The Impact of Nonresponse in Different Survey Phases on the Precision of Prevalence Estimates

Ming Ma (ming.ma@ucdenver.edu)
University of Colorado Anschutz Medical Campus

Research note

Keywords: Nonresponse, Precision, Survey study

Posted Date: March 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-286116/v1

The Impact of Nonresponse in Different Survey Phases on the Precision of Prevalence Estimates

Ming Ma[a], MD, MPH

[a] *Community and Behavioral Health, Colorado School of Public Health, University of Colorado Anschutz Medical Campus*

Address: 13055 E 17th Pl., MS F542. Aurora, CO 80045. USA

Email: ming.ma@cuanschutz.edu

Abstract

Objective: Survey research is widely used in social studies. Whereas it has been widely known that nonresponse might produce biased results and impair the precision, the pattern of the impact on the precision of the estimate due to the non-response in the different survey stages is historically overlooked, though such information is essential to guide the recruitment plan. This study proposed to examine and compare the effect of first and second level nonresponse on the precision of prevalence estimates in the multi-stage survey studies. Based on the benchmark dataset from a state level survey, we used simulation approach to create datasets with different first and second level nonresponse rates and then compare the margin of error (an indicator for the precision) for the 12 outcomes between datasets with first vs. second level nonresponse.

Results: At the same nonresponse rate, the mean margin of error was greater for the data with first level nonresponse, compared to the data with second level nonresponse. As the nonresponse rate increased, the loss of precision was more inflated with the data with first level nonresponse, suggesting that the effort for recruiting primary sampling units is more crucial to improve the estimate precision in survey studies.

Key Words: Nonresponse, Precision, Survey study

Introduction

Survey studies are commonly conducted to collect data from a population of interest, such as individual's attitude, opinions, social and health behaviors, etc. The information obtained from a representative pool of sample helps researchers identify and monitor the trend of some characteristic of interest about the large population and make further research plan. The prevalence of the indicators can also provide the relative agencies with useful information for making data-driven policies, implementing, and evaluating the intervention programs, etc. Thus, it is important to report estimates with sufficient precision because less certainty indicates more fluctuation in the population estimates.

Although it has been well recognized that nonresponse potentially produces the biased estimates [1] and impairs the precision due to the reduced sample size, little studies investigate and compare the change in the precision influenced by the nonresponse in different stages in the multi-stage survey studies. For example, in a health survey study that proposes to sample a group of clinics and then sample a group of patients within each sampled clinic, nonresponse may happen in both stages. The first level nonresponse occurs when some sampled clinics refuse to participate, and the second level nonresponse occurs when some sampled patients fail to complete the surveys. Both levels of nonresponse could result in the reduced sample size and higher variance estimate, whereas it is not clear that whether the nonresponse at the two phases have similar impact on the precision of estimates. For instance, a youth survey study with a 90% of school response rate and an 80% of student response rate has the same overall response rate (72%) as another youth survey that has 80% of schools and 90% of students responded. However, it does not necessarily indicate that the prevalence estimates produced by the two

survey datasets have similar precision. Further insight on the effect of nonresponse in different sampling phases is warranted to advise effort allocation in recruitment for the survey studies. For example, if the first level nonresponse is more likely to result in the inflated variance, researchers need to be aware of that and consider more strategic recruiting plan to improve the response rate of the primary sampling units or to increase the sample size. These methods could avoid the production of survey estimates with lack of certainty, beyond the consideration of the potential biased results.

The survey studies commonly employ the techniques of complex survey design and the calculation of variance involves the "design effect" that depends on multiple components,[2, 3] and thus is not intuitively to compare the impact of different level nonresponse on variance using mathematical derivation. Additionally, to adjust for the nonresponses, large weights are commonly applied and could cause excessive weight variation [4], which is also not feasible to demonstrate using formulas. Most of statistical survey software use Taylor series expansion [5, 6] to estimate the variance for multi-stage sampling design, which is not applicable to be used to compare the impact in the variance due to the nonresponses in different phases. Alternatively, in this study, by using data from a statewide adolescence survey project, we explored the approach of simulation to assess and compare the impact of first and second level nonresponse on sampling variance. The simulation method has been used in previous studies that investigates research questions related to the variance estimation for complex survey data [7, 8]. We expected the findings from this simulation study provides evidence to reinforce the knowledge and to better understand the patterns in the estimate's precision due to different phases of nonresponse with varied response rates.

Main Text

Methods

Healthy Kids Colorado Survey (HKCS)[9]

HKCS is a biennial statewide survey on the health and well-being of young people. The methods of sampling and data analyzing for HKCS are aligned with Youth Risk Behavior Survey (YRBS) conducted by Centers for Disease Control and Prevention's (CDC's) [10] on a two-year cycle since 1991. We chose 2019 HKCS high school dataset as our analytical baseline dataset because of the relatively higher response rate (83% and 71% for school and student's response rates). The survey was administered from September to December in 2019 and included over 120 questions in domains of physical activity, nutrition, bullying, substance use, school and teacher connections, mental health, sexual behaviors, etc.

In the first stage, there were 199 high schools (primary sampling units) systematically sampled from each of the 21 health statistic regions (strata) and 33 of them refused to participate (first stage nonresponse rate = 16.6%). At the second stage, a total of 65,468 students from 9^{th} to 12^{th} grades were sampled and 18,931 of them failed to complete surveys (second stage nonresponse rate = 28.9%). We constructed HKCS weights to account for the selection probability, nonresponse, and the difference in the demographic distribution between the sample and the population of state's high school students [11, 12]. The weighting factors include: the school base weight (W_1); school nonresponse adjustment factor (F_1); classroom selection weight (W_2); classroom nonresponse factor (F_2); an adjustment factor that accounts for students nonresponse (F_3); a post-stratification factor that adjusts the difference between the sample and the population (F_4); The final weights are the products of base weights and adjustment factors (final weight =

$W_1F_1 * W_2F_2F_3 * F_4$). Extreme weights were trimmed.

Baseline Dataset

The sample included 46,537 9th to 12th graders from 166 high schools in 21 health statistic regions. The 12 frequently reported survey questions from several domains and the related constructed binary variables were included in the baseline dataset: (1) Been active 60 minutes on more than 5+ days past 7 days; (2) Had 1+ drinks past 30 days; (3) Ate breakfast on all of the past 7 days; (4) Been bullied at school past 12 months; (5) Fought 1+ times past 12 months; (6) Described grades as mostly A's or B's past 12 months; (7) Used marijuana 1+ times past 30 days; (8) Never/rarely wore seat belt; (9) Ever had sex; (10) Slept 8+ hours/average school night; (11) Smoked 1+ days past 30 days; (12) Attempted suicide 1+ times past 12 months. The weighted prevalence for those indicators based on the original state dataset ranged from slightly lower than 10% to higher than 70%.

Simulation of nonresponding schools and students

We used the simulation to create the datasets with nonresponding schools and students at different rates. For example, to simulate the first stage school nonresponses, a 5%, 10%, 20%, 30%, 40%, 50%, and 60% of schools were randomly dropped from the baseline dataset, respectively. The simulation was repeated 1,000 times at each nonresponse rate and thus created 7,000 datasets with seven different school nonresponse rates. A pre-compiled macro program was applied to re-construct the survey weights for each simulated dataset. For instance, the school nonresponse adjustment factor (F_1) and a post-stratification factor (F_4) were re-calculated, so the sum of the weights between the simulated dataset and the original baseline state dataset would be same. The similar procedures were used to simulate the scenario of second stage student's nonresponse and a total of 7,000 simulated datasets with seven student nonresponse

rates were created.

Statistically analysis

The weighted prevalence and the 95% confidence interval (CI) for 12 binary outcome variables were estimated based on each of the 14,000 simulated datasets. For each outcome variable, the mean margin of error of the point prevalence estimates at each nonresponse rate were calculated separately for the datasets with first (school) and second (student) stage nonresponse respectively. For instance, to calculate the mean margin of error for the outcome variable "Attempted suicide" at a 10% of school nonresponse rate, the margin of error for this binary outcome was obtained from radius of each CI and then was averaged across the 1,000 simulated datasets that had 10% of nonresponding schools. The means of margin of error were plotted and compared at each nonresponse rate. The simulation and survey data analysis procedures were all performed using SAS 9.4 (Cary, NC) [13]. This study was a secondary analysis of the existing data that are publicly available, and thus is exempted from IRB approval.

Results

Figure 1 shows the point prevalence estimates for the 12 outcomes at different school and student nonresponse rates. Because we randomly dropped schools and students from the baseline dataset, there was not substantial fluctuation in the point prevalence estimates for the data with simulated nonresponding schools and students across different nonresponse rates.

Figure 1. Point prevalence estimates for the 12 outcomes at different school and student nonresponse rates.

The mean margin of error increased with the increased nonresponse rates for both the simulated data with nonresponse at first (school) and second (student) stages. At the same nonresponse rate, compared to the survey data with second stage nonresponse, the mean margin of error was

greater for the data with first stage nonresponse (nonresponding schools). Furthermore, with the increased nonresponse rates, the magnitude of the increase in the margin of error were more greatly inflated for the data with first stage nonresponse. The mean margin of error of the 12 outcomes were shown in Figure 2.

Figure 2. Mean margin of error for the estimates of the 12 outcomes at different school and student nonresponse rates.

Discussion

Survey research is a widely used approach in health, social, and other disciplines. Although the adverse consequence of the nonresponses in producing biased outcomes has been widely reported and discussed, the negative effect of the nonresponse of different stages on the variance inflation has not been thoroughly assessed, while the negative relationship between sample size and variance is well known. This study aimed to use simulation approach to assess the impact of nonresponse of different phases on the variance estimation for survey data.

The findings from the simulation indicated that, under the scenarios of the same nonresponse rate, the first phase nonresponse was more likely to impair the precision of the estimates. Furthermore, the magnitude in the difference of the variance was greater at higher nonresponse rate, and this issue was more pronounced for the survey items with higher prevalence estimates (i.e., current smoking).

The findings from the simulation study have reinforced the existing knowledge regarding the variance of the survey data, and besides of that, we further revealed that the nonresponse at different sampling phases had different impact on precision of the estimates. Researchers who conduct survey studies may place extra emphasis on the first phase recruitment and resource may

be leveraged to enhance the response rate from first phase sampling units, especially for the survey studies with instruments that involve common or non-rare items. Though the simulation is based on the dataset from a single survey research, the evidence provided in this study can also be generalized to other multiple stage survey studies.

Limitations

Some limitations of this simulation study need to be noted. First, it is often the case in real world survey studies have both first and second level nonresponding sampling units. It is not practical to simulate the scenarios with mixed nonresponse patterns. Our study simplified the scenario to investigate the nonresponse at the first and second phases separately. Second, because of the considerable amount of computation, this simulation study took long time in the process of generating 14,000 datasets, re-calculating weights, and data analysis.

Declarations

Ethics Approval and Consent to Participate

This study was a secondary analysis of the existing data without individual level identification information, and thus is exempted from IRB approval.

Consent for Publication

Not applicable

Availability of Data and Material

The baseline dataset without school identifiers used in the simulation approach was 2019 HKCS Colorado Survey High School dataset which can be requested from Colorado Department of Public Health and Environment (CDPHE) (https://cdphe.colorado.gov/hkcs), the same dataset

with school identifiers can be requested from the authors upon reasonable request and with permission of Colorado School of Public Health (CSPH). The SAS code used to create simulated datasets and analysis is available from the corresponding author (ming.ma@cuanschutz.edu) on reasonable request.

Competing interests

We do not have any financial or non-financial competing interests to disclose.

Funding

The dataset used in this study was from HKCS Colorado Survey project sponsored by CDPHE (PI: Ashley Brooks-Russell; 19 FHLA 108254)

Author's Contributions

The author of this manuscript conceptualized the research idea/topic and conducted the literature reviews, ran analysis, and wrote the manuscript.

Acknowledgements

Not applicable

Authors' information

The author of this manuscript, Dr. Ma, is a research instructor in Colorado School of Public Health and the statistician for the HKCS Colorado Survey project.

Reference

1. Johnson TP, Wislar JS. Response rates and nonresponse errors in surveys. JAMA. 2012;307(17):1805-6.
2. Heeringa S, West B, Berglund P. Applied Survey Data Analysis: New York: Chapman and Hall/CRC; 2017. 590 p.
3. Burden S, Probst Y, Steel D, Tapsell L. The impact of complex survey design on prevalence estimates of intakes of food groups in the Australian National Children's Nutrition and Physical Activity Survey. Public Health Nutr. 2012;15(8):1362-72.
4. Little RJ, Vartivarian S. Does weighting for nonresponse increase the variance of survey means? 2005.
5. Shah B, editor Linearization Methods of Variance Estimation2005: ew York: John Wiley and Sons; 2005.
6. Wolter K. An Introduction to Variance Estimation: New York: Springer-Verlag; 2007.
7. Sheffel A, Wilson E, Munos M, Zeger S. Methods for analysis of complex survey data: an application using the Tanzanian 2015 Demographic and Health Survey and Service Provision Assessment. J Glob Health. 2019;9(2):020902.
8. Hulliger B, Münnich R. Variance estimation for complex surveys in the presence of outliers. Proceedings of the Section on Survey Research Methods. 2006.
9. Colorado Department of Public Health & Enviroment. Healthy Kids Colorado Survey and Smart Source Information Colorado Department of Public Health & Enviroment [Available from: https://www.colorado.gov/pacific/cdphe/hkcs.
10. Centers for Disease Control and Prevention. YRBSS Methods [Available from: https://www.cdc.gov/healthyyouth/data/yrbs/methods.htm.
11. Kalton G, Flores-Cervantes I. Weighting methods. Journal of official statistics. 2003;19(2):81.
12. Anderson L, Fricker Jr RD. Raking: An important and often overlooked survey analysis tool. Phalanx. 2015;48(3):36-42.
13. SAS 9.4. Cary, NC: SAS Institute Inc.

Figures

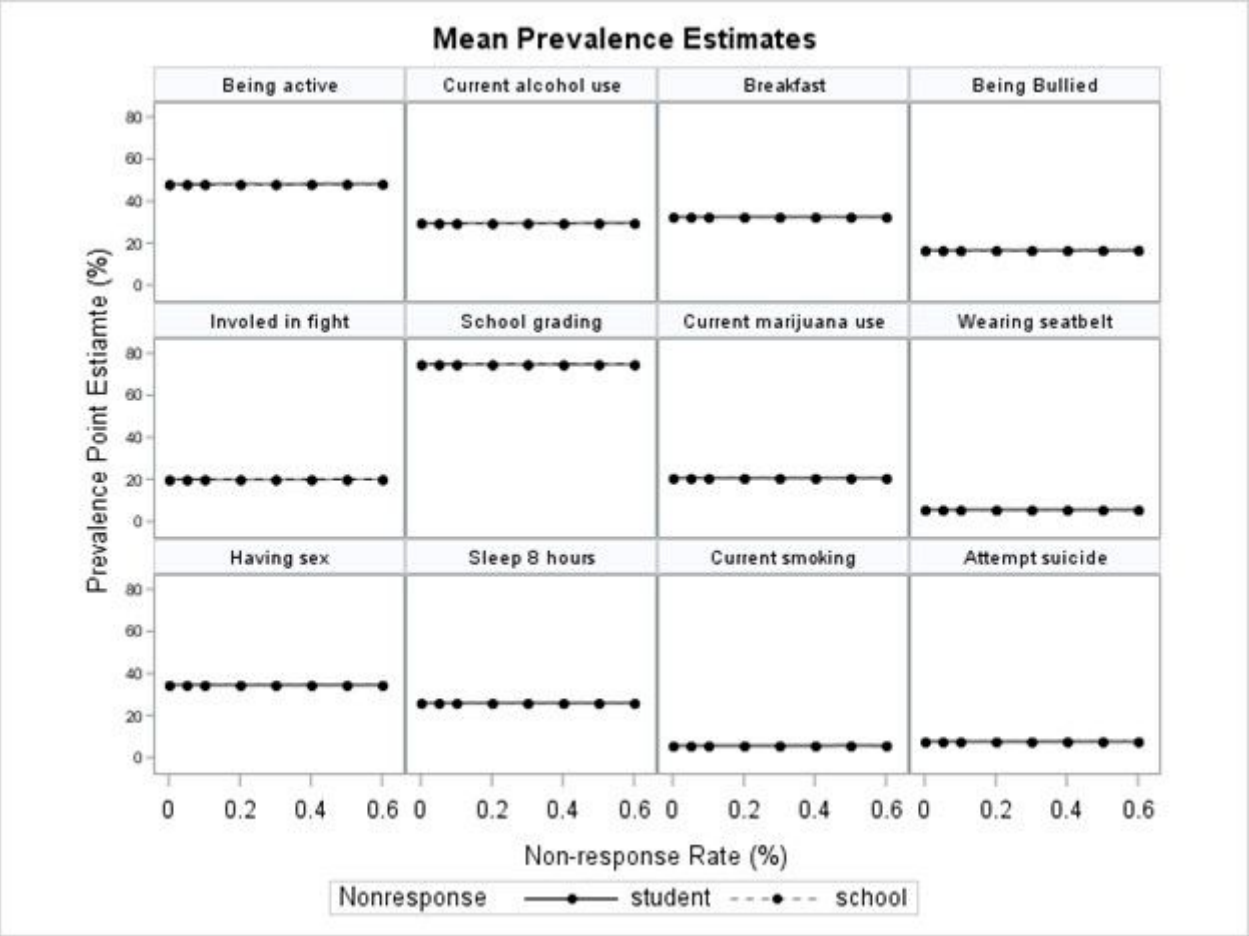

Figure 1

Point prevalence estimates for the 12 outcomes at different school and student nonresponse rates.

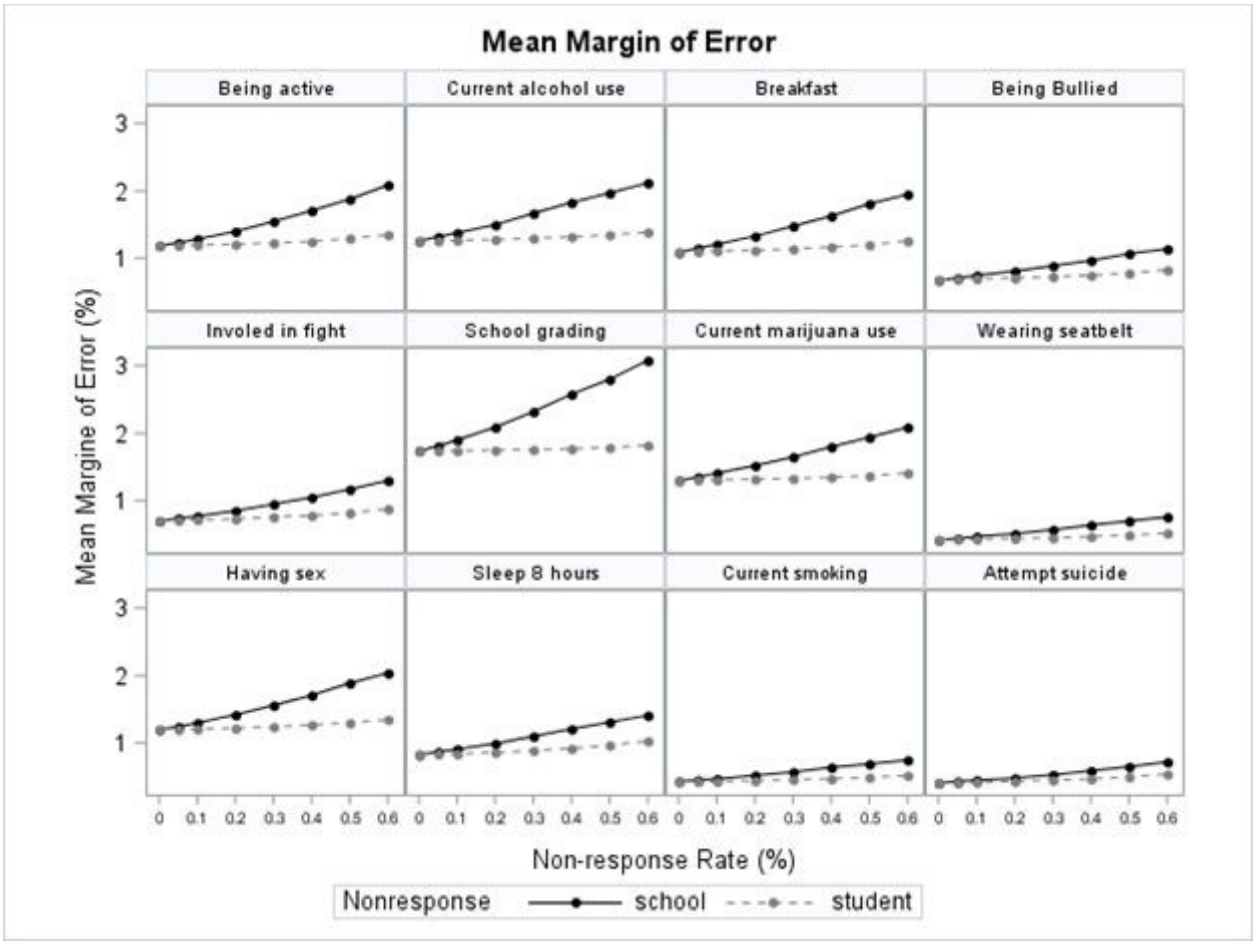

Figure 2

Mean margin of error for the estimates of the 12 outcomes at different school and student nonresponse rates.