Estimating item non-response bias in the 2015 New York City Youth Risk Behavior Survey using multiple imputation

Stephen Immerwahr[1], Lauren Murray[1], Teena Cherian[2]
[1]New York City Department of Health and Mental Hygiene, 42-09 28th Street, Long Island City, NY 11101
[2]Department of Global Health and Social Medicine, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115

Abstract
The Center for Disease Control and Prevention (CDC)'s Youth Risk Behavior Survey (YRBS) is the nation's largest surveillance survey of adolescent health. Item nonresponse is due to many factors, including students skipping questions and post-survey edits. CDC recommends "complete case analysis" (CCA), which assumes missing data are missing completely at random. Item nonresponse in the 2015 New York City (NYC) YRBS was substantially higher than in the national survey. We used multiple imputation (MI) to estimate possible nonresponse bias due to complete case treatment of missing data. We found evidence of item nonresponse bias for multiple measures, including two questions where data was missing from fewer than 10% of respondents. Further evaluation of the sensitivity of YRBS estimates to the treatment of missing data is warranted.

Key Words: Youth Risk Behavior Survey (YRBS), New York City, item nonresponse, nonresponse bias, complete case analysis, multiple imputation

1. Background

The Center for Disease Control and Prevention (CDC)'s Youth Risk Behavior Survey (YRBS) is the nation's largest surveillance survey of the health of adolescents, conducted in odd-numbered years in nearly all states and many municipalities, including New York City (NYC). Using a multi-stage cluster sampling design, all students in sampled classes within participating schools complete a paper-and-pencil survey of up to 99 items, although the national YRBS is only 89 questions in length. Questions are written to allow students to answer every item, but most surveys are affected by item nonresponse. Median item nonresponse in the 2015 NYC YRBS (9.2%) was significantly higher than in the 2015 national YRBS (4.0%) (Wilcoxon Rank-Sum statistic = 6389, ρ <.05). CDC recommends using complete case analysis[1] (CCA), also referred to as listwise deletion, to address missing data, although this approach assumes that missing data are missing completely at random. We estimated potential bias due to CCA treatment of item nonresponse. Although YRBS data are widely used, to the best of our knowledge, this is the first study to estimate bias in the YRBS due to item nonresponse.

[1] https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2015/2015_yrbs-data-users-guide.pdf

1.1 Item nonresponse in the YRBS

Item nonresponse has many causes, including students not answering questions and post-survey edits for consistency applied by CDC. Potential reasons for not answering questions include recall bias, social desirability bias, poor question wording, response fatigue (especially since there are no skip patterns), and survey fatigue. Additionally, some students might not be able to complete the entire survey in the allotted time, which is generally a single class period. CDC drops cases when more than 23 questions are unanswered, or when the same response category has been selected for 15 or more items in a row, i.e. "straight-lining". CDC also applies logical edits by comparing two questions at a time, both of which are set to missing if responses conflict, with the exception of demographic questions. For instance, if a respondent's age at first sex is greater than their current age, both responses are set to missing.

The 2015 NYC YRBS was a stratified, two-stage cluster survey of 8,522 students in 83 public high schools, weighted to be representative of eligible NYC public high school students in grades 9-12. Special education and transfer schools, as well as schools where 30% or more of the students were English language-learners, were not included in the NYC YRBS sample frame. Item nonresponse generally increased during the course of the survey (Figure 1), and was over 20% for the each of the last 20 of 99 survey questions. Shaded bars indicate the placement of questions for which item nonresponse bias is estimated in this analysis.

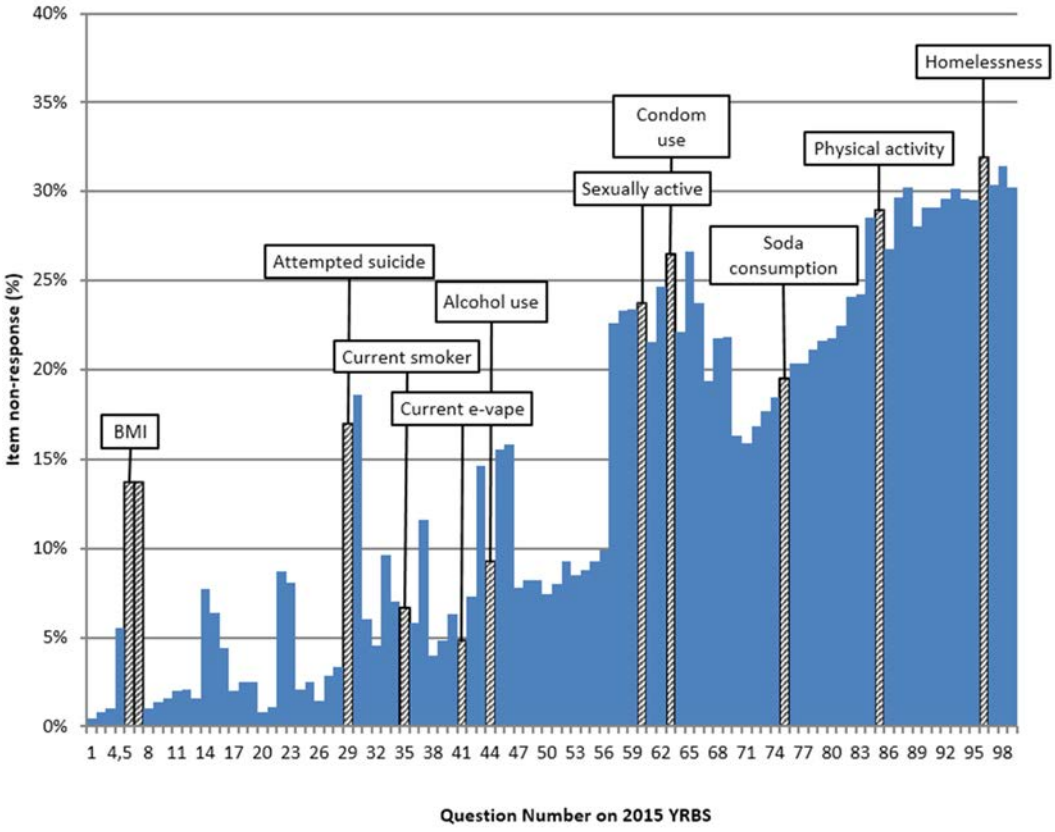

Figure 1: Percent missing data over the course of the 2015 New York City YRBS, by question number. Shaded bars indicate placement of questions for which item nonresponse bias is estimated.

1.1.1 Potential bias due to item nonresponse

The shortcomings of CCA and other naïve treatments of missing data such as "available case" analysis or mean substitution are well known. Unbiased estimates using CCA are only possible when values are missing completely at random, an assumption that is easily challenged and rarely met, with potentially biased point estimates and/or regression coefficients as a result (Frankel et al. 2012; Little & Rubin 2002). Even when data are missing completely at random, CCA reduces efficiency because other information is discarded with the incomplete observations, reducing survey precision due to reduced sample size. However, the focus of this study is on bias in estimates.

1.2 Analysis of item nonresponse and estimation of item nonresponse bias in the 2015 NYC YRBS

1.2.1 General description of missingness and student characteristics

We explored item nonresponse bias in two ways. First, to test the CCA assumption that question responses were missing completely at random, we divided student respondents into quartiles by the total number of questions each student was missing a response. We observed the distribution of demographics across the quartiles and compared health prevalence estimates of students in the lowest and highest quartiles of missing responses. Chi-square tests were used for comparisons of unweighted student demographics across quartiles, while Wald tests were used to compare weighted health estimates for the students in lowest vs. highest quartiles of item nonresponse. For these and for the subsequent analyses, SUDAAN to was used to incorporate the survey design and final sample weights in variance estimation. Statistical significance was set at $\rho < .05$, two-tailed.

1.2.2 Multiple Imputation using IVEware

To estimate bias, we selected 10 health measures from throughout the survey and compared prevalence estimates using CCA to those obtained with multiple imputation (MI) of missing values. *IVEware 0.2* was used to multiply impute missing data at the case level, with imputed values supporting subsequent regression imputations (Raghunathan, Solenberger & Van Hoewyk 2002). A large number of variables were included in the MI process: 28 survey variables, 2 interaction terms (student age by race/ethnicity and student age by ever had sex), and overall school attendance rate. Count variables were imputed using Poisson regression, binary and ordered variables were imputed using logit regression, and continuous variables were imputed using linear regression.

Although *IVEware* provides flexible options for logic and value controls, initial imputed values were strikingly different from observed values when created using a single imputation model. As a result, MI was ultimately conducted in two stages. First, 20 data sets were created with multiply-imputed missing values for all non-sexual behavior questions and lifetime sexual intercourse. Then, the number of sexual partners during the past 3 months and condom use at last sexual intercourse were imputed using the observed and imputed values from the 20 data sets.

1.2.3 Item nonresponse bias

Item nonresponse bias was measured as the absolute and relative differences between the CCA and MI estimates using the final sampling weights. Because the CCA and MI prevalence estimates are made using the same observations, statistical tests to compare two independent samples were not used. Instead, we considered the difference in estimates significant if the imputed values alone differed from the observed values, using chi-square tests.

2. Results

2.1 Overall missingness and the assumption of "Missing Completely at Random"
Median item nonresponse was 9.8% of 99 questions (Mean = 14.1%, SD = 10.2%). At the student level, those in the lowest quartile of item nonresponse were missing data for only 1 of 99 questions; those in the highest quartile were missing data for 24 or more questions (Table 1). There was significant variation in the distribution of student age, sex, race/ethnicity, and grade level across students by quartiles of missingness (Table 1). There were also statistically significant differences in prevalence estimates between students in the lowest and highest quartiles of item nonresponse (Table 2). For instance, those in the lowest item nonresponse quartile were less likely to report attempting suicide in the past 12 months than those in this highest quartile (4.3% vs 14.7%). These differences suggest that the CCA assumption of missing completely at random was not met by the 2015 NYC YRBS.

Table 1: Select demographics of public high school students by quartiles of questions with missing data (out of 99), 2015 NYC YRBS (unweighted)

	Quartile 1 Missing response to 1 item	Quartile 2 Missing responses to 2-5 items	Quartile 3 Missing responses to 6-23 items	Quartile 4 Missing responses to 24+ items
Total sample	2172	2090	2149	2111
	n (%)	n (%)	n (%)	n (%)
Age*				
14 years old or younger	413 (19.0)	400 (19.1)	481 (22.6)	614 (29.3)
15 years old	490 (22.6)	496 (23.7)	542 (25.4)	533 (25.5)
16 years old	543 (25.0)	559 (26.8)	464 (21.8)	478 (22.8)
17 years old or older	726 (33.4)	634 (30.3)	646 (30.3)	467 (22.3)
Sex*				
Female	1257 (54.9)	1196 (57.4)	1051 (49.4)	808 (39.1)
Male	915 (42.9)	886 (42.6)	1078 (50.6)	1260 (60.9)
Race/ethnicity*				
White	395 (18.8)	251 (13.2)	219 (11.3)	197 (10.6)
Black/African-American	455 (21.7)	471 (24.8)	524 (27.1)	507 (27.3)
Asian/Native Hawaiian/ Pacific Islander	355 (16.9)	274 (14.4)	219 (11.3)	99 (5.3)
Hispanic/Latino	825 (39.3)	838 (44.1)	905 (46.8)	984 (53.0)
Multi-racial/Other race	68 (3.2)	65 (3.4)	66 (3.4)	69 (3.7)
Grade*				
9th grade	394 (18.1)	395 (19.0)	538 (25.4)	689 (33.4)
10th grade	496 (22.8)	505 (24.3)	554 (26.1)	546 (26.5)
11th grade	532 (24.5)	551 (26.5)	441 (20.8)	452 (21.9)
12th grade	745 (34.3)	623 (30.0)	577 (27.2)	356 (17.3)

* Chi-square test across quartiles, $\rho < .05$

Table 2: Health prevalence estimates for public high school students by quartiles of questions with missing data (out of 99), 2015 NYC YRBS (weighted)

	Quartile 1 Missing response to 1 item	Quartile 2 Missing responses to 2-5 items	Quartile 3 Missing responses to 6-23 items	Quartile 4 Missing responses to 24+ items
Total sample	2172	2090	2149	2111
	%	%	%	%
Attempted suicide past 12 months*	4.3	6.4	10.6	14.7
Current smoker	5.1	5.1	6.1	7.2
Binge drinking (5+ drinks in a row past 30 days)	21.6	20.7	18.9	22.7
Physical fight past 30 days*	17.4	19.9	23.1	31.0
Condom use last sex (among those having sex past 3 months)	63.4	66.3	68.1	63.2

* Wald test for Q1 vs Q4, $\rho < .05$

2.2 Estimated bias due to CCA treatment of item nonresponse

Item nonresponse for the 10 selected measures (shaded bars in Figure 1, Table 3) ranged from 4.9% for e-vapor products to 31.9% for housing instability in the past 12 months. Estimated bias – the difference between prevalence estimates using CCA and MI – ranged from 1.4% to 3.4% in absolute terms (percentage points) and from -11.3% to 19.2% relative to the CCA estimates (the difference between the CCA and MI estimates as a percent of the CCA estimate). Imputed values were significantly different from observed values for 7 of the 10 measures, including two measures with less than 10% item nonresponse: current cigarette smoking and current e-vapor product use. Calculated relative bias due to item nonresponse was large and significant for two measures in particular: one with moderate nonresponse (obesity: 13.7% missing, CCA = 12.4%, MI = 11.0%) and one with high nonresponse (homelessness: 31.9% missing, CCA = 7.8%, MI = 9.3%).

Table 3: Missingness and prevalence estimates using Complete Case Analysis (CCA) and Multiple Imputation (MI) for select health measures, 2015 NYC YRBS

#	Question topic	Missing %	CCA % (SE)	MI % (SE)	Relative Δ from CCA %
6, 7	BMI - Obese (height, weight)	13.7	12.4 (0.6)	11.0 (0.5)	-11.3*
29	Attempted suicide past 12 months	17.0	8.3 (0.5)	8.9 (0.5)	7.2*
34	Current smoker	7.0	5.8 (0.6)	6.2 (0.4)	6.9*
41	Current use of electronic vapor products	4.9	15.9 (0.8)	15.5 (0.6)	-2.5*
45	Binge drinking (5+ drinks in a row past 30 days)	9.3	8.5 (0.6)	8.6 (0.4)	1.2
60	Had sex during past 3 months	23.8	18.7 (1.5)	19.8 (0.7)	5.9*
63	Used a condom at last sex (among those having sex past 3 months)	26.5	62.2 (2.1)	65.6 (1.5)	5.5
75	Drank soda one or more times per day (not counting diet soda)	19.5	15.8 (0.9)	16.3 (0.6)	3.2
85	Spend 10 minutes or more walking, riding a bike or skateboarding on the way to school (among those usually getting to school by walking, biking, or skateboarding)	28.9	70.3 (0.9)	69.0 (1.0)	-1.8*
96	Lived away from parents or guardians because kicked out, ran away, or were abandoned past 12 months	31.9	7.8 (0.6)	9.3 (0.6)	19.2*

* Wald test for CCA prevalence vs imputed missing values only, not shown, $\rho < .05$

3. Discussion

Item nonresponse in YRBS data sets can be substantial, particularly at the end of the survey, and potential item nonresponse bias in the YRBS has not been previously explored. We found that the assumptions required for CCA to provide unbiased estimates in the presence of item nonresponse were not met for 2015 NYC YRBS data, and evidence of item nonresponse bias for multiple measures, including two questions where data was missing from fewer than 10% of respondents.

The 2015 NYC YRBS had higher item nonresponse than the national YRBS, so the observed size and direction of bias may differ from other populations. However, high

levels of item nonresponse may not be unique to the NYC YRBS and may increase if Web-based administration were to replace the current paper-and-pencil format (Denniston et al. 2010).

Given the widespread use of the YRBS for youth health surveillance at the national and sub-national levels, the presence of item nonresponse bias in the 2015 NYC YRBS, and the paucity of studies on this topic, further evaluation of the sensitivity of YRBS estimates to the treatment of missing data is warranted.

Human Subjects

The 2015 NYC YRBS data were collected under approval from both the New York City Department of Health and Mental Hygiene and New York City Department of Education Institutional Review Boards.

References

Denniston MM, Brenner ND, Kann L, Eaton DK, McManus T, Kyle TM, Roberts AM, Flint KH, Ross JG. 2010. Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Participation, data quality, and perceived privacy and anonymity. Computers in Human Behavior 26(5), September 2010, Pages 1054-1060.

Frankel MR, Battaglia MP, Balluz L, Strine T. 2012. When data are not missing at random: implications for measuring health conditions in the Behavioral Risk Factor Surveillance System. BMJ Open 2012;2:e000696.

Little RJ, Rubin DB. Statistical Analysis With Missing Data. 2nd ed. Hoboken, NJ: Wiley Interscience; 2002.

Raghunathan TE, Solenberger PW, Van Hoewyk J. IVEware: Imputation and Variance Estimation Software user guide. Michigan: University of Michigan; 2002 [accessed September 10, 2018]. Available at http://www.isr.umich.edu/src/smp/ive/