modeling variable that attempts to overcome these problems and improve the use of substitution in probability samples is proposed here.

As before, let s be a probability sample from a finite population $U = \{1,...,i,...N\}$, in which units are selected with known inclusion probabilities $\pi_i = P(i \in s)$, so that design weights can be computed as $d_i = \pi_i^{-1}$, with $D = (d_1, d_2, ..., d_n)'$. The set of respondents and nonrespondents are denoted by r and m, respectively. The nonrespondents in m are substituted according to a matching variable X. The set of matched substitutes is denoted by q and each one of its units has a one-to-one correspondence to a unit in the nonrespondents set m.

In imputation, the design-weights d_i of the subject are attributed to the imputed data. Since substitution can be considered as a form of imputation, where missing data for the nonresponding unit is replaced by data from the substitute, the design-weights of the nonrespondents can also be assigned to their corresponding substitutes. Alternatively, these design-weights can be computed as if the substitutes were the originally selected units. Simulation results (not shown here) suggest that in terms of bias both approaches lead to similar performances. Throughout this study, the former alternative for computing design-weights of the substitutes is used.

To adjust the matched substitutes according to a variable Z, observed for both nonrespondents and their substitutes, a calibration approach (Deville and Särndal, 1992) is proposed. The objective is to find for the substitutes a new set of weights $W = \left(w_1, w_2, ..., w_n\right)'$ that minimizes a distance measure $G_i\left(W, D\right)$ under the following restriction:

$$\sum_{i \in q} w_i z_i = \sum_{i \in m} d_i z_i \, , \tag{7}$$

such that the calibrated-weighted total for the variable Z over the substitutes will be the same as the design-weighted total over the nonresponding units. While Rubin and Zanutto call Z a modeling covariate, it is denoted here as a calibration covariate.

Once the calibrated weights for the substitutes are found, the combined set of responding and matched substitute units $s^* = \left(r, q\right)$ is used to estimate the finite population mean for a variable Y as

$$\overline{y}_{Cal.MSub} = \frac{\sum_{i \in s^*} w_i^* y_i}{\sum_{i \in s^*} w_i^*} \tag{8}$$

where $w_i^* = d_i$ for $i \in r$ and $w_i^* = w_i$ for $i \in q$.

The calibration restriction given above can be further extended to:

$$\sum_{\substack{i \in r \\ i \in q}} w_i z_i = \sum_{i \in s} d_i z_i \, , \tag{9}$$

that is, the total for the calibration covariate Z over the set of all respondents, including both the originally selected units and the nonrespondents' substitutes, is calibrated to the design-weighted total over all units (respondents and nonrespondents) selected in the original sample s. This restriction is more general in the sense that it can also be used when the substitution is not fully successful (when it is not possible to find a responding matching substitute for every nonrespondent). In this case, only the responding substitutes in q would be used in the left-hand side of the calibration restriction above.

Unlike Rubin and Zanutto's MMM method, this calibration approach does not require the selection of substitutes for respondents, either from the unsampled population or from the pool of respondents in the sample, thus avoiding any additional operational costs. Further, it does not discard the substitute data prior to the estimation of the population parameters. Instead, it uses them, along with a calibration-weighting adjustment, to account for possible differences in the calibration variable Z between nonrespondents and their substitutes.

4.5 Simulation Studies

4.5.1 Simulation Design

A series of simulations were conducted to evaluate and compare the performance of the methods discussed previously to other commonly used nonresponse adjustment procedures. These simulations were performed under two different contexts:

- Simulation Study 1: Populations containing variables with hidden clustering effects induced by a matching covariate, and
- Simulation Study 2: Explicitly clustered populations.

In both studies, the objective was to estimate the finite population mean $\bar{Y}$ in a set of $K = 5,000$ repeated samples. All simulations and analysis were conducted in `R` (R Core Team, 2014) and the calibration adjustments were performed using the `survey` package (Lumley, 2012; Lumley, 2004).

Simulation Study 1 followed a setting similar to the one designed by Rubin and Zanutto (2002). Simple random samples of size $n = 500$ were drawn from three different artificial finite populations of size $N = 10,000$, with each sample generated according to the survey variable and nonresponse mechanism models in Table 4.1. The matching covariate X was a dichotomous variable that induced a hidden clustering effect in the following way: $x_i = 0$, for $i = 1 + 100c, \ldots, 75 + 100c$, and then $x_i = 1$, for $i = 76 + 100c, \ldots, 100 + 100c$, with $c = 0, \ldots, 99$. That is, the population consisted of 100 sequences of 75 units with $X = 0$, followed by 25 units with $X = 1$. This covariate was indirectly used in the substitution process by matching nonrespondents to substitutes on their index number, as units with close index numbers likely had the same value on X.

Variable Z was the modeling/calibration covariate, while U was an unobserved covariate that cannot be used for matching or modeling/calibration (i.e., U generated a missing not at random nonresponse mechanism). In this study, both Z and U were independent $N(0,1)$ random variables. The probability of response is denoted by p and for the all populations the nonresponse mechanism model gives a response rate of approximately 70%.

Table 4.1: Populations for Simulation Study 1.

Population	Survey variable model	Nonresponse mechanism model
1	$y_i = z_i + 5x_i + \varepsilon_i$	$p_i = \frac{0.95}{1 + \exp(-1.25 + z_i + x_i)} + 0.05$
2	$y_i = z_i + \varepsilon_i$	$p_i = \frac{0.95}{1 + \exp(-0.95 + z_i)} + 0.05$
3	$y_i = z_i + 5x_i + u_i + \varepsilon_i$	$p_i = \frac{0.95}{1 + \exp(-1.35 + z_i + x_i + u_i)} + 0.05$

Each population corresponds to situations of different performance for the methods studied here. Population 1 is a situation in which the survey variable and the response propensity both depend on matching variable X and modeling/calibration covariate Z. Methods that compensate for both variables are expected to have less bias than methods that do not. For instance, the MMM method is expected to have smaller bias because it accounts for both matching and modeling/calibration values in estimation. On the other hand, a standard nonresponse propensity weighting adjustment that only uses Z as a predictor is expected to have larger bias.

Population 2 represents the scenario where both the survey variable and the response propensity depend solely on the modeling/calibration covariate Z. For this reason, methods that use only this variable in adjustments, such a nonresponse propensity weighted-mean, are expected to perform just as well as methods that also adjust for the matching covariate, like the MMM. However, a matching substitution without any further adjustment to account for Z is expected to produce biased estimates.

Population 3 corresponds to a non-ignorable nonresponse situation, in which the survey outcomes and the response propensity depend on an unobserved variable U. In this case, none of the methods studied here are well suited to adjustment and estimates are expected to have some degree of bias. However, because the matching covariate X and modeling/calibration covariate Z also explain the survey variable and the response propensity, methods that adjust for both of them will tend to lead to smaller biased estimates than methods that do not.

Simulation Study 2 was motivated by studies conducted by Yuan and Little (2006) and Skinner and D'Arrigo (2011), and involves a more complex structure where the clustering was explicit. Four artificial finite population of size $N = 40,000$, each consisting of $A = 400$ equal size clusters of $B = 100$ elements each, were generated using the models for the survey variable Y (at the cluster- and element-level) and the nonresponse mechanism given in Table 4.2.

Table 4.2: Populations for Simulation Study 2.

Population	Survey variable model (cluster level)	Survey variable model (element level)	Nonresponse mechanism model
4	$\nu_i = \alpha_i$	$y_{ij} = 5z_{ij} + \nu_i + \varepsilon_{ij}$	$p_{ij} = \frac{\exp(1+u_i)}{1+\exp(1+u_i)}$
5	$\nu_i = \alpha_i$	$y_{ij} = 5z_{ij} + \nu_i + \varepsilon_{ij}$	$p_{ij} = \frac{\exp(0.5z_{ij}+u_i)}{1+\exp(0.5z_{ij}+u_i)}$
6	$\nu_i = \alpha_i + 5u_i$	$y_{ij} = 5z_{ij} + \nu_i + \varepsilon_{ij}$	$p_{ij} = \frac{\exp(1+u_i)}{1+\exp(1+u_i)}$
7	$\nu_i = \alpha_i + 5u_i$	$y_{ij} = 5z_{ij} + \nu_i + \varepsilon_{ij}$	$p_{ij} = \frac{\exp(0.5z_{ij}+u_i)}{1+\exp(0.5z_{ij}+u_i)}$

The modeling/calibration covariate is $z_{ij} \overset{iid}{\sim} N\left(2,1\right)$, for clusters $i=1,..,A$ and elements within clusters $j=1,..,B$. Z was generated using the `R` package `truncnorm` (Trautmann et al, 2014), and truncated below by 0 and above by 4. The random variables α_i, u_i and ε_{ij} are independent $N\left(0,1\right)$, for $i=1,..,A$ and $j=1,..,B$.

Therefore, each population represents a different clustering structure. While Populations 4 and 5 have a lower intra-cluster correlation (as defined by Kish, 1965) of about 4%, Populations 6 and 7 present a stronger clustering effect, with a intra-cluster correlation of approximately 48%. Since the cluster indicator is used as the matching covariate X in this second simulation study, it is expected that methods that solely rely on matching substitution will not to perform as well in population 4 and 5 than they would perform in populations 6 and 7. Moreover, methods that do not use matching substitution should present some substantial bias in these latter two populations.

Each population in this simulation study also corresponded to a different missing mechanism. Population 4 is missing completely at random (MCAR). Since the response propensities in Population 5 depend only on the fully-observed covariate Z, the data are missing at random (MAR). Populations 6 and 7 correspond to different types of cluster-specific non-ignorable nonresponse (CSNI), another form of missing mechanism proposed by Yuan and Little (2006) for cluster sampling settings. As described by these authors, CSNI occurs when the response propensities of the elements in a cluster sample depend on the cluster means. Although the cluster membership is fully observed for nonrespondents, the missingness is not MAR, because the cluster means are in fact unobserved random effects in this type of setting.

The sample design of Simulation Study 2 was a two-stage cluster sample of size $n=1,200$ with simple random sampling at both stages of $a=60$ clusters and $b=20$ elements within sampled clusters. In this case, the nonresponse occurs at the element-level, and, as in the first simulation study, the overall response rate was approximately 70%. The matching covariate in this set of simulations is the cluster indicator. Therefore, in most cases there were multiple candidates of substitutes for the nonrespondents, so that within a cluster the substitutes were randomly selected.

In both simulation studies, seven nonresponse adjustment methods were compared. All the methods used a target sample size of $n = 1,200$ and ultimately, on average, this same sample size is used for the estimation of the population mean $\bar{Y}$. However, the MMM methods use data from an additional of 30% of the number of nonrespondents (an average of $n^* = 0.3(1-\bar{p})n =$ $= 0.3(1-0.7)1200 = 108$ in these simulations) to allow estimation of their multiple imputation models. The initial sample size could be adjusted to account for these additional units, but, because they are not ultimately used for the estimation of the population mean, the target sample size was kept the same to be comparable with the other adjustment methods. Although the first method described below is very rarely used in practice -- as it will be most likely biased under the presence of nonresponse -- it is included here as a baseline measure of the amount of nonresponse bias the other adjustment methods are able to reduce. Further, a sample mean assuming complete response is used to estimate the sampling variance under ideal conditions. Each of the seven methods is described below.

1. **Inflated sample size (ISS)**: This is the unadjusted respondent mean where the sample size is inflated by the expected response rate, $\bar{p}$. That is, a sample of size $n' = n/\bar{p}$ in Simulation Study 1 and $b' = b/\bar{p}$ in Simulation Study 2 is selected and the mean of the respondents is used as an estimate of the population mean. The known value of the response rate $\bar{p}$ is used in these simulations.

2. **ISS adjusted by nonresponse propensity Weight (ISS.W)**: Similar to the previous method, the sample size is inflated by the expected response rate, $\bar{p}$. The respondents are then weighted by the inverse of their predicted response propensities , $\hat{p}_i$, estimated using the modeling/calibration covariate, Z, as predictor of the response indicator in the following logistic regression model:

$$\operatorname{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 z_i,\ i \in s$$

3. **Matching Substitution (MSub)**: An initial sample of size n is selected. Each nonrespondent is substituted by a unit from the pool of unsampled units that is the closest to the nonrespondents in terms of the matching covariate (index number in the Simulation Study 1 and cluster indicator in Simulation Study 2). If the substituted unit turns out to be a nonrespondent as well, the next closest unsampled unit is selected as the substitute, repeating this process until a responding substitute is chosen. If there is more than one unit that can be used as a substitute for a given nonrespondent, the substitute is randomly selected from among these units. No further adjustments are done to take into account possible differences in the modeling/calibration covariates.

4. **Matching Substitution adjusted by nonresponse propensity Weight (MSub.W)**: Following MSub, with respondents (original selected and substitutes units) weighted by the inverse of their predicted response propensities, $\hat{p}_i$, estimated using the modeling/calibration covariate, Z, as predictor of the response indicator in the following logistic regression model.

   $$\operatorname{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 z_i,\ i \in s \bigcup q$$

   Notice that such model is estimated using the data from all the respondents and nonrespondents in the original sample s and substitute set q. Therefore, these predicted response propensities account for both the original and substitute nonresponse.

5. **Matching, Modeling and Multiple Imputation (MMM)**: The method proposed by Rubin and Zanutto (2002). Similar to MSub and MSub.W, an initial sample of size n is selected and a substitute for every nonrespondent is chosen by matching on the matching covariate (index number in the first simulation study and cluster indicator in the second study). Following Rubin and Zanutto's simulation study, substitutes are also selected from the pool of *unsampled* units for a sub-sample of $n^* = 0.3n_m$ respondents, where n_m

is the number of nonrespondents. Then, $M = 10$ multiple imputations sets for the missing data are created using the method described in section 2.

6. **Modified Matching, Modeling and Multiple Imputation (MMM.M)**: The modification of Rubin and Zanutto's MMM method proposed in section 4.3. It follows the same steps of MMM, but instead of selecting substitutes for the sub-sample of $n^* = 0.3n_m$ respondents from the pool of *unsampled* units, these substitutes are selected from the pool of the *remaining respondents* (without replacement). Again, $M = 10$ multiple imputations were used.

7. **Calibrated Matching Substitution (MSub.C)**: The approach proposed in section 4.4 using for the calibration a chi-square distance measure $G_i\left(W, D\right) = \left(w_i - d_i\right)^2 \Big/ 2d_i$, one of the most often used distance measure in calibration applications (Deville and Särndal, 1992; Särndal, 2007).

For each of the seven methods, an estimate of the population mean is computed from each of 5,000 repeated samples. The seven methods were compared using the following measures:

1. Relative change of the empirical bias of the estimate of $\overline{Y}$ compared to the unadjusted respondent mean using an inflated sample size (method ISS):

$$RB_m = 100 \times \frac{\text{Bias}\left(\overline{y}_{ISS}\right) - \text{Bias}\left(\overline{y}_m\right)}{\text{Bias}\left(\overline{y}_{ISS}\right)} = 100 \times \frac{\left(\sum_{k=1}^{5000} \frac{\overline{y}_{ISS,k}}{5000} - \overline{Y}\right) - \left(\sum_{k=1}^{5000} \frac{\overline{y}_{m,k}}{5000} - \overline{Y}\right)}{\left(\sum_{k=1}^{5000} \frac{\overline{y}_{ISS,k}}{5000} - \overline{Y}\right)} \qquad (10)$$

where $\overline{y}_m$ denotes the estimate of $\overline{Y}$ for method $m =$ ISS.W, MSub, MSub.W, MMM, MMM.M, & MSub.C .

2. Relative change in the empirical sampling variance of the estimate of $\overline{Y}$ compared to the empirical variance of the complete response (CR) estimate $\overline{y}_{CR} = \sum_{i=1}^{n} y_i \big/ n$:

$$RV_m = 100 \times \frac{\mathrm{Var}\left(\overline{y}_m\right) - \mathrm{Var}\left(\overline{y}_{CR}\right)}{\mathrm{Var}\left(\overline{y}_{CR}\right)} = 100 \times \frac{\left(\sum_{k=1}^{5000} \frac{\left[\overline{y}_{m,k} - \mathrm{E}\left(\overline{y}_m\right)\right]^2}{5000} \right) - \left(\sum_{k=1}^{5000} \frac{\left[\overline{y}_{CR,k} - \mathrm{E}\left(\overline{y}_{CR}\right)\right]^2}{5000} \right)}{\left(\sum_{k=1}^{5000} \frac{\left[\overline{y}_{CR,k} - \mathrm{E}\left(\overline{y}_{CR}\right)\right]^2}{5000} \right)} \qquad (11)$$

where $\mathrm{E}\left(\overline{y}_m\right) = \sum_{k=1}^{5000} \frac{\overline{y}_{m,k}}{5000}$ and $\overline{y}_m$ denotes the estimate of $\overline{Y}$ for method $m =$ ISS, ISS.W, MSub, MSub.W, MMM, MMM.M, & MSub.C.

3. Empirical root mean square error:

$$RMSE_m = \sqrt{\sum_{k=1}^{5000} \frac{\left(\overline{y}_{m,k} - \overline{Y}\right)^2}{5000}} \qquad (12)$$

where $\overline{y}_m$ denotes the estimate of $\overline{Y}$ for the method $m =$ ISS, ISS.W, MSub, MSub.W, MMM, MMM.M, & MSub.C.

It is worth noticing that each of these measures has different reference points. RB_m uses the unadjusted respondent mean of the ISS method as a baseline, since this approach is most likely to produce the largest bias across all the studied methods when the missing mechanism is not MCAR. RV_m sets the sampling variance of the complete response mean as a benchmark, because, under the presence of nonresponse, it will be the smallest value that could be obtained among the evaluated approaches on this study. Finally, to measure the total error each one of the methods can incur, the $RMSE_m$ measure the deviation of their estimates to the true population parameter $\overline{Y}$ they are attempting to estimate.

4.5.2 Simulation Results

Simulation Study 1

Table 4.3 summarizes the results for Population 1. The missing data mechanism in Table 4.1 for Population 1 leads to respondents with smaller values on the survey variable Y, because respondents are more likely to have units with the matching covariate $x_i = 0$ and smaller values on the modeling/calibration covariate Z. Therefore, the unadjusted respondent mean of the ISS method is severely biased, underestimating the population mean by about 37%, with the nonresponse bias dominating the RMSE of this estimation method. The bias in ISS.W is much smaller than that in ISS, but still substantial, underestimating the population mean by 21%. This is because the response propensities used to make the nonresponse adjustment in this method are estimated using only the modeling/calibration covariate Z, while the matching covariate X, which explains both the nonresponse mechanism and the survey variable Y, is not used.

The matching substitution method MSub takes into account X and ignores variable Z. Since substitutes are respondents, they tend to have smaller values on Z, and consequently on Y, which is not adjusted by using only a matching substitution based on X. Also because both matching and modeling/calibration variables have the same association level with the survey variable and the nonresponse mechanism, the bias on this method is similar to ISS.W. Furthermore, both methods have virtually the same sampling variance and RMSE. Obviously, if the matching or the modeling/calibration covariate had a larger predictive power to explain either the survey outcome or the nonresponse mechanism, one method would lead to estimates with smaller bias and sampling variance than the other.

MSub.W takes into account both X and Z in the nonresponse adjustment. Although the bias is largely reduced, it is still not completely eliminated. MMM produces an essentially unbiased estimate for the population mean in this population (the empirical absolute relative bias in this simulation was under 2%). However, MMM is more costly because data on an additional 108 units, on average, are needed for the matches to the sub-sample of respondents. Despite the larger cost, the MMM sampling variance is about 58% larger than the complete response sampling variance. This variance can be reduced by increasing the size of the sub-sample of respondents to have substitutes selected, as discussed by Zanutto (1999) and Rubin and Zanutto (2002), increasing survey costs by additional data collection. Such a trade-off between bias and variance can be better compared through the RMSE. Despite producing an unbiased estimate, the sampling variance of MMM is much larger than the sampling variance of MSub.W; their RMSE are almost equivalent.

The MMM.M method also leads to essentially an unbiased estimate of the population mean without the additional cost of data collection for substitutes for a sub-sample of the respondents. Unsurprisingly, the MMM.M sampling variance is even larger than MMM's since the substitutes for the sub-sample needed in MMM.M is obtained from the pool of remaining respondents and, therefore, no new information is added to the sample.

Not only the calibrated matching substitution MSub.C decreases the bias as much as the two MMM methods, but it produces estimates with much smaller sampling variance. As a result, MSub.C has sampling variance and RMSE smaller than any of the other methods considered for population. Moreover, MSub.C does not require the collection of additional data for its estimation.

Table 4.3: Simulation 1, Population 1: *RB*, *RV*, and *RMSE* by Method

Method (*m*)	RB_m (%)[1]	RV_m (%)[2]	$RMSE_m$ (x 100)
ISS	0*	-17.2	47.4
ISS.W	43.8	-13.9	28.0
MSub	45.1	-7.0	27.6
MSub.W	81.7	-3.5	14.0
MMM	95.8	57.8	14.3
MMM.M	92.4	135.0	17.7
MSub.C	88.5	18.4	13.4
CR	99.9	0*	11.3

[1] Compared to bias in ISS.

[2] Compared to the CR sampling variance.

* Zero by definition

Respondents in Population 2 tend to have smaller values of Z and, therefore, smaller values of Y (see Table 4.1).Therefore, again, ISS underestimates the population mean by about 44%. Since the matching covariate X is not associated with either the survey outcome or the nonresponse mechanism, the nonresponse adjustment in ISS.W completely eliminates the nonresponse bias by adjusting respondent data according to estimated response propensities using Z (see Table 4.4). The matching substitution MSub, on the other hand, does not reduce the nonresponse bias at all since the responding substitutes have smaller values of Z , which are not adjusted in MSub. MSub.W substantially improves the method, although not quite as effectively as the weights in ISS.W. Both MMM and MMM.M lead to unbiased estimates and similar sampling variances. The calibrated matching substitution MSub.C completely removes the nonresponse bias, but it does not lead to the smallest RMSE among the methods studied here, which is obtained by the ISS.W method. The difference between these two methods is due their sampling variance, with ISS.W giving slightly more precise estimates than the MSub.C.

Table 4.4: Simulation 1, Population 2: *RB*, *RV*, and *RMSE* by Method

Method (m)	RB_m (%)[1]	RV_m (%)[2]	$RMSE_m$ (x 100)
ISS	0*	-3.5	24.5
ISS.W	98.6	-1.3	6.0
MSub	1.0	-2.3	24.3
MSub.W	85.2	-2.0	6.9
MMM	98.9	46.7	9.9
MMM.M	98.7	37.3	9.3
MSub.C	99.2	3.4	6.5
CR	99.4	0*	6.2

[1] Compared to bias in ISS.
[2] Compared to the CR sampling variance.
[*] Zero by definition

None of the methods studied here are suitable to handle the nonresponse mechanism in Population 3. The difference between respondents and nonrespondents is not only explained by the matching and modeling/calibration covariates, but also by an unobserved variable U . Table 4.5 confirms that none of these methods completely eliminate the nonresponse bias. Aside from that, however, the patterns of the results in Population 3 are very similar to the results in Population 1. First, the methods that adjust for both Z and X covariates – MSub.W, MMM, MMM.M and MSub.C – tend to be more successful in decreasing the nonresponse bias. Only adjusting for one of these covariates, as in ISS.W and MSub, leads to a smaller (though still substantial) bias, compared to the ISS baseline, with both approaches having an empirical absolute relative bias of about 20% in this simulation. The sampling variance of MMM.M method is again much larger than the original Rubin and Zanutto method. However, unlike the results in Population 1, MSub.C produced an estimate for the population mean with a bias equivalent to either of the MMM methods. Moreover, the MSub.C was again the method with the smallest RMSE among the procedures evaluated on these simulations, with the MMM also performing very well.

Table 4.5: Simulation 1, Population 3: *RB*, *RV*, and *RMSE* by Method

Method (*m*)	RB_m (%)^1	RV_m (%)^2	$RMSE_m$ (x 100)
ISS	0*	-21.6	62.5
ISS.W	26.3	-19.1	46.7
MSub	29.0	-14.1	45.2
MSub.W	51.1	-12.5	32.2
MMM	58.9	82.2	29.9
MMM.M	58.0	152.8	31.8
MSub.C	56.2	6.6	29.7
CR	99.9	0*	12.2

[1] Compared to bias in ISS.

[2] Compared to the CR sampling variance.

[*] Zero by definition

Simulation Study 2

Tables 4.6, 4.7, 4.8 and 4.9 show the results for four populations studied in Simulation 2. These are explicitly clustered populations with a two-stage cluster sample design, in which the matching substitution is performed using cluster indicators seeking to adjust the clustering effect on the nonresponse.

The nonresponse mechanism in Population 4 is MCAR. Thus, the unadjusted respondent mean in ISS is an unbiased estimator of the population mean. All the other methods also produce unbiased estimates for the population mean in this population. For this reason, the relative bias reduction result is not shown on Table 4.6. The effect of the modeling/calibration covariate Z over the survey outcome is much larger than the clustering effect (since the intra-cluster correlation of the survey variable in this population is approximately zero), which might explain why the sampling variance of ISS.W and MSub.W is smaller than the other methods, especially compared to the MMM methods that intrinsically have larger sampling variance due to the imputation variability. Hence, overall, ISS.W is the method with the best performance under this population, but the difference relative to the other methods is quite small.

Table 4.6: Simulation 2, Population 4: *RV* and *RMSE* by Method

Method (m)	RV_m (%)[1]	$RMSE_m$ (x 100)
ISS	1.9	17.9
ISS.W	-12.0	16.6
MSub	8.4	18.4
MSub.W	-6.6	17.1
MMM	8.2	18.4
MMM.M	6.2	18.2
MSub.C	0.8	17.7
CR	0*	17.7

[1] Compared to the CR sampling variance.
[*] Zero by definition

In Population 5, methods that adjust the respondents by the modeling/calibration covariate Z, like ISS.W, MMM, MMM.M, and MSub.C, completely remove the nonresponse bias present in the unadjusted respondent mean in ISS. However, the nonresponse adjustment in MSub.W is not as effective in reducing bias as the others. Similar to previous results in the Simulation Study 1, the calibration matching substitution led to smaller sampling variances, and thus also smaller RMSE, than the MMM methods. Interestingly, just as in Simulation 1 Population 2, MMM.M produced a slightly smaller sampling variance than the MMM, because in these two populations both the survey variable and the nonresponse mechanism are explained only by the modeling/calibration covariate and not by the matching variable. This suggests not collecting additional information for substitutes of the sub-sample of respondents might not decrease the efficiency of the MMM.M, both in terms of bias reduction and sampling variance, when the matching covariate is not an important explanatory factor for Y and nonresponse. As in Simulation 2 Population 4, the nonresponse weighted-adjusted mean ISS.W led to the smallest RMSE due to the smaller sampling variance produced by this method.

Table 4.7: Simulation 2, Population 5: *RV* and *RMSE* by Method

Method (*m*)	RB_m (%)[1]	RV_m (%)[2]	$RMSE_m$ (x 100)
ISS	0*	5.7	53.4
ISS.W	98.8	-11.5	16.6
MSub	-16.4	12.1	61.3
MSub.W	63.0	-5.8	25.3
MMM	99.2	8.8	18.4
MMM.M	99.3	7.0	18.3
MSub.C	99.1	1.5	17.8
CR	99.3	0*	17.7

[1] Compared to bias in ISS.

[2] Compared to the CR sampling variance.

[*] Zero by definition

With the CSNI missing mechanism in Population 6, in which the nonresponse was caused solely by the clustering effect, it is not surprising that the nonresponse adjustment using the modeling/calibration variable Z in ISS.W is totally ineffective in reducing the nonresponse bias (see Table 4.8). On the other hand, the matching substitution methods are essentially equally effective in bias reduction. The sampling variances of these methods are all very similar, such that no method clearly outperforms the others in terms of RMSE.

Table 4.8: Simulation 2, Population 3: *RB*, *RV* and *RMSE* by Method

Method (*m*)	RB_m (%)[1]	RV_m (%)[2]	$RMSE_m$ (x 100)
ISS	0*	-19.4	143.0
ISS.W	-0.2	-20.4	143.1
MSub	99.6	0.6	60.3
MSub.W	99.7	-0.7	59.9
MMM	99.6	0.8	60.4
MMM.M	99.5	0.5	60.3
MSub.C	99.7	-0.2	60.1
CR	99.6	0*	60.2

[1] Compared to bias in ISS.

[2] Compared to the CR sampling variance.

[*] Zero by definition

Population 7 presents the most complex structure in the simulation. Both the CSNI nonresponse mechanism and the survey outcome are explained by the clustering effect and the modeling/calibration covariate Z. For this reason, procedures that adjust only by one of these variables, such as ISS.W or MSub, fail to completely eliminate the nonresponse bias, although they do reduce it to some extent (see Table 4.9). The matching substitution MSub followed by nonresponse adjusted MSub.W are more successful, but also don't completely remove the nonresponse bias. The two MMM procedures and the calibrated matching substitution MSub.C are the only methods to produce essentially unbiased estimates for the population mean in this population. Overall, in terms of sampling variance and RMSE, the three of them also perform very similarly, with a slight advantage of the MSub.C method due its smaller sampling variance.

Table 4.9: Simulation 2, Population 7: *RB*, *RV*, and *RMSE* by Method

Method (m)	RB_m (%)[1]	RV_m (%)[2]	$RMSE_m$ (x 100)
ISS	0*	-27.3	187.5
ISS.W	25.6	-20.7	144.5
MSub	67.7	-11.1	81.3
MSub.W	87.8	-7.2	62.0
MMM	99.7	1.1	60.5
MMM.M	99.7	0.6	60.4
MSub.C	97.4	-0.8	60.1
CR	99.7	0*	60.2

[1] Compared to bias in ISS.
[2] Compared to the CR sampling variance.
[*] Zero by definition

4.6 Discussion

In general, the results of Simulation Study 1 show that the calibrated matching substitution is a strong candidate to adjust for potential differences between nonrespondents and their substitutes on variables that cannot be used in the matching procedure when the nonresponse is caused by hidden clustering. Although the calibrated matching substitution method led to a slightly smaller reduction of bias compared to the MMM methods in two of the three populations evaluated, it also produced a more precise estimate of the population mean, such that, overall, its RMSE was substantially smaller than most of the other alternatives across the three populations.

Further, the MMM.M method proved to be a viable alternative to Rubin and Zanutto's original method, achieving similar levels of bias reduction. Despite producing estimates with a larger sampling variance, this modified MMM does not require additional units to be collected for estimation purposes, which would incur in extra data collection costs for the survey operation. The trade-off of sample size between these two methods is the key motivation for the development of MMM.M. The cost savings could purchase additional sample selection and data collection, reducing the MMM.M sampling variances further. Of course, such cost savings depend on how large a sub-sample of the respondents would be selected to be substituted in the MMM method. Zanutto (1998) gives a brief discussion about this choice, concluding that it is another trade-off problem between sampling variance and survey costs. That is, the larger this sub-sample is, the smaller the sampling variance will be, at a cost of larger survey costs. This can indicate that the losses in precision in MMM.M may actually be compensated for when compared to certain sizes for the sub-sample of respondents to be substituted in MMM. Moreover, for a fixed total survey cost, MMM.M may actually yield estimates with smaller sampling variances than MMM. More research on these cost trade-offs should be addressed in future studies of these MMM methods.

In Simulation Study 2 the calibrated matching substitution MSub.C did not provide as favorable results as in Simulation Study 1, though it performed just as well as the alternatives. In particular, it continued to achieve the same levels of bias reduction as the MMM methods and still led to smaller sampling variances, but to a lesser degree among these populations. Interestingly, the MMM.M not only kept the same levels of bias reduction, but it also presented a slightly smaller sampling variance than the original method proposed by Rubin and Zanutto. This difference, however, is so small that it may be due to simulation error. Nonetheless, the MMM.M method remains the more affordable alternative to MMM for bias reduction.

Although unit nonresponse has received a lot of attention in recent decades in the survey community, through numerous studies and research on weighting, imputation and field method to increase rates, substitution has been mostly neglected by the field and often considered an illegitimate method for dealing with this problem. While substitution may not necessarily reduce nonresponse bias, under certain conditions, it can perform just as well as any other statistical adjustment that uses the same information.

This paper presented two alternatives for the MMM method that avoid collecting additional substitutes beyond those for the nonrespondents. First, a minor modification of Rubin and Zanutto's method was suggested, in which substitutes for the sub-sample of respondents are selected from the pool of existing respondents. Because substitutes are also respondents, this modification was hypothesized to have the same bias reductions properties than the original procedure. However, because these substitutes for the sub-sample of respondents are already part of the sample, losses in precision were expected compared to Rubin and Zanutto's method.

The simulation studies confirmed both expectations. The bias reductions were virtually the same for these two methods across all simulations and in most scenarios the sampling variance of MMM.M was larger than MMM's. In the second simulation study, however, the two methods led to estimates with very similar levels of variability, indicating that there are situations in which the losses in precision on the modified version of the method are not substantial. Therefore, if the extra cost associated with the selection of additional substitutes for the sub-sample of respondents is prohibitive, the proposed modified version of the MMM method can be considered for the same levels of nonresponse bias reduction, but with some loss in precision.

The calibrated matching substitution was proposed as an attempt to overcome the two major disadvantages of the MMM methods: (i) the cost of the additional substitutes for the sub-sample of respondents and (ii) the inflation of the sampling variance due to the multiple imputation variability. While the MMM.M method solved the first problem, it may lead to inflation of the sampling variance, as discussed above. As the simulation results showed, using calibration to adjust for differences between nonrespondents and their substitutes not only reduces the nonresponse bias to levels comparable to the MMM methods, but manage to keep the sampling variance to similar levels of a complete response estimate (or at least did not lead to substantial increases).

This new proposed method also has some disadvantages compared to the MMM methods. First, it may not always reduce nonresponse bias to the same extent as the MMM method, as can be observed by the results of the simulation on Population 1 in Table 4.3. This is due to the fact that the adjustment between the nonrespondents and their substitutes in terms of the modeling or calibration covariates happens at the aggregate level, that is, for the totals in the sample, whereas in the MMM method this adjustment is much finer since it occurs at the element level. Nonetheless, the simulations showed that in general, these small differences in bias reduction between these two methods are countered by the gains in precision given by the calibration procedure, making the calibrated matching substitution overall a more accurate method than the MMM methods.

An important advantage of MMM methods over calibrated matching substitution is model flexibility. In general, the calibration procedure generates a set of weights based in a single model that is used for the estimation of every survey statistic. Although, in theory, different sets of weights could be computed assuming different models for each survey variable, this would be impractical for most surveys, which require the estimation not only of descriptive single-variable statistics, but also multiple variable estimates, such as regression or correlation coefficients. In that sense, the MMM methods are much more flexible, because they allow different models for the imputation of each variable in the survey. In this paper, this feature was not very evident because the simulation studies were evaluating only a single-variable population mean, but this is an important practical component, as most of surveys are multi-purpose and multi-variable. On the other hand, having a single set of weights that can be applied to every survey statistic is more convenient than having to model every single variable in a survey.

Although variance estimation was not discussed in this paper, it is another important problem that should be addressed in future research. Under a MAR mechanism, the standard error of an estimate that uses substitutes for nonrespondents is approximately the same as a complete response sampling variance estimate (Vehovar, 1999). Therefore, standard techniques for sampling variance estimation can be applied for the substitution methods reviewed in this study. For the MMM method, since it relies on imputation, proper variance estimates can be obtained by multiple imputation using Rubin’s combining rule, as suggested by Rubin and Zanutto (2002). The variance of the calibrated matching substitution should adequately take into account the calibration procedure, which might not be as straightforward as the multiple imputation approach. One alternative is to use the GREG sampling variance estimate approximation (Deville and Särndal, 1992) usually used for sampling variance estimation of calibrated estimates. Another alternative is to use repeated replication methods such as jackknife or bootstrap. The properties of these methods for sampling variance estimation of the proposed calibrated matching substitution should also be studied in future research, particularly when the data is MNAR.

Throughout the simulations conducted in this paper, it was assumed that the substitution procedure of the nonrespondents for all the methods evaluated was fully successful. That is, for every nonrespondent, it was possible to find a responding unit to substitute it. In practice, however, it is very likely that no responding units are found to substitute for some of the nonrespondents, even after multiple attempts with different substitute candidates. The calibration matching substitution can still be applied in this case using restriction (9) without making any modifications. The MMM methods, on the other hand, would need to be altered to take into account the nonrespondents for which there were no responding substitutes available, which could potentially make the procedure more complicated. The properties of these methods under these conditions should also be topic of future research.

Finally, this paper considered the case in which there is only one matching covariate and one (quantitative) modeling/calibration covariate. The methods proposed here can be readily extended to situations with multiple variables for matching and modeling (or calibration), either quantitative or qualitative (categorical). The general results observed on the simulations conducted in this study are not likely to change significantly, but it would be important to conduct further research on these methods under these more general circumstances. Furthermore, future evaluations of these methods should also analyze other, more complex, estimators, such as the median and regression coefficients.

References

Chiu, W. F., Yucel, R. M., Zanutto, E. and Zaslavsky, A. M. (2005). Using Matched Substitutes to Improve Geographically Linked Databases. *Survey Methodology*, Vol. 31, No. 1, pp. 65-72.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methodology Section*, ASA, 41–49.

Deville, C. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

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

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.

Lumley, T. (2012). survey: analysis of complex survey samples. R package version 3.28-2.

Lumley T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*. 9 (1): 1-19

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*, 29, pp. 159-183.

Rubin, D. B. (1987). *Multiple Imputation for Survey Nonresponse*. New York: John Wiley and Sons.

Rubin, D. B., and Zanutto, E. (2002). Using Matched Substitute to Adjust for Nonignorable Non response through Multiple Imputation. In *Survey Nonresponse*, edited by R. Groves, R. J. A. Little, and J. Eltinge. New York: John Wiley, pp. 389-402.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, *33*(2), 99-119.

Skinner, C. J., and D’Árrigo, J. D. (2011). Inverse probability weighting for clustered nonresponse. *Biometrika*, 98, 4, pp. 953-966.

Trautmann, H., Steuer, D., Mersmann, O. and Bornkamp, B. (2014). truncnorm: Truncated normal distribution. R package version 1.0-7. http://CRAN.R-project.org/package=truncnorm

Vehovar, V. (1999). Field Substitution and Unit Nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, pp. 335-350

Yuan, Y., and Little, R. J. A. (2007). Model-based estimates of the finite population mean for two-stage cluster sample with unit non-response. *Applied Statistics*, 56, Part 1, pp. 79-97.

Zanutto, E. (1998). Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes. Doctorate thesis submitted for the graduate faculty of Harvard University, May, 1998.

CHAPTER V

A Substitution Procedure for Missing Not at Random Mechanism

Summary

Although commonly used in many surveys as a strategy to deal with unit nonresponse, substitution is frequently criticized and has received very little attention by survey researchers. In fact, there is evidence to suggest that the performance of substitution as a strategy to mitigate nonresponse is comparable to other adjustment methods, such as weighting or imputation (Vehovar, 1999; Rubin and Zanutto, 2002). However, as with many other nonresponse adjustment methods, research on and applications of this method has been limited to ignorable nonresponse mechanisms. This paper presents a new substitution procedure that incorporates a nonignorable nonresponse mechanism in the selection of the substitutes of the nonresponding units through the use of pattern-mixture models. This method can be employed to perform sensitivity analysis of a range of missingness models using additional real data of the substitutes, as opposed to other methods that use predicted values under a model or data from hot-deck donors already present in the responding sample. This new methodology is evaluated and compared to other nonresponse adjustment methods through a simulation study.

5.1 Introduction

Nonresponse occurs when a sampled unit fails to provide either part (item nonresponse) or all (unit nonresponse) of the information requested in a survey. This may be due to noncontact, refusal, or other reasons, such as an inability to understand the request. Nonresponse can lead to bias in survey estimates, and as a result, this source of error, with regards to unit nonresponse in particular, has been increasingly studied in statistics and survey methodology. The study has become more intense as response rates have declined in recent decades (de Leeuw and de Heer, 2002; Rand, 2006, Bethlehem et al., 2011). While the size of the error is related to the level of nonresponse, the relationship between response rates and nonresponse error has been called into question by several studies (Keeter, et al., 2000; Merkle and Edelman, 2002; Curtin, Presser and Singer, 2005; Keeter, et al., 2006; Groves and Peytcheva, 2008), highlighting the importance of a careful exploration of all existing methods for dealing with nonresponse.

Research has also examined alternative methods for compensating for nonresponse and the capacity of those methods to reduce error associated with it. Substitution is a widely used approach to dealing with unit nonresponse at the fieldwork stage of a survey. This method consists of replacing nonresponding sampled units with replacement units not originally selected in the sample. Despite its popularity among practitioners, most survey methodology and sampling textbooks either ignore (e.g., Cochran, 1977; Särndal et al., 1992; Groves et al., 2009) or present only a brief discussion of substitution (e.g., Kish, 1965; Lessler and Kalsbeek, 1992; Lohr, 1999; Little and Rubin, 2002). The literature, in general, tends to criticize substitution and recommends avoiding its use, although no conclusive evidence suggesting it performs worse than competing alternatives, such as weighting or imputation, has been found. In fact, some studies have demonstrated that, under certain conditions, substitution performs just as well as other nonresponse adjustment procedures (Vehovar, 1999; Rubin and Zanutto, 2002).

One limitation of substitution methods used in practice is that they assume that nonresponse is ignorable (Rubin, 1987), operating by a missing data mechanism that is either missing completely at random (MCAR) or missing at random (MAR). Under these mechanisms, the distribution of missingness depends only on a set of variables observed for respondents, nonrespondents, and their substitutes. These variables are usually used either to find a substitute – as in a matching substitution – or in post-data collection nonresponse adjustments, such as weighting or imputation.

Rubin and Zanutto (2002) proposed a method that attempts to adjust for nonignorable nonresponse through the use of matched substitutes with multiple imputation. However, their method is only successful in addressing the nonignorability problem if the variables used for selecting the substitutes incorporate unobserved characteristics – through hidden clustering or contextual effects – that are related to both the survey variables and the nonresponse mechanism. If there are other unobserved variables that explain the differences between nonrespondents and their substitutes, their method is unable to completely eliminate nonresponse bias. This is the case when nonresponse is missing not at random (MNAR). An extreme example occurs when missingness depends on the survey variables themselves.

Nonignorable nonresponse poses a challenging problem for survey statisticians because it assumes that missingness depends on unobserved variables and, therefore, requires strong and untestable assumptions. Nonetheless, there have been methods proposed to assess the impact of nonresponse in this kind of situation through sensitivity analysis with pattern-mixture models (Little, 1993; Little and Rubin, 2002; Andridge and Little, 2011). This paper proposes adapting the pattern mixture hot-deck imputation procedure developed by Sullivan and Andridge (2015) to substitution by using pattern-mixture models (PMM) to assist in the selection of substitutes. Obtaining the most accurate estimates for population parameters is an important objective for every nonresponse adjustment method. However, due to the nonignorable nature of the missing mechanism assumed here, a primary concern of this method is the detection of the risk of nonresponse bias through sensitivity analysis. This method can also be used in nonresponse follow-up, particularly when there are persistent nonrespondents or nonresponse in longitudinal surveys. The performance of the proposed method is evaluated through a simulation study.

5.2 Pattern-Mixture Model

Let X be a fully observed auxiliary variable and Y be a survey variable not observed for nonrespondents, in which the missingness distribution is given by the indicator variable M ($M=1$, if Y is missing and $M=0$ if Y is observed). Little (1994) proposes the following bivariate normal pattern-mixture model for the joint distribution of Y, X and M:

a) $$\left(Y,X\mid M=m\right)\sim N_2\left(\begin{pmatrix}\mu_y^{(m)}\\ \mu_x^{(m)}\end{pmatrix},\begin{pmatrix}\sigma_{yy}^{(m)} & \rho_{yx}^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}}\\ \rho_{yx}^{(m)}\sqrt{\sigma_{yy}^{(m)}\sigma_{xx}^{(m)}} & \sigma_{xx}^{(m)}\end{pmatrix}\right),m=0,1$$

b) $$M\sim Bernoulli\left(\pi\right),\quad \pi=\Pr\left(M=1\right)$$

Little (1994) notes that from the eleven parameters $\phi^{(m)}=\left(\mu_y^{(m)},\mu_x^{(m)},\sigma_{yy}^{(m)},\sigma_{xx}^{(m)},\rho_{yx}^{(m)}\right)$, $m=0,1$, and π of this model, only eight are identified from the data:

$$\phi_{id} = \left(\mu_y^{(0)}, \mu_x^{(0)}, \sigma_{yy}^{(0)}, \sigma_{xx}^{(0)}, \rho_{yx}^{(0)}, \mu_x^{(1)}, \sigma_{xx}^{(1)}, \pi \right)$$

This is because the data do not provide any information about some of the parameters of the conditional distribution of Y given X for nonrespondents ($M = 1$). Only the parameters of the respondent distribution ($M = 0$) and those of the marginal distribution of X of the nonrespondents are identified and readily estimable.

While this model is under-identified, under certain restrictions on these parameters, and under assumptions about the missing data mechanism, this model can become identified. For example, the assumption that the nonresponse mechanism is MAR implies that the distribution of Y given X is the same for respondents and nonrespondents, identifying the remaining three parameters in the model.

Little (1994) proposes a more general restriction, in which the missingness of Y given $\left(X, Y \right)$ depends only on a linear combination of Y and X. More specifically, he proposes assuming that, for some function f,

$$P\left(M = 1 | Y, X \right) = f\left(X + \lambda Y \right).$$

Under the assumption that $\left(X, Y \right)$ is independent of $M \,|\, X + \lambda Y$, the parameters of the pattern-mixture model are identified.

Since the data do not provide any information about the parameter λ, Little (1994) suggests evaluating the estimates of the substantive parameters of interest over a range of different plausible values for λ to assess the sensitivity of inferences to the missing mechanism assumptions. For example, if $\lambda = 0$, the missing mechanism is MAR. On the other hand, if $\lambda = \infty$, all the missingness will depend on the Y variable, an “extreme” case of MNAR.

Andridge and Little (2011) use $\lambda = \{0,1,\infty\}$ to perform a sensitivity analysis of model performance. They suggest using the intermediate case of $\lambda = 1$, in which the auxiliary variable X and the survey outcome Y have the same weight in explaining the nonresponse mechanism, because in this case the standardized bias of the respondent mean of Y is equal to the standardized bias of the respondent mean of X, that is, $E\left(\bar{y}_R - \mu_y\right)\Big/\sqrt{\sigma_{yy}} = E\left(\bar{x}_R - \mu_x\right)\Big/\sqrt{\sigma_{xx}}$, regardless of the estimated correlation between X and Y.

If more than one fully observed auxiliary variable is available, say a set of variables $Z = \left(Z_1, Z_2, ..., Z_p\right)'$, is available, Andridge and Little (2011) suggest using a *proxy* pattern-mixture model to account for the nonresponse mechanism. This method consists of creating a "proxy" variable X by first regressing Y on Z using the respondent data and then taking X to be the predicted values of Y under this model based on Z, available for both respondents and nonrespondents. The bivariate normal pattern-mixture model proposed by Little (1994) can then be employed using the proxy variable X. Moreover, to improve interpretability, Andridge and Little (2011) suggest rescaling the proxy variable X on the distribution of the missingness of Y given $\left(X, Y\right)$ to have the same variance of Y, that is, $P\left(M = 1 | Y, X\right) = f\left(X\sqrt{\sigma_{yy}^{(0)}\big/\sigma_{xx}^{(0)}} + \lambda Y\right)$.

Sullivan and Andridge (2015) proposed adapting the proxy pattern-mixture model to hot-deck imputation, which they called proxy pattern-mixture (PPM) hot-deck, to accommodate nonignorable missing data to this imputation procedure, extending the work by Siddique and Belin (2008) of hot-deck imputation to nonignorable nonresponse. The premise of this method rests on the computation of predictions for nonrespondents' outcome variable and a bootstrap sample of respondents based on a pattern-mixture model conditional to a value of λ. The predicted values for Y are used to calculate distances between donors on the bootstrap sample of respondents to nonrespondents and select donors for the nonrespondents based on probabilities inversely proportional to the k^{th} power of those distances (they use $k = 3$ in their simulations and application). This process is repeated D times as in a multiple imputation procedure. The method also employs different values of λ in the model, allowing the subsequent multiply imputed values to incorporate sensitivity to nonignorable nonresponse.

5.3 Pattern-Mixture Model Substitution

Substitution is somewhat similar to hot-deck imputation. While hot-deck imputation selects donors for nonrespondents from among respondents already in the sample, substitution seeks donors from among the unsampled units in the population. This suggests that Sullivan and Andridge's method can be used to accommodate a nonignorable nonresponse mechanism in substitution for nonresponse.

Hence, the objective of this study is to adapt the PPM hot-deck method proposed by Sullivan and Andridge (2015) to a substitution procedure that can accommodate a variety of assumptions about the nonresponse mechanism, encompassing MAR and different degrees of MNAR. This procedure will be called pattern-mixture model (PMM) substitution hereafter. As in the matched substitution method proposed by Rubin and Zanutto (2002), it is assumed there is at least one auxiliary variable X observed for all the units in the population. In the survey sampling literature, such an auxiliary variable is sometimes referred to as a frame variable. If there is a vector of auxiliary variables, Z, they can be reduced to a single "proxy" variable using the method proposed by Andridge and Little (2011). These types of variables are not usually available for units like households or individuals in multistage surveys, but they are fairly common for primary and secondary sampling units such as enumeration areas, counties, census tracts, establishments.

While in most applications substitutes are selected for nonrespondents during the fieldwork stage of the survey, a different approach is proposed here. It is assumed that at some point during data collection, before the selection of the substitutes, the data on the survey variable Y for the respondents are available to fit a pattern-mixture model, conditional on a value for the parameter λ. Under this model, predicted values for the nonrespondents and the unsampled units in the population are computed and substitutes are selected based on some measure of distance of these predictions. Compared to the standard use of substitution, this approach has the advantage of incorporating a nonignorable nonresponse adjustment through a matching substitution on the predictive values under the PMM. Such adjustment is based on the association of the auxiliary variable X and the survey outcome Y among the respondents, and the differences between respondents and nonrespondents on X. Below, the implementation of this procedure is described in detail.

For simplicity, assume that a simple random sample of size n is drawn from a finite population of size N, and r units are respondents. The survey variable Y is observed only for these r units.

For a given value of the parameter λ:

1. Compute the predicted values for the $n-r$ nonresponding units in the sample and the $N-n$ unsampled units in the population based on the pattern-mixture model and parameter restriction given above. Sullivan and Andridge (2015) use the conditional expected value $E\left[Y \mid X, M=m\right]$ for these predicted values. Under this approach, the predictive value for the i^{th} nonrespondent on the sample is

$$\begin{aligned}\hat{y}_i^{(\lambda)} &= \bar{y}_R + \sqrt{\frac{s_{yy}^{(0)}}{s_{xx}^{(0)}}}\left(\frac{\lambda+\hat{\rho}^{(0)}}{1+\lambda\hat{\rho}^{(0)}}\right)\left(\bar{x}_{NR}-\bar{x}_R\right) \\ &\quad+\left[\frac{\hat{\rho}^{(0)}\sqrt{s_{xx}^{(0)}s_{yy}^{(0)}}}{s_{xx}^{(1)}}+\left(\sqrt{\frac{s_{yy}^{(0)}}{s_{xx}^{(0)}}}\frac{\lambda+\hat{\rho}^{(0)}}{1+\lambda\hat{\rho}^{(0)}}\right)\left(1-\frac{s_{xx}^{(0)}}{s_{xx}^{(1)}}\right)\right]\left(x_i-\bar{x}_{NR}\right)\end{aligned}$$

where $\bar{y}_R$ is the respondent sample mean of Y, $\bar{x}_R$ and $\bar{x}_{NR}$ are the respondent and nonrespondent sample means of X, $s_{yy}^{(0)}$ and $s_{xx}^{(0)}$ are the respondent sample variances of Y and X, $s_{xx}^{(1)}$ is the nonrespondent sample variance of X, and $\hat{\rho}^{(0)}$ is the sample correlation between Y and X among the respondents. These are all maximum likelihood estimates of the pattern-mixture model parameter under the identifying restriction proposed by Little (1994) that $P\left(M=1 \mid Y, X\right)=f\left(X+\lambda Y\right)$, for some function f.

Since the units to be used as substitutes of the nonrespondents will ultimately be respondents, the predicted values under the pattern-mixture model for the unsampled units are computed assuming they are respondents as

$$\hat{y}_i^{(0)} = \bar{y}_R + \hat{\rho}^{(0)} \sqrt{s_{yy}^{(0)} / s_{xx}^{(0)}} \left(x_i - \bar{x}_R \right)$$

2. Compute the distance between predicted values for *Y* for a given nonrespondent *j* and all the non- sampled units. Any distance measure could be used, but here the absolute difference $D_{jk}^{(\lambda)} = \left| \hat{y}_j^{(\lambda)} - \hat{y}_k^{(0)} \right|, k = N - n + 1, \ldots, N$, is used. If there is more than one survey variable, say $Y = \left(Y_1, Y_2, \ldots, Y_q \right)'$, a multidimensional distance measure such as the Mahalanobis distance can be used.

3. Select the unsampled unit *k* with the smallest distance D_{jk} as a substitute for nonrespondent *j*. In most applications, substitutes will be selected without replacement. That is, the selected unit for the *j*th nonrespondent would be removed from the pool of unsampled units, but this is not a necessary step if units are allowed to substitute for more than one nonrespondent. Repeat steps 2 and 3 for all nonrespondents.

This process is implemented for a given value of λ , which allows sensitivity to different degrees of nonignorability. For most applications, only one substitute for each nonrespondent would be selected. The performance of the method would be conditional on the validity of the nonresponse assumption represented in λ .

Alternatively, this process could be repeated for different values of λ , say $\lambda = \{0, 1, \infty\}$, as suggested by Andridge and Little (2011) and Sullivan and Andridge (2015) – and determining whether there are large differences in terms of which units would be selected as substitutes for the nonrespondents in each case. If the exact same units are designated as substitutes for each nonrespondent for any of the values of λ , a single substitute per nonrespondent would be selected. If, however, for each value of λ there is a different substitute for each nonrespondent, multiple substitutes could be selected, and a sensitivity analysis conducted across different missing mechanism assumptions. Obviously, from a practical point of view, there is a trade-off between the ability to perform a sensitivity analysis and survey costs associated with the selection of multiple substitutes per nonrespondent. This trade-off is briefly discussed in the conclusions to this paper below.

5.4 Simulation Design

A simulation study was conducted to evaluate the performance of the proposed PMM substitution under different population structures and missing mechanisms. The bias, variance, and mean square error properties of the proposed method are examined and compared to other standard approaches to nonresponse adjustments in survey sampling.

Artificial finite populations of size $N = 10,000$ were generated according to the following bivariate normal distribution:

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{yx} \\ \rho_{yx} & 1 \end{pmatrix}\right), \quad i = 1, ..., N .$$

Five levels of correlation between Y and X, $\rho_{yx} = \{0, 0.2, 0.4, 0.6, 0.8\}$, were used to generate five populations. These correlation levels were chosen to evaluate the performance of the proposed method under different situations. With a null correlation the auxiliary variable will not provide any assistance to the adjustment, whereas as the correlation increases the adjustment through substitution will be more influential. In practice, correlations as high as 0.8 are not very common between survey outcomes and auxiliary variables in surveys, especially with unit nonresponse. The expected highest correlations would be of the order of 0.20 to 0.40, but these stronger correlations were included to allow investigation of the potential larger impact of PMM substitution.

For each population, $K = 5,000$ simple random samples of size $n = 500$ were selected. For each replication, every unit in the population was assigned as a respondent ($m_i = 0$) or a nonrespondent ($m_i = 1$) according to the missing data mechanism generated using the following logistic regression model:

$$\text{logit}\left(\Pr\left(m_i = 0\right) | x_i, y_i\right) = \beta_0 + \beta_1 x_i + \beta_2 y_i, \quad i = 1, \ldots, n$$

where the value of the coefficients $\{\beta_0, \beta_1, \beta_2\}$ are shown in Table 5.1. Three different nonresponse mechanisms were investigated based on the choice of the coefficients. The missing at random or MAR mechanism sets $\beta_2 = 0$. The two not missing at random or MNAR mechanisms set $\beta_2 \neq 0$. Each missing data mechanism was examined at two response rates, 50% and 75%, determined by the choice of the intercept β_0. The values of the slope coefficients, β_1 and β_2, were selected so that the odds of the unit be a respondent are approximately 22% higher by a one unit increase in the predictors.

For the two MNAR mechanisms, different values of λ were used. For a MNAR mechanism in which the nonresponse is explained by both the outcome and auxiliary variables, $\lambda = 1$, and when the nonignorable nonresponse is explained only by the survey variable Y, $\lambda = \infty$.

For simplicity, the same missing mechanism was employed for both the units originally drawn in the sample and for units selected to be substitutes. This implies that that the same survey protocol is applied throughout the fieldwork. While this might not hold true in some instances, it is not feasible to simulate a more general condition without making further assumptions.

Table 5.1. Coefficients of the nonresponse mechanism models

Missing mechanism	Model	Corresponding λ	Response rate	β_0	β_1	β_2
MAR	[X]	0	50%	-0.2	0.2	0
			75%	0.9	0.2	0
MNAR	[X+Y]	1	50%	-0.4	0.2	0.2
			75%	0.7	0.2	0.2
MNAR	[Y]	∞	50%	-0.2	0	0.2
			75%	0.9	0	0.2

This simulation setting also tested how sensitive the proposed PMM substitution method is to violations of the distributional assumptions of the pattern-mixture model. The selection model used here implies that the marginal joint distribution Y and X is normal, whereas the pattern-mixture model assumes conditional normality, given the missing indicator M. Therefore, the correlation between Y and X for the entire sample, ρ_{yx}, might not have been the same from the corresponding correlations for respondents ($\rho_{yx}^{(0)}$) and nonrespondents ($\rho_{yx}^{(1)}$) under the pattern-mixture model. However, because in these simulations the missing mechanism is a linear function of Y and X, these correlations will be the same, as in Andridge and Little (2011).

For each of the 30 combinations of ρ_{yx} and the nonresponse mechanism, K = 5,000 simple random samples of size n = 500 were selected. The inferential objective was to estimate the finite population mean of a survey variable, $\bar{Y} = \sum_{i=1}^{N} y_i \big/ N$, using an auxiliary variable X observed for all the units in the populations. As previously suggested, the proposed PMM substitution method was applied with $\lambda = \{0, 1, \infty\}$. The PPM substitution method was evaluated in terms of the following empirical measures:

1. The empirical bias, $\text{Bias}(\bar{y}) = \text{E}(\bar{y}) - \bar{Y} = \sum_{k=1}^{5000} \frac{\bar{y}_k}{5000} - \bar{Y}$;
2. The empirical sampling variance, $\text{Var}(\bar{y}) = \sum_{k=1}^{5000} \frac{\left[\bar{y}_k - \text{E}(\bar{y})\right]^2}{5000}$ where $\text{E}(\bar{y}_m) = \sum_{k=1}^{5000} \frac{\bar{y}_{m,k}}{5000}$; and
3. The empirical root mean square error, $\text{RMSE}(\bar{y}) = \sqrt{\sum_{k=1}^{5000} \frac{\left(\bar{y}_k - \bar{Y}\right)^2}{5000}}$

The PPM substitution was employed to obtain an estimated mean for each sample. The properties of the PMM substitution mean were compared on the empirical criteria to the following alternative methods:

1. The **Inflated Sample Size mean (ISS)** is the unadjusted respondent mean where the sample size is inflated by the expected response rate $\left(1-\pi\right)$. That is, a sample of size $n' = n/\left(1-\pi\right)$ is selected and the mean of the respondents is used as an estimate of the population mean. Here the known value of the response rate $\left(1-\pi\right)$ is used in the simulations even though in practice this average response rate is estimated beforehand, usually based on previous surveys of a similar target population.

2. The **Inflated Sample Size mean adjusted by nonresponse propensity Weight (ISS.W)** is similar to the ISS where the sample size is inflated by the expected response rate, $\left(1-\pi\right)$, but the respondents are then weighted by the inverse of their predicted response propensities using the auxiliary variable covariate, X, as a predictor of the missing indicator M in a logistic regression model.

3. The **Matching Substitution (MSub) mean** is based on an initial sample of size n where each nonrespondent is substituted with a unit selected from the pool of non-sampled units, chosen by matching each nonrespondent with the unsampled unit to which it is closest in terms of the auxiliary variable X. If the substitute unit turns out to be a nonrespondent, the next closest unsampled unit is selected as the substitute, and this process is repeated until a responding substitute is chosen. If there is more than one unit that can be used as a substitute for a given nonrespondent, the substitute is randomly selected among these units. No further adjustments are done to take into account for possible remaining differences in the auxiliary variable X.

4. The **Matching Substitution with nonresponse propensity Weighted (MSub.W) mean** is similar to MSub, but the values of Y for originally selected and substituted units that respond are weighted by the inverse of their predicted response propensities using the auxiliary variable X as predictor of the missing indicator M in a logistic regression

model. This model is estimated using the data from all respondents and all nonrespondents (originally selected and substitutes units).

The ISS method assumes that the missing mechanism is MCAR, while the ISS.W, MSub, and MSub.W methods assume a MAR mechanism. Therefore, none of them are expected to perform well under a MNAR mechanism, but they serve as a basis of comparison for the MAR case and as a baseline for the level of improvement that can be expected from the proposed method under a nonignorable nonresponse mechanism. With only a single auxiliary variable, MSub is exactly the same as using the PMM substitution with $\lambda = 0$. For the sake of completeness, the results of these two methods are shown separately.

As mentioned previously, if there is more than one auxiliary variable available, the proxy pattern-mixture model approach suggested by Andridge and Little (2011) can be employed for the PMM matching substitution and a distance measure, such as the Mahalanobis distance, can be used for the traditional matching substitution. In this case, the PMM matching substitution and MSub will not necessarily lead to exactly the same results when $\lambda = 0$, but they will likely be very close.

5.5 Results

Figures 5.1 and 5.2 present the empirical expected values of the population mean estimates across the 5,000 simulation replications for ISS, ISS.W, MSub, MSub.W, and the PMM substitution method for a 50% and 75% response rate, respectively. The horizontal red line corresponds to the true population mean and can be used as a basis for evaluation of the estimates' bias. Empirical sampling variances for the estimates of these methods are shown in Figures 5.3 and 5.4, and the empirical root mean square errors are displayed in Figures 5.5 and 5.6. In these these figures the horizontal red line corresponds to the sampling variance or root mean square error of a population mean estimated under complete response. While not actually observed in practice, it serves as a benchmark for the methods evaluated in this study.

The patterns of the results under a 50% response rate are essentially the same as those under a 75% response rate; the only difference between these two response rates is in terms of magnitude. For example, when a method led to biased estimates, the bias was larger under a 50% response rate, as would be expected. In addition, due to larger sample sizes, in general, the sampling variances of the estimates under 75% response rate were smaller. This indicates that the response rate does not change the properties of the methods investigated here other than their magnitude. Therefore, the subsequent discussion of these results will not differentiate between the response rates.

Missingness model [X]

Under a MAR mechanism, in which the missingness mechanism depends solely on the auxiliary variable X, respondents are different from nonrespondents in terms of this covariate. In these simulations, respondents tend to have larger values of X and, because of the positive association with the survey outcome, they also tend to have larger values of Y. Therefore, the respondent mean of the inflated sample size method (ISS) produces estimates with a positive bias for all correlation levels except $\rho_{yx} = 0$, when the bias is null as can be seen in Figures 5.1 and 5.2. Also, as would be expected, this bias increases as the correlation strengthens.

Regardless of the strength of the correlation between the outcome and the predictor, this bias is essentially eliminated for methods that adjust the respondents using the auxiliary variable X -- ISS.W and MSub. Curiously, further nonresponse adjustments on the sample that already used a matching substitution on the same variable in MSub.W produces estimates which are no longer unbiased.

As mentioned previously, the PPM substitution with $\lambda = 0$ is equivalent to MSub and, therefore, also produces unbiased estimates under the MAR mechanism. As the pattern-mixture model is misspecified for other values of λ, it is not surprising that the PMM substitution produces biased estimates of the population mean. The exception is when $\rho_{yx} = 0$, where the estimate that uses $\lambda = 1$ leads to an essentially unbiased estimate. Similarly, when $\rho_{yx} = 0$ and $\lambda = \infty$ is used in the model, the PMM substitution method generates slight overestimates of the population mean.

Figures 5.3 and 5.4 show that under the MAR nonresponse mechanism ISS, ISS.W, MSub, and MSub.W methods produced empirical sampling variances very close to the complete response sampling variance. The exception is for high correlations between X and Y ($\rho_{yx} = 0.6$ and $\rho_{yx} = 0.8$), when the nonresponse weighted-adjusted mean actually showed a slight gain in precision.

The PMM matching substitution with $\lambda = 0$ also led to an empirical sampling variance very similar to the complete response sampling variance. The sampling variances of this method using $\lambda = 1$ were also in general similar to the complete response case, except for $\rho_{yx} = 0.2$ and $\rho_{yx} = 0.4$, when they were slightly larger. On the other hand, the PMM substitution using $\lambda = \infty$, produced estimates with much more variability than all the other methods for intermediate correlations ($\rho_{yx} = 0.2, 0.4, 0.6$). This result demonstrates that the instability of the pattern-mixture model estimates when λ is set to infinity observed by Andridge and Little (2011) carries over to the PMM substitution method.

As a consequence of the bias and variance properties described above, under MAR, the methods that led to the smallest RMSE across all the correlations were the ISS.W, MSub, and the PMM substitution with $\lambda = 0$ (see Figures 5.5 and 5.6). The first two methods are expected to work well under a MAR mechanism. Since the PMM substitution with $\lambda = 0$ employs the correct model for the missing mechanism, it also provides good results, with unbiased estimates and sampling variances similar to a complete response case.

On the other hand, as a result of model misspecification in this missing mechanism, using the PMM with $\lambda = 1$ and $\lambda = \infty$ leads to a larger RMSE, especially for non-zero correlations. Assuming a stronger nonignorable missing mechanism, with $\lambda = \infty$ when the missingness is actually ignorable, clearly leads to larger bias and sampling variance. Therefore, PMM substitution with $\lambda = \infty$ should only be used when there are compelling reasons to believe that nonresponse is driven entirely by the missing variable itself.

Figure 5.1: Empirical expected values of population mean estimates over 5000 simulation replications with a 50% response rate. Red horizontal line denotes the true population mean.

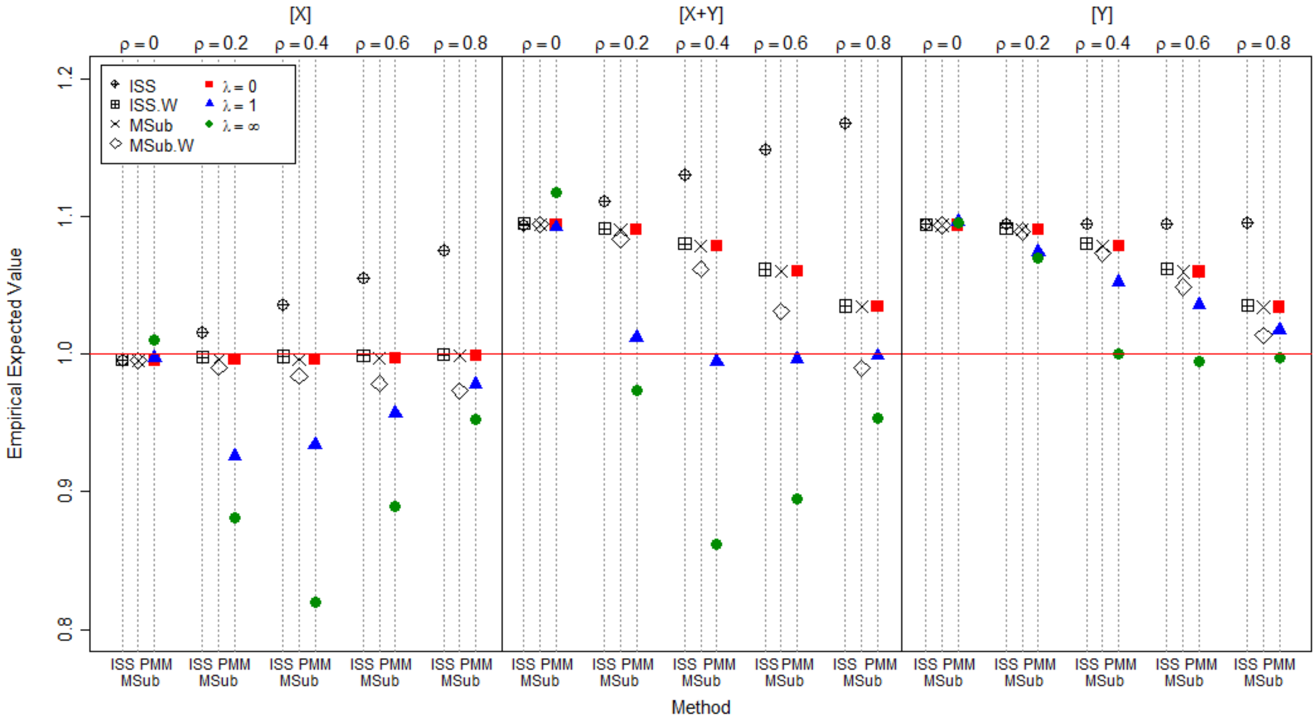


Figure 5.2: Empirical expected values of population mean estimates over 5000 simulation replications with a 75% response rate. Red horizontal line denotes the true population mean.

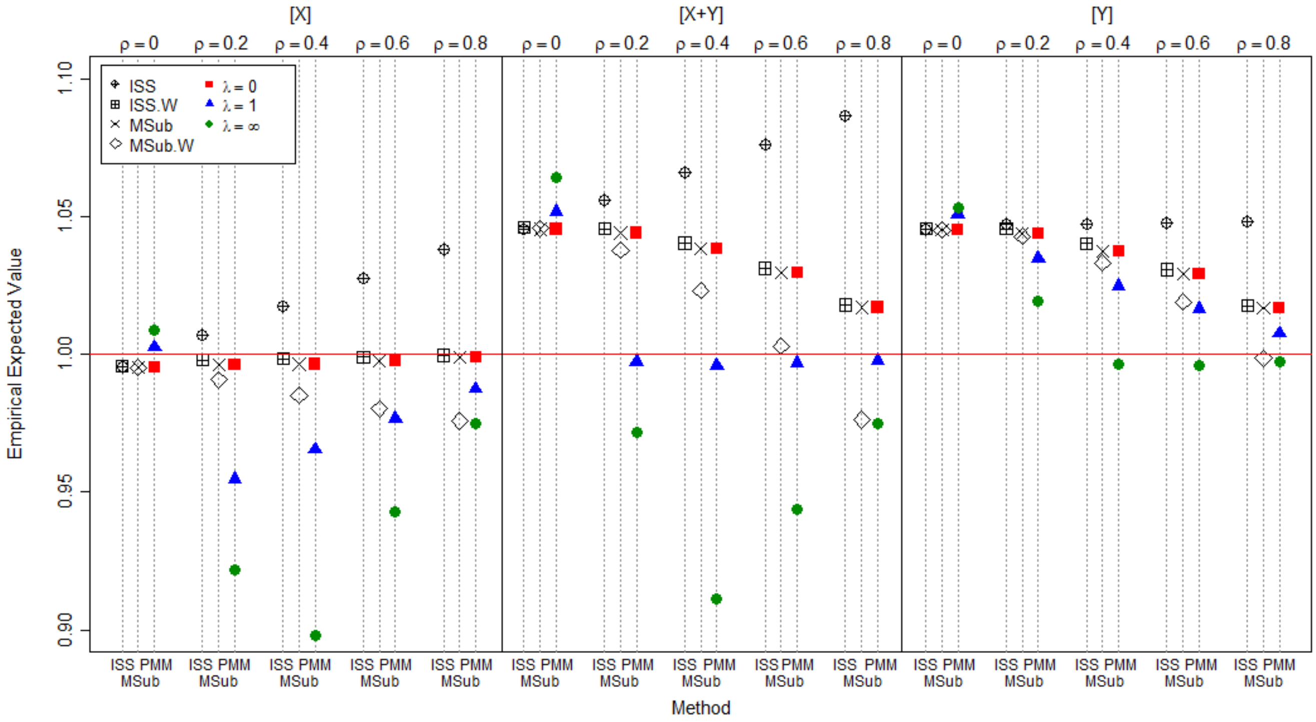


Figure 5.3: Empirical sampling variances of population mean estimates over 5000 simulation replications with a 50% response rate. Red horizontal line denotes sampling variance under complete response.

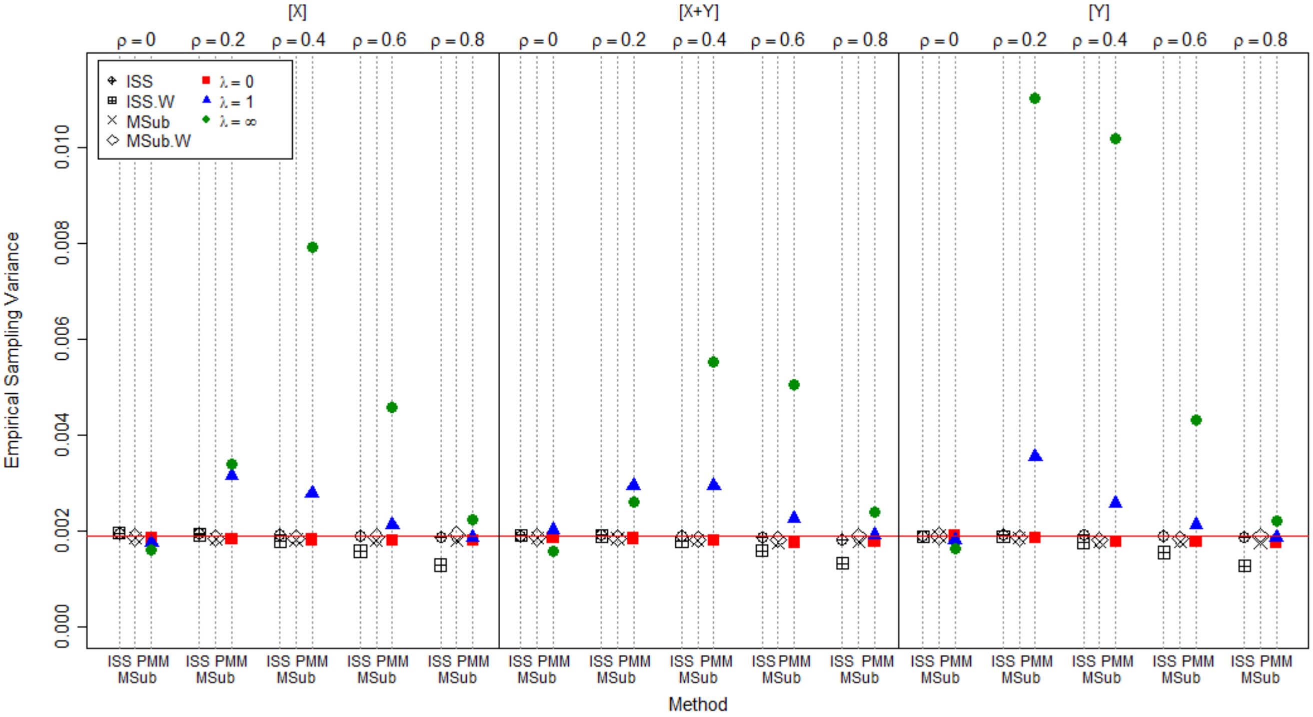


Figure 5.4: Empirical sampling variances of population mean estimates over 5000 simulation replications with a 75% response rate. Red horizontal line denotes sampling variance under complete response.

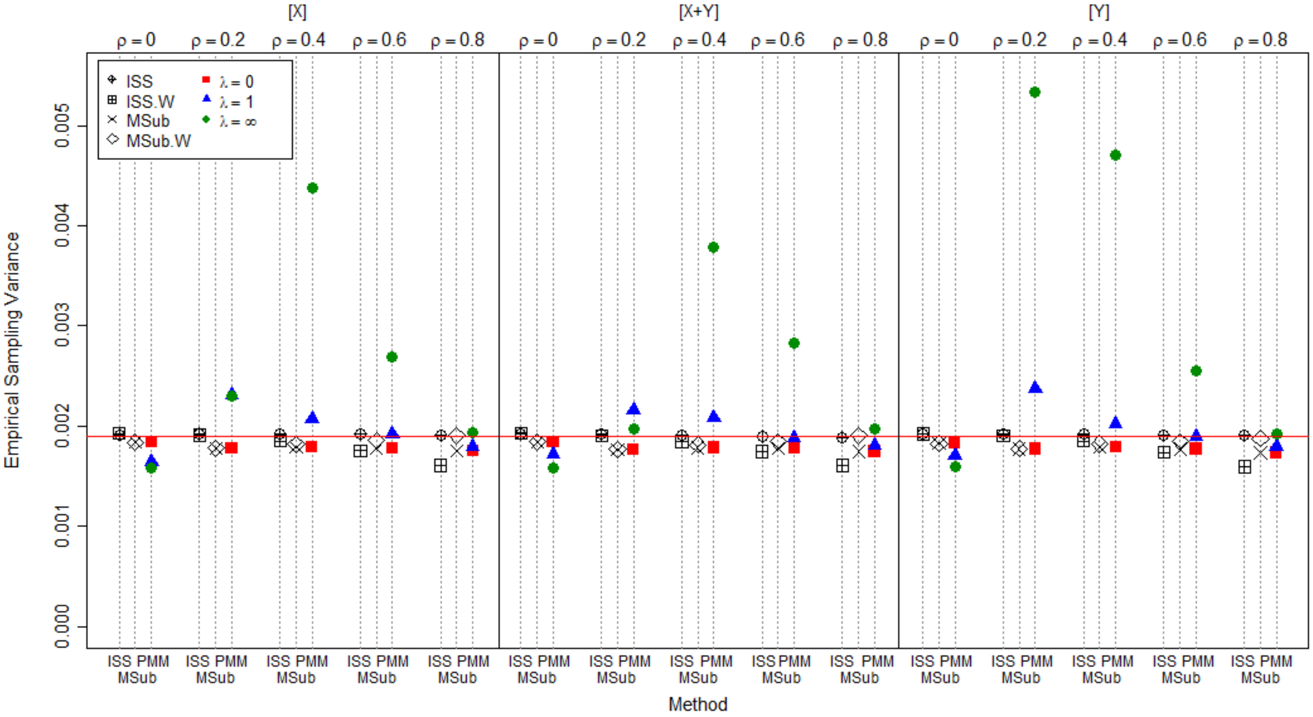


Figure 5.5: Empirical root mean square errors of population mean estimates over 5000 simulation replications with a 50% response rate. Red horizontal line denotes root mean square error under complete response.

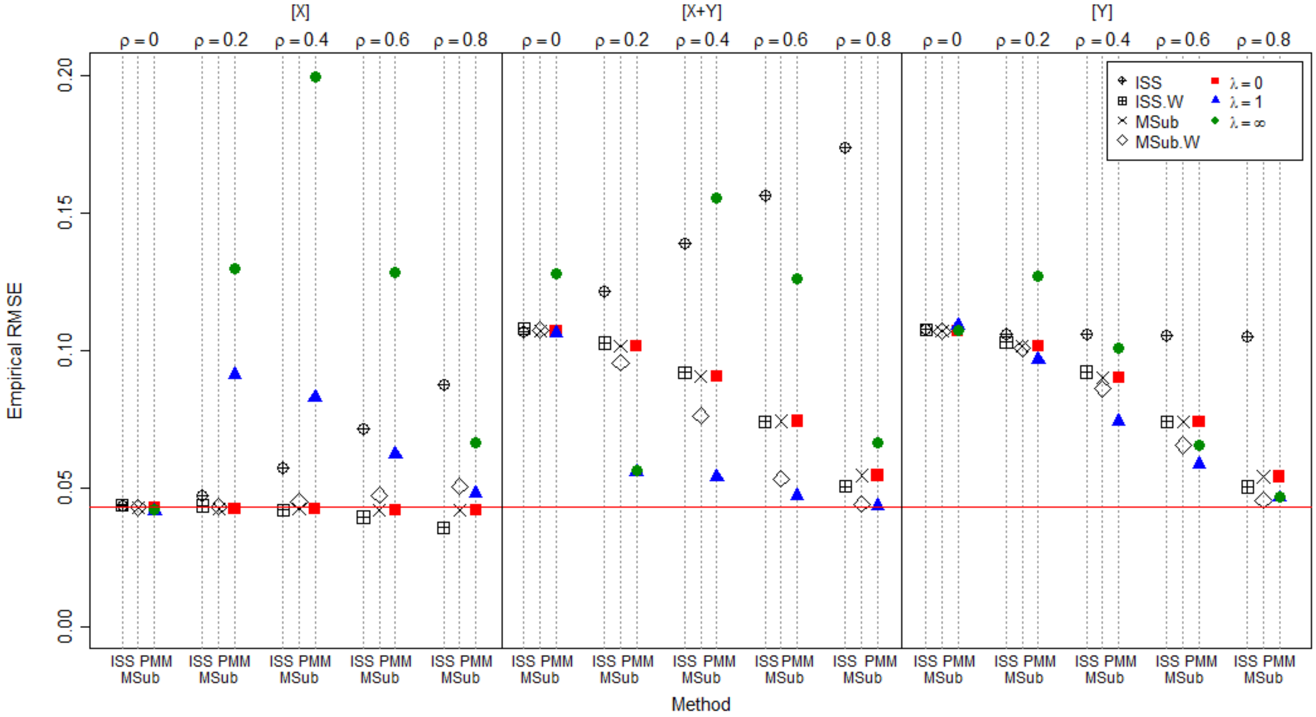
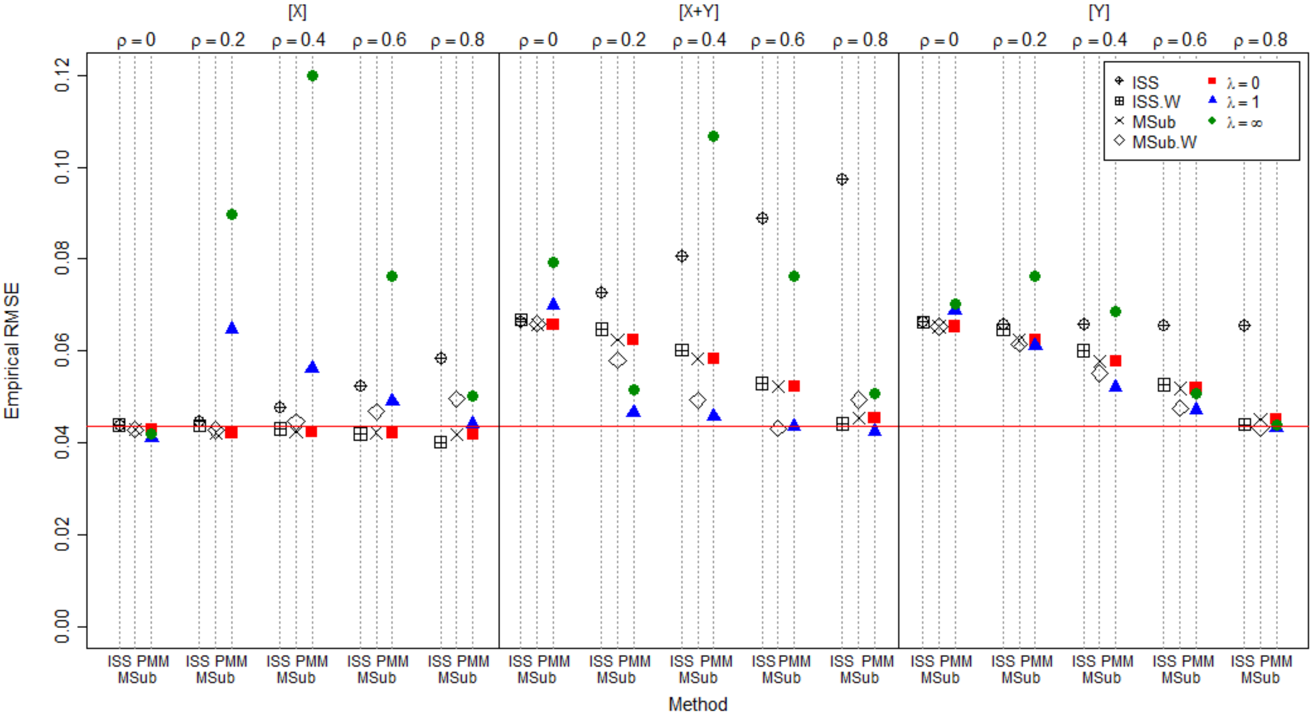


Figure 5.6: Empirical root mean square errors of population mean estimates over 5000 simulation replications with a 75% response rate. Red horizontal line denotes root mean square error under complete response.



Missingness model [X+Y]

When nonresponse depends on both the auxiliary variable X and the survey variable Y, it is expected that a pattern-mixture model with $\lambda = 1$ would produce unbiased estimates for the population mean, whereas other approaches that do not take into account the nonignorability feature of this mechanism would perform poorly. As Figures 5.1 and 5.2 show, this is true across all correlations, except $\rho_{yx} = 0$, where all methods led to equally biased estimates, illustrating again the importance of having good predictors of the survey outcome for nonresponse adjustments. For all the other correlations, the PMM substitution with $\lambda = 1$ produces essentially unbiased estimates, while the estimates of the other approaches have substantial bias. Although the PMM substitution method with $\lambda = \infty$ does account for a nonignorable nonresponse, it employs a misspecified model in which the nonignorability is much stronger than actually it is. Thus the mean under this PMM substitution substantially underestimates the true population mean when the correlation between X and Y is not zero.

Figures 5.3 and 5.4 show that the sampling variances of estimates under the $\left[X + Y\right]$ missingness model were quite similar to those observed under the MAR mechanism. That is, the ISS, ISS.W, MSub, and MSub.W methods lead to estimates with variability similar to the complete response case. The PMM substitution methods for $\lambda = 1$ and $\lambda = \infty$ provide estimates with slightly larger sampling variance for intermediate levels of correlation. This is evidence that the missing mechanism does not have much impact on the overall behavior of sampling variability, other than their magnitudes essentially dictated by the response rate.

Overall, for the $\left[X + Y\right]$ missingness model, the PMM substitution assuming $\lambda = 1$ was the method that led to the smallest RMSE (Figures 5.5 and 5.6). This is particularly true for the intermediate correlations ($\rho_{yx} = 0.2, 0.4, 0.6$). For the zero and highest level of correlation, all tested methods led to estimates with similar levels of error, with the PMM substitution assuming $\lambda = \infty$ giving slightly worse results. In fact, for the $\left[X + Y\right]$ missingness model, assuming such a strong nonignorable nonresponse produced estimates with higher RMSE than when it was assumed that the missing mechanism was ignorable ($\lambda = 0$).

Missingness model [Y]

The nonresponse mechanism induced by this model corresponds to the extreme case in which missingness depends solely on the survey outcome itself. It can generally pose a difficult challenge for standard nonresponse adjustments, since the auxiliary variables usually used in these types of methods are not directly related to nonresponse.

As can be noted in Figures 5.1 and 5.2, standard approaches such as ISS.W, MSub, and MSub.W produce estimates with substantial bias across all levels of correlation. The PMM substitution assuming $\lambda = 0$ and $\lambda = 1$ also leads to biased estimates for the $\left[Y\right]$ missingness model. The correctly specified model in this case would assume $\lambda = \infty$, yet, only for moderate to high correlations ($\rho_{yx} = 0.4,\ 0.6, 0.8$) the PMM substitution under $\lambda = \infty$ provides unbiased estimates for the population mean. For small correlations, this method performs just as well as the others in terms of bias, with a slight advantage when $\rho_{yx} = 0.2$. However, with the exception of ISS, all tested methods reduce bias as the correlation between X and Y increases, reinforcing the importance of using good predictors for nonresponse adjustments, regardless of the missing mechanism.

Despite being the most appropriate model for this missing mechanism and leading to unbiased estimates, setting λ to infinity in the PMM substitution method in this case produces the least stable estimates, just as in the other missingness models (Figures 5.3 and 5.4). Although the general pattern in the variability of the estimates across the methods is essentially the same as the other two cases previously analyzed, the difference in the variability of the estimates of the PMM substitution method with $\lambda = \infty$ and the estimates of the other approaches is much larger, especially for the moderate levels of correlation ($\rho_{yx} = 0.2, 0.4, 0.6$).

The variance inflation of PMM substitution with $\lambda = \infty$ is so large that it cancels the bias reductions obtained by this method. The RMSE of its estimates are not the smallest for any of the correlations analyzed in this study. In fact, for two of the correlations ($\rho_{yx} = 0.2$ and $\rho_{yx} = 0.4$), PMM substitution setting $\lambda = \infty$ presented the largest RMSE (Figures 5.5 and 5.6). Overall, PMM substitution with $\lambda = 1$ performs slightly better across most levels of correlation under this

MNAR nonresponse mechanism. However, this is mostly due to the sampling variability of the estimates, since this approach also produces substantially biased estimates.

5.6 Discussion

PMM substitution performed well when λ corresponds to the underlying missing mechanism. Not only was it the only method that led to unbiased estimates across most missing mechanisms and correlations, but it also gave the most accurate estimates for almost all scenarios. The only exceptions, as described above, occurred when (1) there was no association between the survey variable Y and the auxiliary covariate X, and (2) for the missingness model $\left[Y\right]$.

Exception (1) , in general, is challenging for any nonresponse adjustment given that having good predictors of the outcome variable is a key factor for reducing nonresponse bias and sampling variance (Little and Vartivarian, 2005). Exception (2), on the other hand, presents an interesting bias/variance trade-off, in which adopting the appropriate model (i.e., using $\lambda = \infty$) leads to unbiased estimates with large variability, whereas using a model that does not reflect exactly the true missing mechanism (setting λ to one when missingness model is $\left[Y\right]$) generates biased estimates, but with smaller variances. When taking both bias and variance into account, the latter approach does give more accurate estimates. However, since ultimately the interest here is to identify and minimize nonresponse bias, using the appropriate model at the cost of less stable estimates would usually be preferred over the alternative. Moreover, as will be discussed, the PMM substitution approach can be implemented for a sensitivity analysis using different values of λ, each of which produces estimates that together portray the impact of nonresponse in a more complete manner.

Although substitution of nonrespondents is a commonly used approach to mitigate unit nonresponse in many surveys, it has been largely neglected by the survey statistics and methodology literature, with few investigations to understand and improve the method. All substitution methods available until now assume a MCAR or MAR mechanism. Nonignorable nonresponse is problem that has never been directly tackled by any of these substitution methods. Although Rubin and Zanutto (2002) did provide an imputation method that uses substitutes for one type of nonignorable nonresponse, their approach only applies when the variables that cause the nonresponse mechanism to be ignorable are indirectly observed, making it closer in reality to an ignorable nonresponse problem. Moreover, when selecting substitutes for nonrespondents, substitution methods do not take into account the survey variables observed for the respondents, a valuable approach for minimizing nonresponse bias when missingness is not ignorable

Pattern-mixture models have been suggested and used to analyze MNAR data when fully observed auxiliary data are available. The applications of such models so far, however, have been restricted to the data analysis, mostly for sensitivity analysis. This paper presented a new application of pattern-mixture modeling, applying it to assist the selection of substitutes for replacing nonrespondents. By doing so, this method incorporates a wider variety of missingness assumptions, ranging from a MAR to different degrees of MNAR mechanisms, into the sample selection process. This enables the possibility of performing sensitivity analysis using real additional data, as opposed to predicted values under a model or values already observed in the sample (such as from hot-deck donors), by selecting substitutes under different assumptions about the missing mechanism.

Another feature of the proposed PMM substitution method not present in some standard nonresponse adjustments, such as weighting or the standard substitution methods, is that it also takes into account the information of both the auxiliary variables -- assumed to be available for every unit in the population, such as frame variables -- and the survey outcomes, available from the respondents. This is particularly important since nonresponse error is variable and statistic dependent, and therefore its adjustments should consider the information on the outcome variables and their relationship with the auxiliary covariates.

The simulation results showed that the proposed method tends to eliminate nonresponse bias when the missing mechanism matches the value that the λ parameter is set to on the pattern-mixture model and the correlation between the survey and auxiliary variables is at least moderate. Moreover, when these conditions hold, PMM substitution presented the smallest RMSE, with the exception of the $[Y]$ missing model, in which assuming $\lambda = 1$ led to more accurate estimates than using $\lambda = \infty$, the correct value for λ under the nonresponse model. This puzzling finding is explained by the high instability of the estimates when using a pattern-mixture model with $\lambda = \infty$. Since a primary objective of this method is to detect bias through sensitivity analysis, such variance inflation is not a primary concern, although it should be considered when choosing an approach for nonresponse adjustments.

In practice, the value of λ that matches the nonresponse mechanism is rarely known and the respondent data do not provide any information about this parameter. Therefore, it is not possible to choose one single value for λ to eliminate nonresponse bias. However, expert knowledge on the substantive variables and their interaction with nonresponse may provide guidance on the nature of the missing mechanism, allowing researchers to make educated guesses about the values of λ more suitable to nonresponse adjustment. Moreover, as suggested above, PMM substitution method can be used as a tool to detect potential nonresponse problems through a sensitivity analysis. To do so, previous studies have suggested using different values for λ to compute the population estimates and evaluate how much they change according to each value. Andridge and Little (2011) and Sullivan and Andridge (2015), for example, recommended using $\lambda = \{0,1,\infty\}$.

However, implementing this sensitivity analysis in PMM substitution may pose an operational challenge, since different values of λ can lead to different substitutes for a given nonrespondent. One could select multiple substitutes for each nonrespondent, one for each value of λ, but this would impact survey costs. This would not be a problem if the substitutes selected for the nonrespondents are all the same for different values of λ. But in this case the sensitivity analysis should not provide any additional insights, given that estimates would be roughly equal if the same cases are selected as substitutes over different missingness assumptions. The ability to perform such sensitivity analysis in the PMM substitution and the costs associated with it raises a trade-off between nonresponse bias and survey costs. There might be a point that the PMM substitution could lead to reductions in the RMSE large enough to justify the added costs of selecting multiple substitutes per nonrespondent. While this type of investigation is beyond the scope of this study, it merits further investigations, which is left for future research.

One approach to using PMM substitution for sensitivity investigation is to select sub-samples of nonrespondents using different values for λ parameter for each sub-sample. For example, for $\lambda = \{0,1,\infty\}$ nonrespondents would be randomly allocated, controlling for the auxiliary variables, to three sub-samples corresponding to each value of λ. Once substitutes for each sub-sample are collected, estimates of the population parameters would be computed separately for each sub-sample, but also using data from the original respondents. If the correlation between the survey variable and the auxiliary covariate is at least moderate, large differences across the estimates of the sub-samples might be an indication of nonresponse bias. In this case, a more substantive understanding of the relationship between the survey variable and the missing mechanism would be necessary to decide which estimate is more plausible. On the other hand, small differences between the estimates would indicate that there may not be problems of nonresponse bias with that survey variable, unless the correlations between the survey outcomes and the auxiliary variables are low. When that is the case, there is not much that sensitivity analysis, or in fact any nonresponse adjustment, can do to assess nonresponse bias. This reinforces the important role of strong predictors of the survey variables in such adjustments.

Another operational challenge of the PMM substitution method is associated with the requirement of having to have respondent data before the selection of the substitutes. This may prove operationally inconvenient, as it would make prompt action for nonresponse through substitution at early stages of the fieldwork impossible, potentially extending the data collection period. On the other hand, using respondent data with pattern-mixture models enables substitute selection to be performed in a more informed fashion, taking all of the available information up to that moment into account in this process, as mentioned before. Also, waiting a period of time during data collection before selecting substitutes would allow more time for extra efforts to get the cooperation of late respondents, that otherwise might be prematurely substituted, a concern raised by Vehovar (1994) and Chapman and Roman (1985a, 1985b). Moreover, other than a standard application of substitution, this could be implemented as an alternative to deal with persistent nonrespondents or refusals after a nonresponse follow-up, for example, or to handle attrition in longitudinal surveys, in which there are data for the survey outcomes from previous waves.

This paper illustrated the use of the proposed method using one single auxiliary variable. As described previously, if more than one covariate is available, this same procedure can be employed, but using a “proxy” auxiliary variable which can combine the auxiliary covariates through a principal component analysis or linear predictors, as suggested by Andridge and Little (2011). Also, the PMM substitution method proposed here initially assumed only one normally distributed survey outcome. Surveys contain multiple outcome variables of different types, each of which may have a different relationship with the auxiliary variables and the missing mechanism. An advantage of PMM substitution is that unlike other substitution and some nonresponse adjustments, it takes this variable-statistic dependent nature of nonresponse bias into account. It also means that, if implemented variable-by-variable, some of the nonrespondents might be assigned for different substitutes because for each survey outcome, even under the same value of the λ parameter. Clearly applying PMM substitution for each variable separately would not be feasible in practice. To accommodate multiple survey variables, practitioners may compute the predicted values under the pattern-mixture model for each variable separately and then use a multidimensional distance measure, such as a Euclidean or Mahalanobis distance, to select the substitutes. While this will probably not be the optimum for any combination of the individual survey outcomes, it might be an acceptable compromise to detect patterns of nonresponse bias across all or most of the survey variables. Moreover, extensions for other types of survey outcomes, such as binary variables, for example, could be developed, as has been done for the proxy-pattern mixture models (Andridge and Little, 2009) and the PPM hot-deck (Sullivan, 2014).

For simplicity, in this study, the PMM substitution was proposed and evaluated through simulation studies under a simple random sample design. This method, however, can be extended to more complex sample designs, involving stratification, clustering and unequal selection probabilities. Also, it was assumed that the same missing mechanism operates over original selections and substitutes. This may not hold true in many applications. For instance, it is often the case that there is less time to spend to obtain the responses of the substitutes than of the originally selected units. Even if all the other survey protocols are the same, the substitutes will likely have a smaller response propensity than the original units and, therefore, a different nonresponse mechanism. Therefore, more investigations extending the results of the simulations of this study to more general nonresponse mechanisms are needed. Such extensions on the sample design and missing mechanism assumptions should be developed in future studies.

A challenging problem for PMM substitution, and for substitution procedures in general, is variance estimation. While under an MAR mechanism this can be accomplished using standard variance estimation techniques, as it would under complete response, there has not been almost no research on how to estimate sampling variance using substitute data when the missing mechanism is nonignorable. The approach of Rubin and Zanutto (2002) uses substitutes to multiply impute the nonrespondent missing data and then, using Rubin's combining rule, estimate sampling variance using multiple imputation. Since substitutes can be selected under a nonignorable missing mechanism using PPM substitution, multiple imputation would account for adjustments in terms of differences between respondents and nonrespondents due to the MNAR nonresponse mechanism. The properties of sampling variances using these or any other approach should be addressed in future research.

Finally, it could be argued that the same results obtained by the PMM substitution can be achieved using the PPM hot-deck imputation method proposed by Sullivan and Andridge (2014) with no additional costs attributed to the substitutes. Although this statement may generally be true, the proposed substitution procedure might be preferred over PPM hot-deck when there are not enough potential donors on the covariate space related to missingness, which is particularly important in the case of nonignorable missing data. With substitution, the pool of "donors", or substitutes in this case, is much larger, given that the sampling fraction is small in most applications, even within strata, and therefore the number of unsampled units, $N-n$, is very large.

References

Andridge, R. R., Little, R. J. A. (2009). Extensions of proxy pattern-mixture analysis for survey nonresponse. *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 2468–2482.

Andridge, R. R. and Little, R. J. (2011). Proxy Pattern-Mixture for Survey Nonresponse. *Journal of Official Statistics*, Vol. 27, No. 2, pp. 153-180.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*.John Wiley & Sons, Inc., Hoboken, New Jersey

Chapman, D. W. and Roman, A. M. (1985a). Appendix 6 (Substitution). In *Results of the 1984 NHIS/RDD Feasibility Study: Final Report*, internal U.S. Bureau of Census report, February.

Chapman, D. W. and Roman, A. M. (1985b). An investigation of substitution for an RDD survey. *Proceedings of the Survey Research Methodology Section*, ASA, pp. 269-274.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley & Sons.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly*, 69, pp. 87-98.

De Leeuw, E. and De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D Dillman, J. Eltinge, and R. Little (eds.) *Survey Nonresponse*, pp. 41-54. New York: Wiley.

Groves, R. M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72 (2), pp. 167-189.

Groves, R. M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009).*Survey Methodology*. Hoboken, NJ: John Wiley and Sons.

Keeter, S., Miller, C., Kohut, A., Groves, R. M. and Presser, S. (2000). Consequences of Reducing Nonresponse in a Large National Telephone Survey. *Public Opinion Quarterly,* 64, pp. 125-48

Keeter, S., Kennedy, C., Dimock, M., Best, J. and Craighill, P. (2006). Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey. *Public Opinion Quarterly*, 70, pp. 759-779

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Lessler, J. T. and Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421), 125-134.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3), 471-483.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.

Little, R. J., and Vartivarian, S. L. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*. 31, pp. 161-168.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Merkle, D. M. and Edelman, M. (2002). Nonresponse in Exit Polls: A Comprehensive Analysis. In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, pp. 243-58. New York: Wiley.

Rand, M. (2006). Telescoping Effects and Survey Nonresponse in the National Crime Victimization Survey. Paper presented at the Joint UNECE-UNODC Meeting on Crime Statistics. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.14/2006/wp.4.e.pdf (accessed on March 21st 2014)

Rubin, D. B. (1987). *Multiple Imputation for Survey Nonresponse*. New York: John Wiley and Sons.

Rubin, D. B., and Zanutto, E. (2002). Using Matched Substitute to Adjust for Nonignorable Non response through Multiple Imputation. In *Survey Nonresponse*, edited by R. Groves, R. J. A. Little, and J. Eltinge. New York: John Wiley, pp. 389-402.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Siddique, J. and Belin, T. R. (2008). Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Computational statistics & data analysis*, 53(2), 405-415.

Sullivan, D. (2014). A Hot Deck Imputation Procedure for Multiply Imputing Nonignorable Missing Data: The Proxy Pattern-Mixture Hot Deck. Doctorate thesis submitted for the graduate faculty of The Ohio University, May, 1998.

Sullivan, D. and Andridge, R. (2015). A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. *Computational Statistics & Data Analysis*, 82, 173-185.

Vehovar, V. (1994). Field substitution – a neglected option? *Proceedings of the Survey Methods Section*, ASA, pp. 589–94.

Vehovar, V. (1999). Field Substitution and Unit Nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, pp. 335-350

CHAPTER VI

Discussion

Despite its widespread use in survey practice, substitution remains an understudied topic in the survey statistics and methodology literatures, and is frequently viewed with skepticism in these fields. This dissertation sought to investigate the characteristics of this method and develop new approaches to improve the estimates obtained with its use. The results presented in Chapters III to V have important practical implications for surveys that use substitution as a solution to unit nonresponse. In this concluding chapter, these results are synthesized along with their practical implications and directions for future research.

6.1 Summary of Study Results

Substitution is frequently used in establishment surveys as a remedy for unit nonresponse, particularly in school-based surveys where students are the target population. In many of these cases, nonresponse occurs at the cluster-level, such as schools in the example given above. The few existing studies on the use of substitution examine only element-level nonresponse. Moreover, there has not been much research done about nonresponse at the cluster-level in multi-stage cluster samples. The study conducted in Chapter III investigated both of these problems. More specifically, it studied the problem of nonresponding Primary Sampling Units (PSU) in probability-based two-stage cluster samples.

The derivation of the nonresponse bias due to PSU nonresponse of an unadjusted respondent mean in a two-stage equal-sized cluster sample provided a new and interesting result: such bias is an increasing monotonic function of the intra-cluster correlation coefficient, as shown by the expression below:

$$Bias\left(\bar{y}_r\right) \doteq \frac{1}{\bar{p}} Cov_a\left(Y, p\right) = \frac{1}{\bar{p}} Corr_a\left(Y, p\right) \sigma_a\left(p\right) \sqrt{\frac{\sigma^2\left(Y\right)}{B}\left[1+\rho\left(B-1\right)\right]},$$

where $Corr_a\left(Y,p\right)$ is the correlation between the survey variable, Y, and the response propensity, p; $\sigma_a\left(Y\right)$ and $\sigma_a\left(p\right)$ are the standard deviations of the survey variable and response propensity, respectively; ρ is the intra-cluster correlation of the survey variable (the subscript a denotes that these statistics are being evaluated at the cluster level), and B is the cluster size. This means that survey variables with larger intra-cluster homogeneity are more susceptible to cluster nonresponse bias. A similar result was found for unequal-sized cluster populations.

These results illustrate the importance of examining cluster nonresponse and its difference comparatively to element nonresponse. Therefore, further investigations on methods dealing with this problem are needed. The simulation studies conducted in Chapter III evaluated a few of such methods. Particularly, these studies investigated the properties of substitution and nonresponse weighting methods for the estimation of a finite population mean.

An important general result was that the performance of the substitution methods depends on how it is implemented; the results demonstrated that the selection of the substitutes is one of the most important factors. If substitutes are randomly selected, its properties are similar to the naïve unadjusted respondent mean. This is true even if the substitutes are randomly selected within the same stratum of their corresponding nonrespondents – a procedure known as stratified random substitution (Lynn, 2004). On the other hand, if substitutes are selected by a matching procedure – that is, the unit with the smallest distance for a given set of measures to the nonrespondent is selected as its substitute – the performance of the substitution methods is almost equivalent to a standard nonresponse propensity weighting adjustment (Cassel, Särndal and Wretman, 1983; Groves et al, 2002), as long as the variables used in the matching procedure are the same as the ones used in the estimation of the nonresponse propensity.

The intuition behind these results is that by randomly selecting a unit to substitute a nonrespondent, on average, the substitutes will be more similar to the responding units in the sample than the nonrespondents. This will not necessarily decrease the level of bias, just as an unadjusted respondent mean with an inflated (by an unexpected response rate) sample size will not lead to a bias reduction. On the other hand, if substitutes are matched to nonrespondents, then they are expected to be more like them and, to the extent that the matching covariates are related to the survey variables, there will be some bias reduction, as is seen when nonresponse weighting adjustments are used. Also, as expected from results by Little and Vartivarian (2005), the stronger the association between the matching and survey variables, the larger the bias reductions are in both matching substitution and nonresponse weighting adjustment methods. This is also an important result because it shows that any criticism about such substitution methods in terms of the statistical properties should also be directed toward other commonly used techniques, such as nonresponse weighting.

A second factor that impacts the performance of the substitution methods is the number of substitution iterations used. This number represents the number of times a substitution is attempted to be made (before a responding unit is finally selected or the substitution procedure is terminated) if a selected substitute for an original nonresponding unit turns out to be a nonrespondent as well. This is an overlooked featured in previous studies of substitution, but it is directly related to the degree of this method's success. As would be expected, as the number of substitution iterations increases, the more likely all nonrespondents will be successfully substituted, which consequently results in a reduction of nonresponse bias. Obviously, as the number of substitution iterations increases, the number of responding units also increases and, therefore, the sampling variance decreases. Hence, increasing the number of substitution iterations has positive impacts on survey estimates both in terms of bias and sampling variability.

Another important contribution from the study in Chapter III is related to sampling variance estimation. Many surveys use as many strata as possible while still enabling design-based sampling variance estimation, which is known as deep stratification (Kish, 1965). In stratified multi-stage samples, this technique tends to create to designs with two PSUs per stratum. In such cases, it is very common that, after nonresponse, some strata end up with only one or no responding PSUs, thus prohibiting design-based sampling variance estimation. Vehovar (1999) pointed out that one of the potential advantages of using substitution is its avoidance of such situations. However, he also mentioned that a comparison with strata collapsing, a commonly used technique to deal with this problem, would be needed to evaluate the real efficiency of the substitution method for that purpose. This comparison was conducted in the simulation studies of Chapter III. In general, the sampling variance estimates of the substitution methods were less biased and more accurate than those of the strata collapsing technique.

Therefore, the general results of Chapter III show that substitution is a valid alternative for dealing with PSU nonresponse, as long as a matching procedure is implemented and the substitution is carried out a sufficient number of iterations to ensure the substitution of as many nonrespondents as possible. Another necessary condition for the successful use of substitution is that the variables used in the matching procedure are correlated with the survey variables. When these conditions are met, substitution provides the same levels of bias reduction as other standard methods, such as nonresponse weighting adjustments. Further, it can produce less biased and more accurate sampling variance estimates, using a standard Taylor Series approximation method, compared to strata collapsing, when the sample (without substitution) turns out to have strata with no or one PSU after nonresponse.

In some instances, however, it is not possible to match nonrespondents and substitutes on some important variables that can explain the survey outcomes, either because they are not readily available for every unit in the population or because they can only be measured during data collection, such as paradata. Therefore, there might be some systematic differences between nonrespondents and their corresponding substitutes that are not taken into account in the matching procedure. These differences, consequently, might diminish the potential bias reductions generally provided by the substitution procedure. With that in mind, Rubin and Zanutto (2002) proposed a method to take these differences into account by modeling them and multiply imputing the nonrespondents using a procedure they named Matching, Modeling and Multiple Imputation (MMM).

Rubin and Zanutto’s method succeeded in taking into account differences between nonrespondents and their substitutes on auxiliary covariates (modeling or calibration covariates) observed only for these two subsets and, consequently decreasing nonresponse bias. However, it also translated into larger survey costs, due to the need to also select substitutes for a sub-sample of the respondents to estimate the imputation model, and an increase in the sampling variances of the survey estimates, due to the imputation procedure. To overcome these two problems, Chapter IV presented a modification to Rubin and Zanutto's method, as well as a new method to adjust for differences between nonrespondents and substitutes using a calibration procedure.

In the modified version of Rubin and Zanutto's MMM method proposed in Chapter IV, instead of selecting substitutes for a sub-sample of respondents from the unsampled population to estimate the imputation model, they were selected from among the remaining set of respondents, in a manner similar to a hot-deck imputation procedure. By doing so, it is no longer necessary to select additional units into the sample, thus avoiding an increase in the survey costs, and at the same time still producing estimates with the same bias reduction levels observed in the original method. A disadvantage, however, of this variant of the MMM method is that it can further increase the sampling variance, as, contrary to Rubin and Zanutto's procedure, there is no new information coming into the sample. A set of simulation studies confirmed these results: while the levels of bias reduction of the alternative version of MMM were equivalent to its original version, it also produced slightly less precise estimates. However, in some situations, such losses in precision were not observed, particularly when the modeling covariates are more strongly correlated to the survey variables than the covariates used in the matching procedure.

The new method introduced in Chapter IV proposed using a calibration weighting procedure (Deville and Särndal, 1992) to take into account differences between nonrespondents and their substitutes. This method rests on the calibration of the substitutes to the nonrespondents in terms of the modeling (or calibration) covariates. That is, to create a new set of weights that make the sample total of the substitutes on these covariates to match the corresponding sample totals of the nonrespondents. Similar to the modified version of Rubin and Zanutto's MMM method, this calibrated matching substitution procedure does not require the selection of additional substitutes for a sub-sample of the respondents, and therefore, does not produce an increase in the survey costs aside from the selection of the substitutes for the nonrespondents. Moreover, because this adjustment relies on a calibration procedure, instead of data imputation, it is expected that the estimates produced by this methods would be more precise than Rubin and Zanutto's MMM (either the original or modified version).

The results of the simulation studies conducted in Chapter IV consistently confirm this hypothesis. In some cases, the bias reductions of the calibrated matching substitution method are not as high as the ones found in the MMM methods. This is due to the fact that the adjustments conducted in the calibration procedure are performed at the aggregate-level (i.e., sample totals), whereas in the imputation procedure they are done at the element-level, making a finer adjustment. Such differences, however, were not substantial in most of the cases in the simulation studies.

These results show that there are multiple methods to improve the quality of survey estimates using substitution procedures to deal with unit nonresponse, provided certain conditions are met. First, the covariates used both in the matching and modeling/calibration procedures must be associated with the survey outcomes. Second, and possibly most important, the data should be Missing At Random (MAR), as in many of other nonresponse adjustment methods (Little and Rubin, 2002). This means that the missing mechanism should depend only on fully observed variables, such as the matching and modeling/calibration covariates. However, in many applications, it is important to evaluate what the consequences would be in case the missing data is non-ignorable.

While there have been a few methods proposed in the literature to analyze Missing Not At Random (MNAR) data, there has not been so far much research done using substitution methods in this area. Although Rubin and Zanutto (2002) developed a method that can address some forms of non-ignorable nonresponse, a substitution procedure that can handle a more general form of non-ignorability still lacks in the survey sampling literature. For this reason, Chapter V proposed the use of Pattern-Mixture Models (PMM) to assist in the selection of substitutes. The idea was motivated by the Proxy Pattern-Mixture Hot Deck imputation method, developed by Sullivan and Andridge (2015), in which they use a normal PMM (Little, 1994) to predict the outcomes of respondents and nonrespondents, and then, based on a distance measure of these predicted values and under an assumed missing mechanism, they find responding donors to impute the missing data of the nonrespondents.

The PMM substitution method proposed in Chapter V follows a similar structure. First, a PMM is fitted and used to predict the survey outcome for nonrespondents and unsampled units in the population, using auxiliary information available for every unit, such as frame data, under an assumed missing mechanism (designated by the λ parameter in the PMM). Then, for each nonrespondent, substitutes are selected from the unsampled population based on a distance measure on the predicted survey outcomes. This method has the advantage of offering a more flexible way to select substitutes, without having to rely on a MAR assumption. On the other hand, such selections are made under some assumptions about the missing mechanism. Results from a simulation study conducted in Chapter V showed that if the missing mechanism is correctly specified, the PMM substitution lead to the less biased estimates compared to a nonresponse weighting approach or a standard matching substitution method. However, if the missingness model is incorrectly specified, this method does not perform very well. Moreover, under a strong non-ignorable missing mechanism (i.e., when $\lambda = \infty$), the estimates produced by this proposed methods are quite unstable, as previous research of other estimation techniques that also use PMM have shown.

Since the true missing mechanism in most, if not all, practical cases is unknown and the observed data do not provide any information about the λ parameter in the PMM, the use of the PMM substitution using a single value for this parameter may prove difficult. However, as suggested in other applications (Little, 1994; Andridge and Little, 2011,Sullivan and Andridge, 2015), the PMM can be used for sensitivity analysis, to evaluate how the estimates change according to different missing mechanism assumptions. For instance, Andridge and Little (2011) suggested using $\lambda = \{0,1,\infty\}$ for that purpose, since these values can portray a wide range of possible missing mechanisms. The data are assumed to be MAR if $\lambda = 0$. When $\lambda = 1$, the model suggests the (unobserved) survey variable and the (observed) auxiliary variables have the same weight in explaining the missing mechanism. Assuming $\lambda = \infty$ is the most extreme case of non-ignorable nonresponse, in which the missingness depends solely on the (unobserved) survey variable.

For the PMM substitution, this sensitivity analysis may be performed by selecting multiple substitutes for the nonrespondents, in which each of them would be selected under a given value of the λ parameter in the PMM. Different survey estimates would be then obtained using each set of substitutes. Unfortunately, this approach would lead to a substantial increase in survey costs if all nonrespondent were substituted multiple times. A more affordable alternative would be to select a sub-sample of the nonrespondents for each value assigned for the λ parameter. For instance, if $\lambda = \{0,1,\infty\}$, the nonrespondents would be randomly partitioned into three balanced sub-samples in terms of the auxiliary variables and for each sub-sample a value for the λ parameter would be used. With this approach it would be possible to perform sensitivity analysis computing different estimates using each set of substitutes separately, while keeping the survey costs similar to a standard substitution method.

Like other survey sampling techniques, the PMM substitution method was developed using only one single survey variable. However, this methodology can be readily adapted to a multivariate setting by predicting each of the survey variables separately using the PMM and then selecting the substitutes using a multidimensional distance measure, such as a Mahalanobis distance. Also, if more than one auxiliary covariate is available for every unit in the population, they can be summarized in a single “proxy” auxiliary variable through principal component analysis or linear predictors, as suggested by Andridge and Little (2011), for example.

In summary, the studies performed in this dissertation show that substitution can be a useful method to address unit nonresponse, particularly for cluster nonresponse and when there are auxiliary information available for all units in the population that may not be viable to use in statistical modeling, due to high-dimensionality issues, for instance. The methods proposed here still deserve more theoretical and empirical investigation and they can be further developed. The next section describes some problems in substitution that should be addressed in future research.

6.2 Future Research

Substitution is a much neglected topic in the survey sampling and methodology literature, but hopefully the studies performed in this dissertation will motivate further research in this area. There are still many areas in need for further research and development. Some of them are described here.

Previous studies on substitution investigated its use when nonresponse occurs at the element-level. However, as pointed out in Chapter III, many of the applications of substitution are done when there is cluster nonresponse. In fact, nonresponse at these higher-level stages of the sampling process is another understudied topic in the survey sampling and missing data literature and, therefore, deserves further investigation into other possible methods to deal with it. The study in Chapter III examined the use of different substitution methods at the Primary Sampling Unit (PSU) level in a two-stage cluster sampling, compared to other standard approaches, such as an unadjusted respondent mean and a nonresponse weighting adjustment procedure. While this is a very common setting for this problem and can possibly be extended to other situations, nonresponse at other sampling stages in more general multi-stage designs should be addressed in future studies. Moreover, Chapter III assumed nonresponse only at the PSU-level and complete response at the element-level. In most surveys, nonresponse will occur at multiple stages of the sampling process. The impact on survey estimates of nonresponse on multiple stages of the sampling process and the performance of different adjustment methods remains a problem that needs further attention in future research.

Although the simulation studies in Chapter III showed consistent patterns across different parameters and clearly depicted what should be expected from the performance of the methods evaluated in those studies, further analytical developments are needed. Particularly, it would be important to derive the nonresponse bias of statistics computed under the substitution procedure over a more general missing mechanism assumption. In this study, the nonresponse mechanism that operates over the substitutes is assumed to be the same as the originally selected units. In many applications, this might not be true. For instance, it is often the case that there is less time to spend to obtain the responses of the substitute units than of the originally selected units. Even if all the other survey protocols are the same, the substitutes will likely have a smaller response propensity than the original units and, therefore, a different nonresponse mechanism. Vehovar (1999) provided a nonresponse bias expression for a respondent mean under a substitution procedure using a deterministic nonresponse approach. While that expression can provide important insights on how different nonresponse mechanisms over the original and substitute units might impact bias, an expression derived under a stochastic nonresponse approach may prove more useful for practical purposes, such as further nonresponse adjustments and responsive designs.

Also related to missing mechanism assumptions, the simulation studies in Chapter IV and V assumed the substitution procedure to be fully successful, that is, after enough iterations of the substitution process, every nonrespondent is substituted by a responding unit. In practice, however, there are always some units that cannot be substituted, mostly due to cost and/or time restrictions of data collection. An extension of the calibrated substitution presented in Chapter IV, where all the respondents (original and substitute units) are calibrated to the entire original sample (respondents and nonrespondents), can be used to address this problem. For the PPM substitution method proposed in Chapter V, other adjustment methods, such as imputation or modeling, may be employed to make further adjustments in the sample if the substitution procedure is not successful in some of the cases. The performance of such extensions should be evaluated in future studies.

Another prominent area that could be very useful for practitioners is how to fit the substitution procedure into a more general responsive design framework. As stated in the introduction of this dissertation, substitution may be seem as a form of responsive design, since it is a proactive reaction to nonresponse during the data collection process. It was only more recently that a responsive design framework was formally developed by Groves and Heeringa (2006). Therefore, being able to fit substitution into that framework can provide some further guidance to practitioners on how to apply this method more properly to their surveys. Moreover, other aspects of responsive design may be applied to the substitution method, such as identifying cases that may pose a nonresponse bias risk if they are substituted or not. In another type of application, a similar procedure was suggested by Peytchev et al. (2010).

Although in Chapter III the properties of the sampling variance estimates of the mean under the substitution procedure were investigated and compared to strata collapsing technique, sampling variance estimation was not the focus of the studies in this dissertation. Moreover, there have not been any studies that looked at this problem in the substitution literature. Since under a MAR assumption the standard error of the sample mean that uses substitution is approximately the same as a complete response standard error estimate, standard techniques, such as Taylor Series approximation, can be used. However, variance estimation of substitute estimates under non-ignorable nonresponse is still an open problem that deserves further investigations.

The adjustment methods investigated in Chapter IV assumed a linear relationship between the survey outcome and the auxiliary covariates used for matching and modeling or calibration. While it has been shown that this assumption is not necessary for the matching procedure in substitution to be successful (Zanutto, 1998), the model and/or calibration adjustments for the differences between nonrespondents and their substitutes in covariates not used in matching need developments for non-linear relationships.

All the simulation studies in this dissertation used a normally distributed survey outcome. While the general pattern of the results obtained here should follow for other types of variables, it is important to confirm this expectation with further simulations or analytical derivation of the properties investigated here. Empirical studies using the methods proposed in Chapter IV and V should be also conducted to investigate their properties in real settings. Also related to distributional assumptions, the PPM substitution method developed in Chapter V assumes a Gaussian model. Although this method is robust for other forms of continuous variables (Sullivan and Adridge, 2015), extensions of this method for binary and categorical data are needed.

Finally, the studies in this dissertation and previous research on substitution have emphasized the properties of linear statistics, such as estimates for population means or proportions. This is an important first step for the understanding of the performance of the substitution methods, but other types of statistics (e.g., quantiles and regression coefficients) need to be addressed in future studies of these methods. Moreover, the methods proposed in this dissertation assumed a single survey variable and one auxiliary covariate. While these methods can be extended to multivariate settings, further investigations of their properties in such applications should be conducted.

References

Andridge, R. R. and Little, R. J. (2011). Proxy Pattern-Mixture for Survey Nonresponse. *Journal of Official Statistics*, Vol. 27, No. 2, pp. 153-180.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3), 251-260.

Cassel, C.-M., Särndal, C.-E., and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. *Incomplete data in sample surveys: Theory and bibliographies*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin). Academic Press, New York: London, 3, 143-160

Deville, C. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Groves, R., Dillman, D., Eltinge, J. and Little, R.J.A. (2002). *Survey Nonresponse*. New York: John Wiley & Sons, Inc.

Groves, R. M and Heeringa, S. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169 (Part 3), pp. 439-457.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3), 471-483.

Little, R. J., and Vartivarian, S. L. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*. 31, pp. 161-168.

Lynn, P. (2004). The Use of Substitution in Surveys. *The Survey Statistician*. No. 49, pp. 14-16.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias through case prioritization. *Survey Research Methods*, Vol. 4, No. 1, pp. 21-29.

Rubin, D. B., and Zanutto, E. (2002). Using Matched Substitute to Adjust for Nonignorable Non-response through Multiple Imputation. In *Survey Nonresponse*, edited by R. Groves, R. J. A. Little, and J. Eltinge. New York: John Wiley, pp. 389-402.

Sullivan, D. and Andridge, R. (2015). A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. *Computational Statistics & Data Analysis*, 82, 173-185.

Vehovar, V. (1999). Field Substitution and Unit Nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, pp. 335-350

Zanutto, E. (1998). Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes. Doctorate thesis submitted for the graduate faculty of Harvard University, May, 1998.