ISSN:1738-6764, eISSN:2093-7504

Socio-demographic Determinants of Item Nonresponse to a Question on Sexual Orientation: Evidence from the National Health Interview Survey

Seungyeon Cho[1]*

[1] Assistant Professor, Department of Economics, Kyonggi University; scho@kyonggi.ac.kr
* Correspondence

https://doi.org/10.5392/IJoC.2023.19.4.010

Manuscript Received 07 March 2023; Received 05 December 2023; Accepted 05 December 2023

Abstract: *Sexual orientation is a sensitive subject and; consequently, in a survey, nontrivial alternatives of item nonresponse arise. Although handling these item nonresponses is particularly important when the missing information is not missing at random, little is known about the sources of nonrandomness. Using data drawn from the 2013–2018 National Health Interview Survey and the multinomial logit model, this study results revealed that participants' socio-demographic characteristics are systematically correlated with the occurrences of missing values in the "don't know" and "refused to answer" response alternatives of the sexual orientation question even when there is controlling for other variables of interest. The results suggest a need for greater attention when researchers conduct an analysis using a complete data set for which pairwise or listwise deletion of the item nonresponse is applied.*

Keywords: Item nonresponse; Missing information; Multinomial logit model; Sexual orientation; National Health Interview Survey

1. Introduction

Increased interest in the economic and health status of sexual minorities has emerged in recent years in response to the availability of information on respondents' sexual orientation or behaviors in nationally representative survey data in many countries. Most of these data categorize the respondents' self-reported sexual orientation into "gay/lesbian", "heterosexual", and "bisexual", which helps researchers examine how individuals' economic and health outcomes differ based on sexual orientation.

However, sexual orientation is a private and sensitive subject and participants may prefer to report socially-approved answers rather than reveal straightforward information [1]. Moreover, those who are reluctant to report their sexual orientation are more likely to answer the question as either "Don't Know (DK)" or "Refuse to Answer (RA)". This type of item non-response is often dealt with as missing information in the sample data, and handling the missing information is important in order to obtain efficient and less-biased estimates in survey research [2–7]. Unless the missing completely at random assumption holds, that is the occurrence of item non-response is systematically correlated with the values of other variables of interest, listwise and pairwise deletion of the missing information may produce biased estimates with reduced statistical power due to information loss [8–11].

Participants with certain profiles may be more likely to report their sexual orientation as "DK" or "RA"; thus, item non-response to a sexual orientation question may be correlated to the participants' socio-demographic characteristics [12], such as income, race, and educational attainment. Recently, there have been attempts to find determinants of item non-response to a sexual orientation question focusing on participants' socio-demographic characteristics. Asians, Hispanics, and Blacks in the Washington state of the United States are more likely to report item non-response on the sexual orientation question [13]. About 30 percent of Canadian gay and bisexual men would not disclose their sexual orientation, and the proportion of non-disclosure differs by age, HIV status, income, ethnicity, and home environments [14]. However, little is understood about

the pattern of the item non-response to a sexual orientation question based on variations in the participants' socio-demographic characteristics at the national level not limited to a specific state in the United States. This is a particularly important issue in empirical ap-plications, such as estimating earning differentials or health status by sexual orientation. As participants' race is a correlate of "DK" and "RA" to a sexual orientation question, it is possible that other individual characteristics, such age, income, and family backgrounds, may also be correlates.

This study fills the research gap by examining determinants of item non-response to a sexual orientation question based on participants' various socio-demographic characteristics using data from the 2013–2018 National Health Interview Survey (NHIS). By doing so, this study draws specific attention to the potential bias researchers face when using a self-reported sexual orientation question in survey data.

2. Data and Methods

The NHIS is an ongoing yearly cross-sectional household survey that aims to monitor the health of the civilian non-institutionalized population in the United States; it is administered by the National Center for Health Statistics of the Centers for Disease Control and Prevention.

The NHIS consists of person-, family-, and household-level files, and approximately 85,000 persons in 35,000 households are surveyed each year. Since 2013, the Sample Adult file of the NHIS has included a single question that asks about the sexual orientation of a randomly selected adult 18 years of age or older in each family unit with the following choices: (a) gay/lesbian, (b) straight, (c) bisexual, (d) something else, (e) I don't know the answer, (f) refused to answer, and (g) not ascertained. For the purpose of this study, the answers (a) through (d) are categorized as item responses by which the participants' sexual orientation is directly revealed. It should be noted that (e) don't know and (f) refused to answer are different types of item non-responses that can occur depending on the nature of the question and the participants' profiles, and interviewers are trained to distinguish between them during the survey. It should be noted that more sensitive questions receive more "RA" responses, while questions that are difficult to answer get more "DK" responses [6]. Consequently, (e) and (f) are treated separately. The NHIS sorts contradictory responses regardless of their intent, answers with coding failure, and those answered by inapplicable respondents, such as (g) not ascertained. Since little is known about how to deal with the occurrence of (g) in the NHIS, the cases affirmative to (g) are excluded from the analytical samples.

The various socio-demographic characteristics of the sample adults (adults, hereafter) include age, marital status, race, educational attainment, home ownership, disability, US citizenship, and employment status, earnings in last year of the survey, and census region of residence. Importantly, those who are not cooperative with or are difficult to contact for the survey interview are more likely to provide item non-response [15], and controlling for such circumstances may help obtain more reliable results. For this reason, the analysis uses the NHIS Paradata file of each survey cycle that documents various circumstances during the interview preparation and process, such as the number of contact attempts for an interview, the reason for an incomplete interview, and the assessment of interview cooperativeness. This study uses an indicator of cooperativeness during the interview, which is measured using a 5-point Likert scale: (1) very good; (2) good; (3) fair; (4) poor; (5) very poor.

In light of the nature of sexual orientation, the sample data are divided by the adults' gender, thus the sample contains male adults (N = 70,350) and female adults (N = 86,858). Adults in multi-family households are excluded in the analytical sample since they may represent the same household characteristics.

To find determinants of item non-response to self-reported sexual orientation, I apply the multinomial logit regression model [17–18] to the samples of male and female adults, respectively. In these specifications, the unordered categorical outcome variable takes three values: 1 for "item response", 2 for "DK", and 3 for "RA". The covariates account for the adults' socio-demographic characteristics. Using the parameter estimates of the model, I calculate Relative Risk Ratios (RRRs), which indicate the likelihood of the covariates on each of the item non-response alternatives relative to those of the reference group, item response. In other words, the RRR measures the extent to which the likelihood of the alternative in the comparison group relative to the likelihood of the alternative in the reference group changes for a unit change in the coefficient value. An RRR > 1 (< 1) means that the likelihood of the alternative in the comparison group relative to the likelihood of the alternative in the reference group increases as the coefficient value increases (decreases). To reflect the different sample frames across the survey years, standard errors are clustered at each year of the survey cycles.

3. Results

3.1 Descriptive Statistics

Table 1 presents the descriptive statistics of the sample of male adults in which most of the participants (98.96 percent) fall into the item response category. I find that some of the male adults' demographic characteristics are associated with higher item response percentages. These characteristics include being younger, married, white, non-disabled, college educated, US citizens, home owners, very cooperative in the interview, or higher income earners ($45,000 or more). Higher percentages of male adults that chose "DK" as a response are unmarried, Hispanic or Asian, disabled, less educated, non-citizens, home renters, very cooperative to the interview, or lower income earners than those who chose "RA".

Table 1. Descriptive statistics of the sample of male adults (n = 70,350)

Variables	Item response (n = 69,692)	Item non-response	
		Don't know	Refused to answer
Age	50.41 ± 18.05	52.08 ± 19.14	52.99 ± 17.46
Married	56.02	28.64	32.31
Race			
White	68.39	53.77	65.00
Hispanic	13.70	21.36	14.62
Black	11.37	12.06	12.31
Asian	5.20	11.31	6.15
Other	1.34	1.51	1.92
With disability	19.32	36.93	23.08
Education level			
Less than high school	13.23	29.15	13.46
High school	44.78	39.70	35.00
College or equivalent	30.36	19.35	34.62
Graduate	11.63	11.81	16.92
US citizen	92.84	84.67	92.31
Home owner	64.17	48.99	56.92
Interview cooperativeness			
Very good	87.34	78.39	66.54
Good	11.23	15.83	20.38
Fair	1.33	4.77	10.77
Poor	0.09	0.75	1.92
Very poor	0.02	0.25	0.38
Earnings			
Less than $20K	47.54	68.09	55.38
$20K–less than $45K	20.54	17.34	14.62
$45K–less than $75K	16.30	7.54	13.46
$75K or more	15.62	7.04	16.54

Note. Figures of age indicate the sample mean ± standard deviation. All others indicate the percentage composition in each category. The statistics for survey years and residence regions are suppressed for brevity. Sample adult weights are applied. Due to rounding, totals within each category may not add up to 100 percent.

Table 2 presents the summary statistics of the sample of female adults in which most of the participants (98.96 percent) fall into the item response category. Higher percentages of female adults that demonstrated an item response are younger, married, white or Black, non-disabled, more educated, home owners, very cooperative in the interview, or higher income earners ($45,000 or more) than those that did not respond. Higher percentages of female adults that chose "DK" as a response are younger, married, Hispanic, Asian, or other, disabled, less educated, non-citizens, home renters, very cooperative in the interview, and income earners less than $20,000 than those who chose "RA".

Table 2. Descriptive statistics of the sample of female adults (n = 86,858)

Variables	Item response (n = 85,952)	Item non-response	
		Don't know	Refused to answer
Age	51.51 ± 18.73	53.59 ± 20.16	57.05 ± 17.93
Married	47.96	27.55	26.82
Race			
White	64.95	50.55	64.25
Hispanic	14.76	25.55	15.92
Black	13.85	10.77	12.29
Asian	5.19	11.13	6.70
Other	1.26	2.01	0.84
With disability	21.55	36.31	31.28
Education level			
Less than high school	13.49	29.74	17.04
High school	44.14	41.97	43.85
College or equivalent	31.07	21.53	28.49
Graduate	11.31	6.75	10.61
US citizen	93.03	85.58	95.25
Home owner	61.65	53.28	57.26
Interview cooperativeness			
Very good	87.84	77.55	67.04
Good	10.72	18.61	22.91
Fair	1.33	3.83	9.22
Poor	0.09	0.00	0.84
Very poor	0.02	0.00	0.00
Earnings			
Less than $20K	63.47	76.64	67.32
$20K–less than $45K	19.47	16.06	18.44
$45K–less than $75K	10.73	5.29	9.78
$75K or more	6.33	2.01	4.47

Note. Figures of age indicate sample mean ± standard deviation. All others indicate the percentage composition in each category. The statistics for survey years and residence regions are suppressed for brevity. Sample adult weights are applied. Due to rounding, totals within each category may not add up to 100% percent.

3.2 Estimation Results

Table 3 reports the RRRs of the multinomial logit model for the sample of male adults. Relative to male adults falling into the item response category, those who are Asian, disabled, or less cooperative in the interview are more likely to respond to "DK" than their respective counterpart groups. Those who are married, other races,

more educated, or with earnings of $45,000 or more are less likely to choose “DK” as a response than their respective counterpart groups. Relative to male adults with item response, those who are less cooperative in the interview are more likely to choose “RA” as a response. Those who are married, US citizens, with earnings more than $20,000 and less than $45,000, or residing in the southern or western regions of the US are less likely to choose “RA” as a response than their respective counterpart groups.

Table 3. Relative risk ratios of the multinomial logit model, male adults (n = 70,350)

Variables	Don’t know		Refused to answer	
	Estimates	Standard errors	Estimates	Standard errors
Age	1.000	0.003	1.005***	0.002
Married	0.407***	0.051	0.400***	0.073
Race (ref.: White)				
Hispanic	1.452	0.477	1.015	0.311
Black	1.270	0.450	1.068	0.293
Asian	3.063***	0.736	0.900	0.337
Other	0.260*	0.166	1.552	0.647
With disabilities (ref.: without disabilities)	2.381***	0.403	1.082	0.232
Education (ref.: less than high school)				
High school	0.492***	0.087	0.788	0.197
College or equivalent	0.354***	0.024	1.243	0.413
Graduate	0.556*	0.139	1.340	0.425
US citizen	0.681**	0.116	0.787*	0.099
Home owner	0.853	0.152	0.902	0.121
Interview Uncooperativeness	1.316*	0.180	2.773***	0.277
Earnings (ref.: less than $20K)				
$20K–less than $45K	0.805	0.146	0.601*	0.132
$45K–less than $75K	0.641*	0.129	1.004	0.214
$75K or more	0.630*	0.116	0.950	0.253

Note. The RRRs are estimated on the basis group of item response. Estimates of state-fixed effects and region of residence are suppressed. Standard errors are clustered at the survey year. Sample adult weights are applied. Interview uncooperativeness is treated as a continuum for model parsimony. *p < 0.05, **p < 0.01, ***p < 0.001.

Table 4 reports the RRRs for the sample of female adults. Relative to female adults with item response, those who are Hispanic or Asian, disabled, or less cooperative in the interview are more likely to choose “DK” as a response than their respective counterparts. Those who are married, Black, more educated, US citizens, or with earnings of $75,000 or more are less likely to choose “DK” than their respective counterparts. Relative to female adults with item response, those who are Asian, US citizens, or less cooperative in the interview are more likely to choose “RA” as a response than their respective counterparts. Those who are married, Black or other race, or residing in the southern region are less likely to choose “RA” than their respective counterparts.

Table 4. Relative risk ratios of the multinomial logit model, female adults (n = 86,858)

Variables	Don't know (n = 548)		Refused to answer (n = 358)	
	Estimates	Standard errors	Estimates	Standard errors
Age	0.989**	0.003	1.012	0.006
Married	0.426***	0.030	0.422***	0.071
Race (ref.: White)				
Hispanic	1.725***	0.167	1.164	0.274
Black	0.690***	0.051	0.614*	0.118
Asian	2.730***	0.324	1.599*	0.380
Other	0.260	0.243	0.080**	0.069
With disabilities	2.418***	0.452	1.270	0.184
Education (ref.: less than high school)				
High school	0.594***	0.055	1.041	0.200
College or equivalent	0.520*	0.160	0.913	0.214
Graduate	0.421*	0.164	0.976	0.298
US citizen	0.664*	0.124	1.593**	0.217
Home owner	1.152	0.191	0.913	0.162
Interview Uncooperativeness	1.394***	0.133	2.554***	0.325
Earnings (ref.: less than $20K)				
$20K–less than $45K	1.025	0.173	1.062	0.196
$45K–less than $75K	0.738	0.240	1.163	0.304
$75K or more	0.481*	0.141	0.910	0.188

Note. The RRRs are estimated on the basis group of item response. Estimates of state-fixed effects and region of residence are suppressed. Standard errors are clustered at the survey year. Sample adult weights are applied. Interview uncooperativeness is treated as a continuum for model parsimony. *p < 0.05, **p < 0.01, ***p < 0.001

4. Discussion and conclusions

The subsample analysis results suggest that the likelihood of the each of the item non-response alternatives relative to the likelihood of the item response changes with not only the adults' ethnic backgrounds, but also their other socio-demographic characteristics. Looking at some of these characteristics pertaining to both male and female adults, lower education levels are positively related to the increased likelihood of a "DK" response, which is consistent with the findings reported by [16]. Disability status is very strongly related to the increased likelihood of a "DK" response, which can be attributed to the notion that participants with disabilities may feel more fatigue when answering a sensitive question, such as one about sexual orientation. Thus, more careful attention is needed in an empirical model, for example, examining disability status based on sexual identity. Asians and other ethnic minorities are more likely to report "DK" than members of the other racial groups, which is consistent with the findings of [13] that large proportions of these racial groups consist of non-citizens who are unfamiliar with the context of the question.

Some other factors are related to the increased likelihood of the item non-response alternatives, but these relationships differ based on the adults' gender. Adult male citizens are less likely to report "DK" and "RA" than adult male non-citizens. Adult female citizens are less likely to report "DK" but more likely to report "RA" than adult female non-citizens. As Reference [13] noted, non-citizens are more likely to have poor English language skills; thus, they are more likely to choose "DK" as a response since the sexual orientation question include terms, such as straight and bisexual, that may be unfamiliar to them. However, adult female citizens are more likely to report "RA" than adult female non-citizens, suggesting that male and female adults are different in terms of citizenship status when they refuse to report their sexual orientation. Interestingly, the degree of interview uncooperativeness is related to the likelihoods of each of the item non-response alternatives differently. That is, as adult participants are more reluctant to comply with the interview process, their responses are more likely to fall into the "RA" category rather than the "DK" category.

It should be noted that there are some limitations on the estimation part of this study. To consider the categorially distributed alternatives, this study employed the multinomial logit model that relies on the satisfaction of the unrealistic assumption of the independence of irrelevant alternatives. Although some empirical models, such as nested logit, mixed logit, and multinomial probit models, allow for including the correlated structure of the alternatives, these models were not applicable to the analytical data since no alternative specific covariates are included in the model used in this study. In this regard, it is possible that the model estimates may be biased proportional to the extent of the correlations of the alternatives. Thus, future research will have to apply other estimation techniques when to incorporate the alternative specific covariates in the model.

5. Conclusions

Using the data drawn from the 2013–2018 NHIS and the multinomial logit model, this study investigates the determinants of the occurrences of missing values in the "DK" and "RA" choices in the sexual orientation question based on the participants' socio-demographic characteristics even if controlling for other variables of interest. The results suggest a need for greater attention when researchers conduct an analysis using a complete data set that includes a sexual orientation variable in which the missing information may not be missing at random. In particular, sexual orientation is potentially endogenous when the confounders are correlated with both the indicator of sexual orientation and the outcome variable, which results in the need for additional considerations in empirical applications.

Conflicts of Interest: The authors declare no conflict of interest.

Funding: This research received no external funding.

References

[1] K. B. Coffman, L. C. Coffman, and K. M. Ericson, "The size of the LGBT population and the magnitude of Antigay sentiment are substantially underestimated," Management Science, vol. 63, no. 10, pp. 3168–3186, 2017. doi: https://doi.org/10.1287/mnsc.2016.2503.

[2] E. D. De Leeuw, J. J. Hox, and M. Huisman, "Prevention and treatment of item nonresponse," Journal of Official Statistics, vol. 19, pp. 153–176, 2003.

[3] G. B. Durrant, "Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates," International Journal of Social Research Methodology, vol. 12, no. 4, pp. 293–304, 2009. doi: https://doi.org/10.1080/13645570802394003.

[4] R. J. Little and D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, New York, NY, USA, 1987.

[5] I. O. Oshungade, "Some methods of handling item non-response in categorical data," Journal of the Royal Statistical Society, vol. 38, no. 4, pp. 281–296, 1989. doi: https://doi.org/10.2307/2349061.

[6] P. J. Shoemaker, M. Eichholz, and E. A. Skewes, "Item nonresponse: distinguishing between don't know and refuse," International Journal of Public Opinion Research, vol. 14, no. 2, pp. 193–201, 2002. doi: http://doi.org/10.1093/ijpor/14.2.193.

[7] T. Yan and R. Curtin, "The relation between unit nonresponse and item nonresponse: a response continuum perspective," International Journal of Public Opinion Research, vol. 22, no. 4, pp. 535–551, 2010. doi: https://doi.org/10.1093/ijpor/edq037.

[8] J. D. Dziura, L. A. Post, E. Zhao, Z. Fu, and P. Peduzzi, "Strategies for dealing with missing data in clinical trials: from design to analysis," The Yale Journal of Biology and Medicine, vol. 86, no. 3, pp. 343, 2013.

[9] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomized clinical trials–a practical guide with flowcharts," BMC Medical Research Methodology, vol. 17, no. 1, pp. 1–10, 2017. doi: https://doi.org/10.1186/s12874-017-0442-1.

[10] T. D. Pigott, "A review of methods for missing data. Educational Research and Evaluation," vol. 7, no. 4, pp. 353–383, 2001. doi: https://doi.org/10.1076/edre.7.4.353.8937.

[11] J. R. Van Ginkel, M. Linting, R. C. Rippe, and A. Van der Voort, "Rebutting existing misconceptions about multiple imputation as a method for handling missing data," Journal of Personality Assessment, vol. 102, no. 3, pp. 297–308, 2020. doi: https://doi.org/10.1080/00223891.2018.1530680.

[12] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," Bmj, pp. 338, 2009. doi: https://doi.org/10.1136/bmj.b2393.

[13] H. J. Kim and K. I. Fredriksen-Goldsen, "Nonresponse to a question on self-identified sexual orientation in a public health survey and its relationship to race and ethnicity," American Journal of Public Health, vol. 103, no. 1, pp. 67–69, 2013. doi: https://doi.org/10.2105/ajph.2012.300835.

[14] O. Ferlatte, T. S. Hottes, T. Trussler, and R. Marchand, R, "Disclosure of sexual orientation by gay and bisexual men in government-administered probability surveys," LGBT Health, vol. 4, no.1, pp. 68–71, 2017. doi: https://doi.org/10.1089/lgbt.2016.0037.

[15] S. Lee, K. I. Fredriksen-Goldsen, C. McClain, H. J. Kim, and Z. T. Suzer-Gurtekin, "Are sexual minorities less likely to participate in surveys? an examination of proxy nonresponse measures and associated biases with sexual orientation in a population-based health survey," Field Methods, vol. 30, no. 3, pp. 208–224, 2018. doi: https://doi.org/10.1177/1525822x18777736.

[16] C. S. Craig and J. M. McCann, "Item nonresponse in mail surveys: extent and correlates," Journal of Marketing Research, vol. 15, no. 2, 285–289, 1978. doi: https://doi.org/10.2307/3151264.

[17] W. H. Greene, Econometric Analysis, 8th ed., Pearson Education Limited: London, UK, 2018.

[18] R. Davidson and J. G. MacKinnon, Estimation and Inference in Econometrics, New York, Ofxord University Press, NY, USA, 1993.