An Evaluation of Nonresponse Bias Using Paradata from a Health Survey

Aaron Maitland[1], Carolina Casas-Cordero[2], Frauke Kreuter[2]
[1]National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782
[2]University of Maryland, 1218 Lefrak Hall, College Park, MD 20742

Abstract

Auxiliary variables that are available for all sample units and related to both the probability of response and the survey variables of interest are potential candidates for nonresponse adjustment variables. The National Health Interview Survey (NHIS) paradata file includes two potentially important sets of auxiliary variables measuring the cooperation and contactability of households in the NHIS sample. In an initial study, these paradata variables were found to have moderate correlations with the probability of response, but weaker correlations with a subset of variables on the NHIS family file. This paper expands this previous study by testing social scientific theories that provide a link between the cooperation and contactability paradata variables and a few key health variables on the NHIS family and sample adult files.

Key Words: nonresponse, paradata, call record

1. Introduction

The trend towards declining response rates has concerned survey researchers for the past several years (De Leeuw and DeHeer 2002). Falling response rates increase the potential for nonresponse bias, which arises when the likelihood of participating in a survey is related to the survey variables.

Overall, there are two general strategies for addressing potential nonresponse bias. One strategy is to increase the response rate so that full participation is received from sample members. This might involve strategies such as offering incentives or addressing concerns that sample members might have through other aspects of the research protocol. Although it might be true that a survey with a very high response rate may not have nonresponse bias, few surveys have been able to achieve these levels.

Since full participation is usually unrealistic, practically every survey will have to deal with nonresponse. A second strategy is to find variables that can be used to statistically adjust the estimates. So, what’s required for adjustment variables? First, the variables must be available for both respondents and nonrespondents. For example, there are a few variables that can be attached to most sampling frames. Traditional weighting variables like region or urbanicity that can be found on even relatively weak sampling frames. Second, good adjustment variables would need to be correlated with both the likelihood of participation and the survey variables of interest.

[1] The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Due to the perceived inadequacies of traditional adjustment variables, mainly weak correlations with the survey variables of interest, survey designers have been searching for an additional set of variables. Paradata in some form or another are often mentioned as potential candidates because the information is available on both responding and nonresponding households. We are considering whether paradata, data that are collected in field operations, might be useful for weighting. Specifically, we focus on whether the variables collected on the Census Bureau's Contact History Instrument (CHI) and available on the National Health Interview Survey (NHIS) public use paradata file would be useful for nonresponse weighting adjustment.

Paradata is potentially useful for both increasing participation and nonresponse adjustment. The information collected with the CHI probably already does provide valuable information for managing field operations and maximizing response rates. However, the variables collected with the instrument may also be important in post-processing as nonresponse adjustment variables. Others have considered similar paradata variables weighting purposes (Kreuter et. al forthcoming). The present effort is one of the first efforts to consider the variables on the NHIS paradata file for this purpose.

2. Data and Methods

The data for this paper come from the 2006 and 2007 National Health Interview Survey. We specifically analyzed variables on the NHIS paradata, family, and sample adult files.

2.1 Sampling

The NHIS is an in-person cross-sectional survey with a multistage area probability sample of households. At the first stage, 428 primary sampling units (PSU's) are drawn from approximately 1,900 geographically defined PSU's from the 50 States and the District of Columbia. Area segments are then selected within PSU's. Procedures are followed to oversample blacks, Hispanics, and Asians.

2.2 Paradata

The 2006 and 2007 NHIS Public Use Paradata Files contain information on more than 40,000 families in the NHIS sample. The datafile is arranged so that one family represents one case. As shown in Table 1, we selected approximately 34,000 of these families each sample year where at least one in-person contact was made with a sample unit member. Most of the families excluded from the analyses were screened out of the sample due to race/ethnicity screening requirements or the household was occupied entirely by Armed Forces adults who are not eligible for inclusion in the survey. We excluded these families because the survey variables of interest were not measured. A much smaller number were excluded because the CHI variables were not measured on them. We analyzed two slightly different subsets of cases within each year. One subset consisted of all families that were not screened out of the NHIS sample and had information collected on them from the CHI. This first subset is the appropriate one to use to analyze the paradata measures of contactability since it is not necessary to make contact with a household for interviewers to record these variables. The second subset consisted of cases with at least one in-person contact attempt so that it was at least theoretically possible for the interviewers to have observed the paradata variables pertaining to the level of cooperation of the household. The reported household response rate for the NHIS in each year was approximately 87 percent and the reported response rate for the sample adult was roughly 70 percent each year. Roughly two-thirds of the nonresponse each year was due to refusals.

Table 1. Number of families in analytical samples.

Sample	2006	2007
Total families on paradata file	44,264	44,462
Total in scope families	34,264	34,448
Total in scope and not missing CHI variables	33,575	34,055
Families with in-person contact attempt	32,342	32,786

We considered three different types of paradata variables in our analysis. All of these variables were recorded by interviewers using the U.S. Census Bureau's Contact History Instrument (CHI) and are publicly available online (www.cdc.gov/nchs/nhis.htm). The first set of variables describes the effort involved in making contact with the household. These measures include indications that no one was home, the type of effort that the interviewer was making to contact, and difficulty locating or obtaining access to a household. The second set of variables measure the cooperation of a household. These are mostly reasons that respondents mentioned for not participating in the interview such as they were not interested, too busy, or had privacy concerns. The third theoretically interesting variable that we considered was whether or not the interview was ever broken off or the case required follow-up due to health reasons. The public use file consists of case level summaries of the contact history for each case rather than information on every contact attempt.

2.3 Survey Variables of Interest

Our survey variables of interest come from two different NHIS data files. First, we chose 31 variables from the NHIS Family File. This information was collected from 29,000 families in each sample year. The variables generally measured the health status and utilization of health services by members of the household as reported by a knowledgeable adult from the household. Next, the NHIS randomly selects one adult per family to answer more detailed questions about their own health. We selected 105 variables from the sample adult file that measured health conditions, health status and limitations, health care access and utilization, and health behaviours. This information was collected on approximately 23,000 sample adults each year.

2.4 Analysis Plan

Our research was aimed at answering two general questions:

First, what are the variables on the CHI measuring? In other words, are each of these variables indicators of a unique phenomenon or are there general patterns or factors that can be identified in the data?

Second, how strongly are the CHI variables correlated with participation and the survey variables? These, of course, are the important considerations in determining whether the variables might be useful for nonresponse adjusment.

We conducted factor analyses using the statistical software package Mplus to answer the first question. Mplus is designed to handle latent variable models like the exploratory factor analyses that we were conducting. Importantly Mplus estimates factor models using a matrix of tetrachoric correlations rather than one consisting of the more traditional Pearson correlations to estimate factor models with dichotomous indicators. The argument for the tetrachoric correlations is that the Pearson based correlations are attenuated with respect to dichotomous variables like those on the paradata file (Muthen 1989). Hence, traditional factor analyses that do not take this into account can be misleading. We proceeded with an exploratory factor analysis by examining scree plots and eigenvalues of the resulting factors. Our final models generally included factors with eigenvalues greater than 1. However, we also examined the interpretability of the rotated factor matrix to decide on the final number of factors. We chose an oblique factor rotation method before interpreting the factors due to the nonzero correlations between our resulting factors. Factor analyses were performed separately on the contactability and cooperation paradata variables.

We next examined the bivariate correlations between the paradata variables, survey participation, and our survey variables of interest to answer our second set of questions. It is important to clarify that the paradata variables and survey participation indicators were collected for all families in the sample. The survey variables of interest were only collected for responding families. Therefore, we are assuming that the correlation between the paradata variables and the survey variables of interest are the same for both respondents and nonrespondents. While this evidence alone is insufficient for assessing nonresponse bias, it is a necessary first step for an analysis like the one conducted in this paper (Peytcheva and Groves 2009). The survey variables used in the analyses include a mixture of dichotomous, ordinal, and interval level variables. We maintained the convention of examining tetrachoric correlations between dichotomous variables, polychoric correlations between ordinal variables, and Pearson correlations between continuous variables throughout this paper.

Additionally, we created some composite variables from the individual paradata variables and checked the correlations between these composite variables and the survey variables of interest. We took this approach, because it is possible that the variables may perform better as a composite than individually. One set of composite variables were based on our factor models. Mplus creates a factor score for each individual based on their values on the paradata variables and the weight given to each paradata variable in the factor model. Last, we ran separate logistic regression models predicting survey participation using the cooperation and contactability paradata variables. We then output predicted response propensities for each family based on these models and examined the correlations between these response propensities and the survey variables of interest.

3. Findings

3.1 Factor Models

We started by analyzing the contactability variables, which revealed two factors. Table 1 shows the results. We have bolded all factor loadings of 0.4 and above in the table to facilitate interpretation. The first factor (Contactability Factor 1) measures what we traditionally think of as noncontact and effort made by the interviewer at making contact. Variables indicating that no one was home or the household did not answer the door when there was evidence that someone was home loaded heavily on this factor. Also loading heavily on the first factor were variables indicating effort that was made by the interviewer such as a previous note or letter was taken, driving by the home, or speaking with a neighbour.

The second factor (Contactability Factor 2) measures problems locating or barriers to obtaining access to a household. For example variables indicating that the sample person was away from home or that the interviewer encountered a locked gate loaded heavily on this factor.

Table 2. Loadings from factor analysis of 2006 NHIS contactability variables.

Contactability variable	Contactability Factor 1	Contactability Factor 2
No one home	**.780**	.092
No one home - appointment broke	.399	.000
No one home - previous note/letter taken	**.825**	-.098
Household does not answer door-evidence someone is home	**.563**	-.046
Drive by	**.497**	.131
Multiple drive by	**.447**	.157
Unable to reach/locked gate/buzzer entry	.001	.**626**
Address does not exist/unable to locate	-.134	.394
On vacation, away from home/at second home	.147	**.468**
Spoke with neighbor	**.492**	.390
Building management/doorman contacted	.155	**.527**
Completed case (Type B or C)	-.090	**.511**
Other specify	.254	.300

Note. We repeated this factor analysis using the 2007 NHIS and observed similar results.

We next analyzed the cooperation variables. The results are shown below in Table 3. The first factor (Cooperation Factor 1) clearly consists of time concerns. For example, indicators like too busy, the interview takes too much time, breaks appointments, and scheduling difficulties load strongly on this factor.

The second factor (Cooperation Factor 2) measures a mixture of privacy and content concerns. For example, indicators like privacy or anti-government concerns, and questions were asked about the survey content load strongly on to this factor. This is potentially an interesting factor as it might be indicating that people are learning about survey content and developing concerns.

The third factor (Cooperation Factor 3) measures general resistance to the survey request. The variable “not interested / does not want to be bothered” loads strongly onto this factor. Also loading on this factor is “hang-up/slams the door” on the interviewer.

The final factor (Cooperation Factor 4) measures gatekeeper issues such as a household member told someone not to participate or the interviewer could only talk to a specific household member.

Table 3. Loadings from factor analysis of 2006 NHIS cooperation variables.

Cooperation variable	Cooperation Factor 1	Cooperation Factor 2	Cooperation Factor 3	Cooperation Factor 4
Not interested/Does not want to be bothered	.330	.134	**.678**	-.015
Too busy	**.768**	.019	.027	.026
Interview takes too much time	**.636**	.301	-.035	-.050
Breaks appointments	**.669**	-.111	.106	.119
Scheduling difficulties	**.666**	.022	-.181	.150
Survey is voluntary	.244	**.486**	.334	-.128
Privacy concerns	-.041	**.956**	-.071	.030
Anti-government concerns	-.044	**.585**	.291	.015
Doesn't understand survey/Ask question	.016	**.438**	-.063	.310
Survey content does not apply	.082	**.508**	.127	.037
Hang up/slams door on FR	.056	-.065	**.720**	.234
Hostile or threatens FR	-.133	.027	**.741**	.257
HH members tell not to participate	-.015	.267	.187	.380
Talk only to specific household member	.158	-.024	-.032	**.572**
Family issues	.182	.100	.019	.375
No concerns	-.238	**-.615**	-.073	-.231
Other specify	.023	.069	.056	**.451**

Note. We repeated this factor analysis using the 2007 NHIS and observed similar results.

In conclusion, the factor analysis results suggest that the paradata variables do logically reduce down to a smaller set of variables.

3.2 Relationships between paradata variables, participation, and survey variables of interest

Anyone who examines the variables on the paradata file might predict that they would be more strongly correlated with participation than the survey variables. As the CHI documentation explains "The data include strategies used for gaining participation and reasons for respondent reluctance." However, it is still worth exploring whether the CHI variables are also correlated with the survey variables. Although many of the variables appear to be measuring general reluctance, they could also be proxies for deeper concerns about the subject matter that might arise in the interview.

We began by looking at the correlations of the factor scores from our factor models with survey participation and the survey variables. This aids in reducing the CHI data to a more manageable set of variables. However, it might also make sense, because the individual variables on the paradata file might be measured with error. For example, interviewers forget to record variables or some of them might be difficult to observe. The factor analysis approach might help smooth out this measurement error by using multiple variables to measure a latent tendency for someone to express a certain type of concern.

Table 4 shows the resulting correlations of the factor scores with survey participation and the survey variables. As expected, the factors based on the individual paradata variables are more strongly correlated with survey participation than the survey variables. The general resistance factor is the strongest correlate with participation. The average correlation between the factor scores and the survey variables is approximately 0.03. The largest correlation is .15.

Table 4.Correlation of factor scores with participation and survey variables (2006 NHIS).

Variable set	Correlation with participation	Correlation with survey variables (absolute values)	
		Average	Maximum
Contactability			
Factor 1: Contact problems or effort	-.25	.03	.15
Factor 2: Location or barrier issues	-.24	.02	.12
Cooperation			
Factor 1: Time concerns	-.25	.03	.09
Factor 2: Privacy or content concerns	-.27	.02	.09
Factor 3: General resistance	-.47	.02	.12
Factor 4: Gatekeeper issues	-.30	.02	.08

Note. We repeated this factor analysis using the 2007 NHIS and observed similar results.

We also looked at the correlation of the individual paradata variables with participation. The top half of Table 5 shows that the cooperation variables are better correlates with participation than the contactability variables. The strongest correlates with participation are the indicators for the sample person indicating they are not interested or do not want to be bothered (-0.69) and the sample person hangs up or slams the door on the interviewer (-0.64). The bottom half of the table summarizes the correlations between the paradata variables and the vector of survey variables. The average correlations are only in the 0.02-0.07 range. Only in a few cases are the correlations larger than 0.2. We also ran separate logistic regression models on the contactability and cooperation variables to obtain a response propensity for each case based on these variables. However, these propensities were not correlated any stronger with the survey variables than many of the individual paradata variables.

Table 5. Summary of the correlation of individual paradata variables with participation and the survey variables (2006 data)

Variable set	Average correlation (absolute values)	Maximum correlation (absolute values)
Correlation with participation		
Contactability variables	.28	.43
Cooperation variables	.33	.69
Correlation with survey variables		
Contactability variables	.02 - .06	.11 - .22
Cooperation variables	.02 - .07	.11 - .28

Note. We repeated this factor analysis using the 2007 NHIS and observed similar results.

The low correlations between the paradata variables and the survey variables in the previous instrument are not surprising given the makeup of most of the variables on the CHI. The does not appear to be an obvious relationship between most of the paradata variables and health. Hence the mechanism for how the paradata variables could reduce nonresponse bias is not very easy to uncover. It does not appear from our analyses that the variables are uncovering any kind of latent concerns about the survey content.

There is one exception in the CHI that actually is more of a direct measure of health. This measure indicates whether an interview could not be conducted or completed due to a health problem. We found that this indicator is somewhat correlated with participation (-0.24) and is relatively more correlated with the survey variables of interest when compared to our previous findings. The average correlation between this indicator and our vector of survey variables is 0.10 (absolute value). The maximum correlation with the survey variables is 0.37. It is at least moderately correlated (~0.3) with a few survey variables, particularly those concerning health limitations on both the family and adult file.

4. Discussion

Our conclusion from analyzing the paradata variables on the NHIS public-use file is that most of the variables on the public use file are probably not well suited for nonresponse adjustment variables.

There are some important limitations to our analyses though. One is that different results might have been obtained if our analyses were based on visit level information. The public use file only contains broad summaries of the contact history for each case. Knowing the sequence of events during the process of obtaining the cooperation of a household might lead to a more useful set of indicators. For example, does it matter when someone expressed a concern or how often it was expressed? We also know very little about the measurement error properties of the CHI variables. What types of errors are they prone to? How reliable are they? Measurement error could attenuate many of the correlations that we examined and should be studied.

It is likely that a different set of variables would need to be included on CHI to be considered for nonresponse adjustment. The CHI was obviously developed in an environment where we have sought to learn about survey participation (Groves and Couper 1998). However, research is shifting solidly towards measuring and reducing error (Groves 2006). This means that we need to consider bias and variance tradeoffs and essentially need better predictors of the survey variables. An important simulation study by Little and Vartivarian (2005) demonstrated that variables that are correlated with the survey variables will at least lead to a reduction in variance even when not correlated very strongly with participation. In contrast, the use of variables that are strong predictors of participation, but not the survey variables will lead to increased variability in the weights without reduction in bias.

It may be worthwhile to think about whether we could measure better correlates of health. This approach obviously brings up a number of measurement questions, but might provide better data for the purposes outlined in this paper. We can not ignore that CHI in its current form has utility as a data management tool and may help understand other aspects of data quality. However, it might be worth considering whether the CHI can be used as an instrument to collect better information that can be used in post-processing of survey data.

References

Dahlhamer JM, Simile CM, Taylor B. Do you really mean what you say? Doorstep concerns and data quality in the National Health Interview Survey (NHIS). Proceedings of the ASA Section on Survey Research Methods, 2008.

De Leeuw E, DeHeer W. Trends in household survey nonresponse: A longitudinal and international comparison. Survey Nonresponse, Groves R, Eltinge J, Little R, (eds.). Wiley: New York, 2002; 41-54.

Groves, RM. Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly 2006; 70: 646-675.

Groves RM, Couper MP. Nonresponse in Household Surveys. Wiley: New York, 1998.

Kreuter F, Olson K, Wagner J, Yan T, Ezzati-Rice TM, Casas-Cordero C, Lemay M, Peythcev A, Groves RM, Rhagunathan TE. Using proxy measures and other correlates of survey outcomes to adjust for nonresponse: examples from multiple surveys. Journal of the Royal Statistical Society; forthcoming.

Little RJ, Vartivarian S. Does Weighting for Nonresponse increase the variance of survey means? Survey Methdology 2005; 31:161-168.

Muthen B. Dichotomous Factor Analysis of Symptom Data. Sociological Methods and Research 1989; 18:19-65.

Peytcheva E., Groves RM. Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. Journal of Official Statistics 2009; 25: 167-191.