Current Population Survey Veterans Supplement Nonresponse Bias Analysis August 2018

Morgan Earp[1]

[1]Bureau of Labor Statistics, PSB Suite 5930, 2 Massachusetts Avenue NE, Washington, DC 20212

1. Introduction

The Veterans Supplement (CPS-VS) is an annual supplemental survey to the Current Population Survey sponsored by the Department of Labor and Department of Veterans Affairs. The supplemental survey is conducted by the U.S. Census Bureau and is used to estimate demographic, employment, and service related disability status rates among U.S. veterans ages 17 or older (U.S. Bureau of Labor Statistics, 2016). The CPS-VS data are collected from CPS respondents 17 years of age or older, who identify as a veteran or identify someone else in the household as veteran. Because the CPS-VS sample is drawn from CPS respondents, CPS respondent data are available for both the CPS-VS respondents and nonrespondents and therefore can be used to compare the characteristics of CPS-VS respondents and nonrespondents, as well as assess the potential for CPS-VS nonresponse bias. No information about the households that did not respond to the basic CPS survey is available in the CPS-VS public access dataset, so all conclusions on CPS-VS nonresponse bias are based on the households responding to the basic CPS.

2. Methodology

Data

Between August 14-23 of 2016, 8,549 veterans or their proxies were administered the 2016 CPS-VS, with 7,108 (83 percent) responding to the Veterans Supplement. CPS data were available for the entire 2016 CPS-VS sample. The U.S. Census Bureau provides a single file of all persons selected to receive the Veterans Supplement, already containing all household and respondent level CPS and CPS supplement variables responses; therefore, no files needed to be merged to create our analysis file (U.S. Census Bureau, 2017).

Nonresponse Assessment

We used several CPS variables to model CPS-VS response (see Table 1). Using a regression tree model, we were able to automatically detect main, mediating, and interaction effects for all the variables shown in Table 1. A regression tree model is constructed by recursively splitting the data based on characteristic variables and response propensities. At each iteration, the regression tree model automatically selects the variable and breakpoint to best maximize the heterogeneity across subgroups and the homogeneity within groups with regard to response propensities. In other words, the regression tree model identifies the variables and variable breakpoints that are most strongly related to survey response in a sequential order, starting with the full dataset and then each of the resulting response propensity subgroups (or nodes as they are referred to in the literature). The variables identified and their breakpoints provide insight into the characteristics most related to survey response, and can indicate whether the respondents and nonrespondents are systematically different, or if there is likely nonresponse bias. The resulting end nodes of the regression tree can be used to classify and describe varying response propensity subgroups for purposes of adaptive/responsive design, analysis of nonresponse bias, and/or nonresponse adjustment.

Table 1: CPS Frame Variables

CPS Household Income (*Less than $5,000 through $150,000 and Over*)
CPS Household Income Missing Indicator (*Reported/Imputed*)
CPS Household Own Indicator (*Own/Not Own*)
CPS Household Ownership Type (*Owned, Rented, Occupied without Payment*)
CPS Household Size
CPS Marital Status (*Married-Spouse Present, Married-Spouse Absent, Widowed, Divorced, Separated, Never Married*)
CPS Marital Status Collapsed (*Married/Not Married*)
CPS Number of Children under 18
CPS Children Indicator (*Children/No Children*)
CPS Proxy Respondent Indicator (*Self, Self and Proxy, Proxy, Proxy Missing*)
CPS Respondent Age
CPS Respondent Disability Flag (*Y/N*)
CPS Respondent Education Level (*Less than 1st Grade through Doctoral Degree*)
CPS Respondent Employment Status (*Employed-At Work, Employed-Absent, Unemployed-Layoff, Unemployed-Looking, Not In Labor Force-Retired, Not In Labor Force-Disabled, Not in Labor Force-Other*)
CPS Respondent Employment Status Collapsed (*Employed/Not Employed*)
CPS Respondent Hispanic (*Y/N*)
CPS Respondent Race (*All Reported Combinations*)
CPS Respondent Race Collapsed (*White/Black/Other*)
CPS Respondent Sex (*Male/Female*)
CPS Veteran Military Service Period [*September 2001 or Later; August 1990 to August 2001; May 1975 to July 1990; Vietnam Era (August 1964 to April 1975); February 1955 to July 1964; Korean War (July 1950 to January 1955); January 1947 to June 1950; World War II (December 1941 to December 1946); November 1941 or Earlier*)]
Veterans Supplement Interview Flag (Supplement Interview/Supplement Non-Interview)

The regression tree model was grown using the CHAID growth model within the SAS HPSPLIT procedure (Cohen & Rodriguez 2013; SAS 2015). A Bonferroni adjustment was applied to prevent the model from favoring variables with more categories. The tree was allowed to continue splitting as long as the response propensity groups contained at least 50 observations. In order to avoid overfitting, we also limited the number of splits that occur under a single branch also known as the depth of the tree to six. After we identified and controlled for characteristics associated with CPS-VS response using the regression tree, we compared response, employment, and disability status rates among the homogenous CPS-VS response rate groups identified by the regression tree model, using the reported CPS employment (CPS Respondent Employment Status Collapsed, Employed/Not Employed) and disability status (CPS Respondent Disability Flag, Y/N), which again are available for both CPS-VS respondents and nonrespondents. We selected employment and disability because they were the closest proxies of the CPS-VS variables, and had data for the entire CPS-VS sample. Using these proxy variables, we assessed the potential for nonresponse bias in relation to veterans' employment and disability status.

3. Results

As noted above, the resulting regression tree was grown using the chi-square automatic interaction detection (CHAID) algorithm (deVille 2006). This method uses a variety of statistical tests to detect significant splits depending on the variable type. In addition to the statistical test used to detect significant variables and breakpoints, there are also multiple stopping criteria that can be used to decide how large to grow a decision tree as described above. The complete regression tree model presented in this paper resulted in a tree with ten leaves (also known as end nodes). Together the leaves comprise the entire CPS-VS dataset, but separately each contain mutually exclusive response propensity groups that vary based on their propensity to respond as well as key characteristics related to response propensity. Each of these end nodes has varying CPS-VS response propensities (see Figure 1, Nodes 1, 7, 8, 10, 11, 13, 14, 16, 17, and 18). A hierarchical description of the tree shown in Figure 1, significant variables (and breakpoints), as well as the resulting end nodes (which are bolded in the text below) are presented below.

Tree Description

The strongest predictor of CPS-VS response was whether or not household income was missing and imputed for the household in which the veteran resided. If household income was reported, the CPS-VS response rate was 89 percent (Node 1), versus a 65 percent response rate for those respondents who resided in households that did not report income on the CPS (Node 2).

Employment Status was the second strongest predictor of CPS-VS response for veterans residing in households that refused to report household income on the CPS. Veterans who were employed, looking for work, or not in the labor force because they were disabled or for some other reason according to the CPS had a lower response rate of 58 percent (Node 3) versus 73 percent for veterans who were laid off or retired (Node 4).

Moving down the branch of veterans who were employed, looking for work, or not in the labor force because they were disabled or for some other reason, home ownership was a significant predictor of CPS-VS response. Veterans who reported not owning their home on the CPS had a lower response rate of 47 percent (Node 5) versus 62 percent for those that reported owning their home (Node 6).

The CPS proxy reporter indicator was a significant predictor of CPS-VS response for veterans who did not own their home. Veterans who self-reported had a higher response rate of 55 percent (Node 9) versus 36 percent for those who a) self-reported and used a proxy reporter, b) had no self-reporting (all data was provided by a proxy reporter on the CPS), or c) were missing data for the CPS proxy indicator (Node 10).

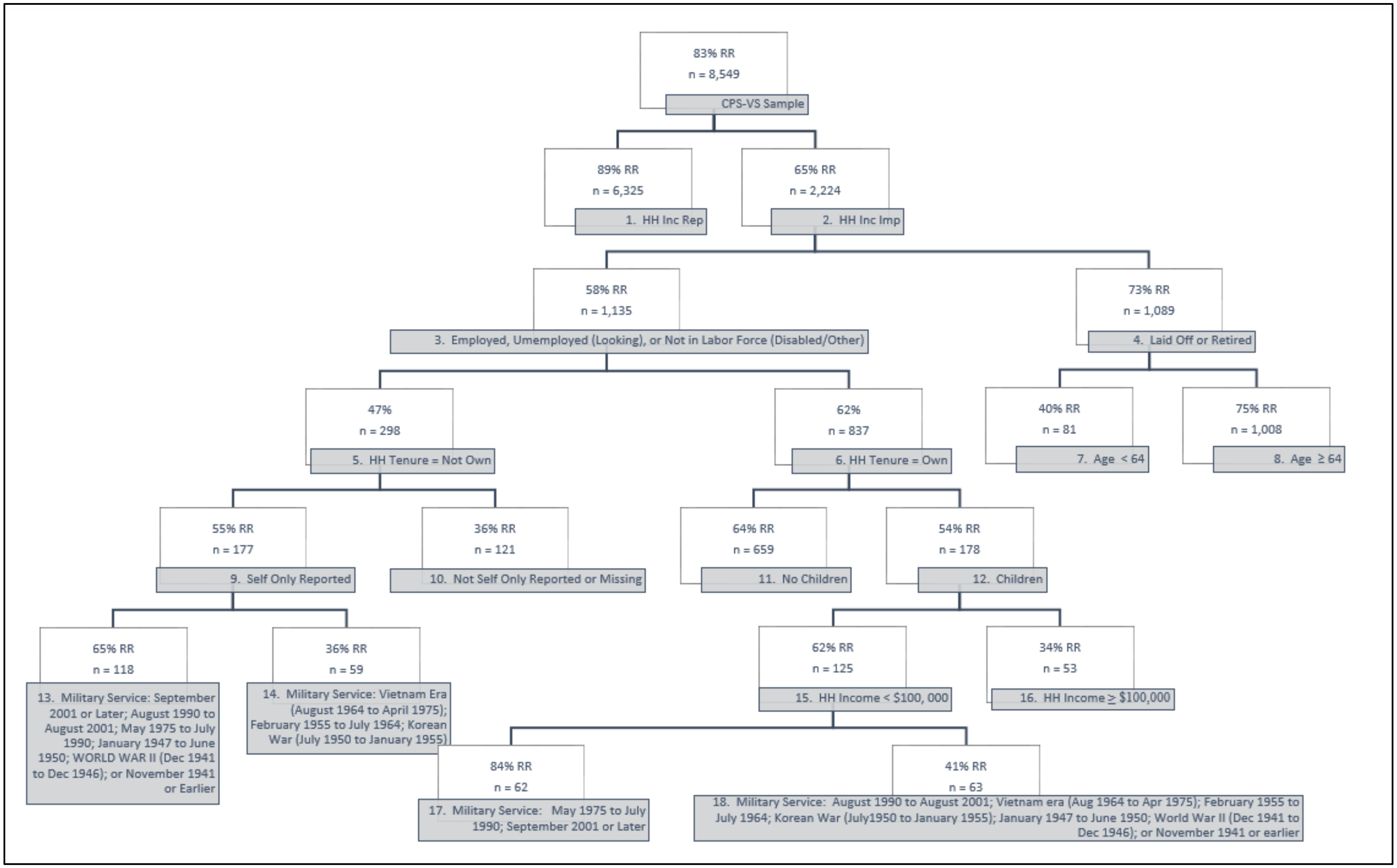

Figure 1: CPS-VS Response Regression Tree Model

Military Service time period reported on the CPS was a significant predictor of CPS-VS response for veterans who self-reported. Veterans reporting military service dates of September 2001 or Later; August 1990 to August 2001; May 1975 to July 1990; January 1947 to June 1950; World War II (Dec 1941 to Dec 1946); or November 1941 or earlier, had a higher response rate of 65 percent (Node 13) versus 36 percent for veterans reporting military service dates during the Vietnam Era (August 1964 to April 1975); February 1955 to July 1964; or the Korean War (July 1950 to January 1955) (Node 14).

Presence of children in the home reported on the CPS was a significant predictor of CPS-VS response for veterans reporting owning their home. Veterans reporting no children on the CPS had a higher response rate of 64 percent (Node 11) versus 54 percent for those with children (Node 12).

Imputed household income greater than $100,000 on the CPS was a significant predictor of CPS-VS response for veterans reporting having children. Veterans with an imputed household income of less than $100,000 had a higher response rate of 62 percent (Node 15) versus 34 percent for those with imputed CPS income of $100,000 or more (Node 16).

Military Service time period reported on the CPS was a significant predictor of CPS-VS response for veterans with an imputed household income of less than $100,000. Veterans reporting military service dates of May 1975 to July 1990; September 2001 or Later, had a higher response rate of 84 percent (Node 17) versus 41 percent for veterans reporting military service dates during August 1990 to August 2001; Vietnam era (Aug 1964 to Apr 1975); February 1955 to July 1964; Korean War (July1950 to January 1955); January 1947 to June 1950; World War II (Dec 1941 to Dec 1946); or November 1941 or earlier (Node 18).

Switching to the right side of the tree, age was a significant predictor of CPS-VS response for veterans who were laid off or retired. Veterans who were less than 64 years of age had a lower response rate of 40 percent (Node 7) versus 75 percent for those 64 years of age or older (Node 8).

Nonresponse Bias Assessment

Again since the CPS-VS is drawn from CPS respondents, we are able to use CPS employment and disability status reported during the basic CPS to assess the potential for nonresponse bias on CPS-VS items related to employment and disability status. In order to assess the potential for nonresponse bias, we compare the reported CPS employment and disability status rates for CPS-VS respondents versus nonrespondents. Persons identifying as being “Employed at Work” or “Employed – Absent” are classified as being employed; everyone else who is identified as being unemployed or not in the labor force, is not classified as being employed. We use the basic CPS disability flag to identify disability status. This flag identifies all persons who report having a disability that affects their hearing, seeing, memory, mobility, and independent living ability.

If there is a significant difference in employment or disability status rates between respondents and nonrespondents within a specific end node, then there is evidence of potential nonresponse bias on any or all CPS-VS items related to either respondent employment or disability (e.g., the variables shown on the regression tree described above). It is important to note that the disability flag constructed using the disability questions from the basic CPS is not highly correlated with any of the CPS-VS disability items. By looking only at regression tree end nodes, we are able to identify potential nonresponse bias that remains after controlling for key CPS-VS respondent characteristics (see Tables 1 and 2).

We first compared reported CPS employment status rates among CPS-VS respondents and nonrespondents within each of the ten terminal nodes using a Chi-Square Test (see Table 2 and Figure 2). By examining the difference between CPS-VS respondents and nonrespondents in CPS reported employment status rates, we identified two nonresponse propensity groups that exhibited significant nonresponse bias: Nodes One and Node Eight (see Table 2 and Figure 3). Node One is veterans in households that reported household income; and Node 8 is veterans in households that did not report household income, who reported being laid off or retired on the CPS, and were 64 years or older.

According to Table 2, we would expect to underestimate employment by about half a percent for veterans in households that reported household income (an error of -14.59 employees), and by 1.78 percent for veterans in households that did not report household income, reported being laid off or retired on the CPS, and were 64 years or older (an error of -6.85 employees) (see Table 2). The overall error for employment was -24.59 employees (an overall nonresponse bias of -0.65 percent). The overall nonresponse bias was calculated as follows:

Total Error = $\sum[(\hat{p}_r - \hat{p}_t) * (n_t)]$

CPS Employment Status Rate $= \frac{\left((\sum n_{emp}) + (Total\ Error)\right)}{(\sum n_t)}$,

CPS-VS Employment Status Rate $= \frac{(\sum n_{emp})}{(\sum n_t)}$

Overall Nonresponse Bias $= \frac{(\text{CPS Employment Status Rate} - \text{CPS-VS Employment Status Rate})}{(\text{CPS Employment Status Rate})}$

Where,

$\hat{p}_r$= the employment status rate of respondents within a given node,
$\hat{p}_t$= the overall employment status rate (including both respondents and nonrespondents) within a given node,
n_t= the total number of persons in a given node,
n_{emp}= the total number of persons employed in a given node.

Next we compared CPS disability status rates among CPS-VS respondents and nonrespondents within each of the ten terminal nodes using a Chi-Square Test (see Table 3 and Figure 4). By examining the difference between respondents and nonrespondents in disability status rates, we determined that none of the response propensity groups exhibited significantly different disability status rates between CPS-VS nonrespondents and respondents (see Table 3 and Figure 5), meaning we did not see any evidence of nonresponse bias with regard to veteran's status. It is worth noting that the p value for

Node 13 was very close at .05; however, the resulting error for this node was the second to the lowest, meaning if it had been significant, the impact would have been rather small (an error of -1.94 persons with disabilities total being overestimated). The overall error for disability was +4.10 veterans (an overall nonresponse bias of 0.19 percent).

Conclusion

Using a regression tree in tandem with Chi Square tests, we identified two groups of CPS-VS nonrespondents who have significant nonresponse bias related to CPS employment status: 1) veterans in households that reported household income (Node 1); and 2) veterans in households that did not report household income, reported being laid off or retired on the CPS, and were 64 years or older (Node 8). We did not identify any groups of CPS-VS nonrespondents who had significant nonresponse bias related to CPS disability status.

4. Discussion

Using a regression tree model, we identified ten mutually exclusive CPS-VS response propensity groups. These groups were defined using a number of CPS household and respondent level characteristics. We found evidence of significant nonresponse bias in two out of the ten groups with respect to employment status rates and in none of the ten groups with respect to disability. We found significant differences between respondents' and nonrespondents' employment status rates for 1) veterans in households that reported household income, and 2) veterans in households that did not report household income, who reported being laid off or retired on the CPS, and were 64 years or older. For both of these groups we would expect to consistently underestimate characteristics related to employment on the CPS-VS, but by only a small percent compared to the overall. The overall relative nonresponse bias across all ten nodes was only -0.65 percent for employment status rate. We found no groups that exhibited significant differences between the respondents and nonrespondents in terms of disability status rates. The overall relative nonresponse bias across all ten nodes was only 0.19 percent for disability status rate. Therefore we conclude that the while we saw some evidence of potential nonresponse bias for employment, the overall relative nonresponse bias is less than one percent for both employment and disability status rates.

Table 2: CPS Employment Rate by Regression Tree End Nodes and Response Status

Node	Resp Status	Count	EMP Rate	EMP Count	EMP Rate LL	EMP Rate UL	NR - RESP	NR - RESP LL	NR - RESP UL	Chi Square	P Value	Significant	Overall EMP Rate	Overall - Resp	Emp Est Error
1	NR	667	49.63%	331	45.83%	53.42%	4.91%	0.90%	8.92%	5.81	0.02	1	45.23%	-0.51%	-14.59
	RESP	5,658	44.72%	2,530	43.42%	46.01%									
7	NR	49	46.94%	23	32.97%	60.91%	12.56%	-9.02%	34.15%	1.25	0.26	0	41.98%	-7.60%	-2.58
	RESP	32	34.38%	11	17.92%	50.83%									
8	NR	250	43.60%	109	37.45%	49.75%	7.19%	0.15%	14.23%	4.12	0.04	1	38.19%	-1.78%	-6.85
	RESP	758	36.41%	276	32.99%	39.84%									
10	NR	78	42.31%	33	31.34%	53.27%	-13.51%	-31.96%	4.95%	2.03	0.15	0	47.11%	8.70%	4.96
	RESP	43	55.81%	24	40.97%	70.66%									
11	NR	236	49.58%	117	43.20%	55.96%	4.90%	-3.05%	12.84%	1.46	0.23	0	46.43%	-1.75%	-5.36
	RESP	423	44.68%	189	39.94%	49.42%									
13	NR	41	48.78%	20	33.48%	64.08%	-1.87%	-20.81%	17.07%	0.04	0.85	0	50.00%	0.65%	0.38
	RESP	77	50.65%	39	39.48%	61.82%									
14	NR	38	47.37%	18	31.49%	63.24%	4.51%	-21.95%	30.97%	0.11	0.74	0	45.76%	-2.90%	-0.78
	RESP	21	42.86%	9	21.69%	64.02%									
16	NR	35	51.43%	18	34.87%	67.99%	-4.13%	-32.43%	24.18%	0.08	0.78	0	52.83%	2.73%	0.76
	RESP	18	55.56%	10	32.60%	78.51%									
17	NR	10	50.00%	5	19.01%	80.99%	1.92%	-31.91%	35.76%	0.01	0.91	0	48.39%	-0.31%	-0.09
	RESP	52	48.08%	25	34.50%	61.66%									
18	NR	37	48.65%	18	32.54%	64.75%	2.49%	-22.54%	27.53%	0.04	0.85	0	47.62%	-1.47%	-0.44
	RESP	26	46.15%	12	26.99%	65.32%									
		8,549		3,817										Total Error	-24.59
														CPS-VS Full Sample Employment Status Rate	44.36%
														CPS-VS Resp Employment Status Rate	44.65%
														Overall Relative Bias	-0.65%

Table 3: CPS Disability Status Rate by Regression Tree End Nodes and Response Status

Node	Resp Status	Count	Dis Rate	Dis Count	Dis Rate LL	Dis Rate UL	NR - RESP	NR - RESP LL	NR - RESP UL	Chi Square	P Value	Significant	Overall Dis Rate	Overall - Resp	Dis Est Error
1	NR	667	23.24%	155	20.03%	26.44%	-2.51%	-5.91%	0.89%	1.98	0.16	0	25.49%	0.26%	4.19
	RESP	5,658	25.75%	1,457	24.61%	26.89%									
7	NR	49	22.45%	11	10.77%	34.13%	-2.55%	-21.57%	16.46%	0.07	0.79	0	23.46%	1.54%	0.29
	RESP	32	25.00%	8	10.00%	40.00%									
8	NR	250	28.80%	72	23.19%	34.41%	2.41%	-4.02%	8.85%	0.56	0.46	0	26.98%	-0.59%	-1.61
	RESP	758	26.39%	200	23.25%	29.52%									
10	NR	78	21.79%	17	12.63%	30.96%	-6.11%	-22.35%	10.13%	0.57	0.45	0	23.97%	3.94%	1.14
	RESP	43	27.91%	12	14.50%	41.31%									
11	NR	236	25.00%	59	19.48%	30.52%	-0.30%	-7.20%	6.61%	0.01	0.93	0	25.19%	0.11%	0.18
	RESP	423	25.30%	107	21.15%	29.44%									
13	NR	41	39.02%	16	24.09%	53.96%	16.95%	-0.63%	34.52%	3.81	0.05	0	27.97%	-5.89%	-1.94
	RESP	77	22.08%	17	12.81%	31.34%									
14	NR	38	23.68%	9	10.17%	37.20%	-0.13%	-22.81%	22.56%	0	0.99	0	23.73%	0.08%	0.01
	RESP	21	23.81%	5	5.59%	42.03%									
16	NR	35	20.00%	7	6.75%	33.25%	-7.78%	-32.35%	16.79%	0.41	0.52	0	22.64%	5.14%	0.62
	RESP	18	27.78%	5	7.09%	48.47%									
17	NR	10	10.00%	1	0.00%	28.59%	-9.23%	-30.69%	12.23%	0.49	0.48	0	17.74%	1.49%	0.16
	RESP	52	19.23%	10	8.52%	29.94%									
18	NR	37	18.92%	7	6.30%	31.54%	-11.85%	-33.62%	9.92%	1.18	0.28	0	23.81%	6.96%	1.04
	RESP	26	30.77%	8	13.03%	48.51%									
		8,549		2,183										Total Error	4.10
														CPS-VS Full Sample Disability Status Rate	25.58%
														CPS-VS Resp Disability Status Rate	25.54%
														Overall Relative Bias	0.19%

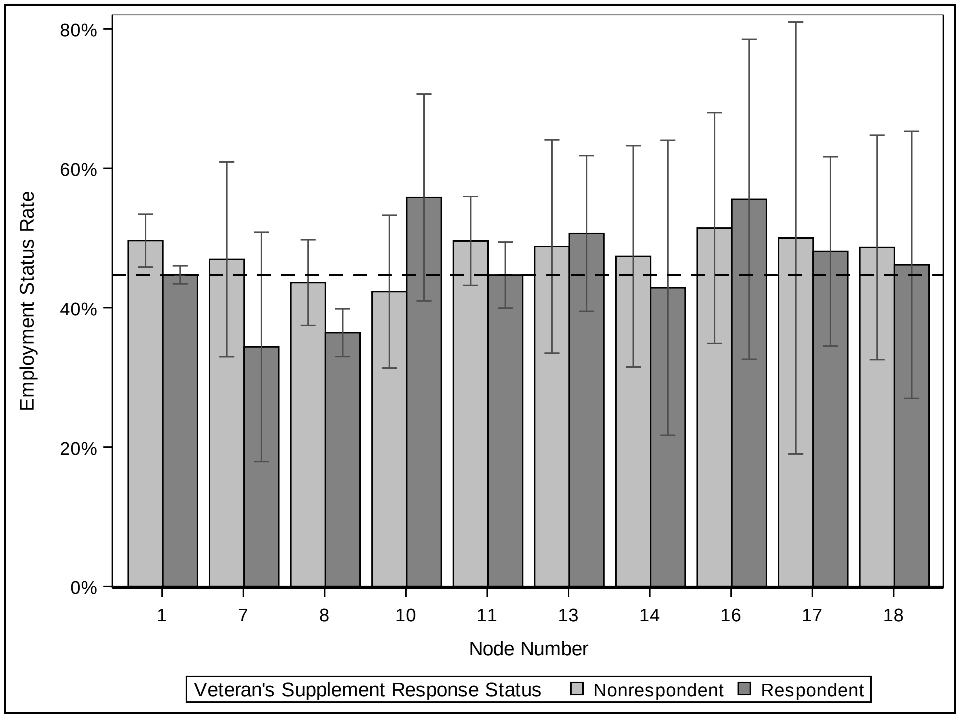

Figure 2: CPS Employment Status Rate by Node for CPS-VS Respondents vs. Nonrespondents

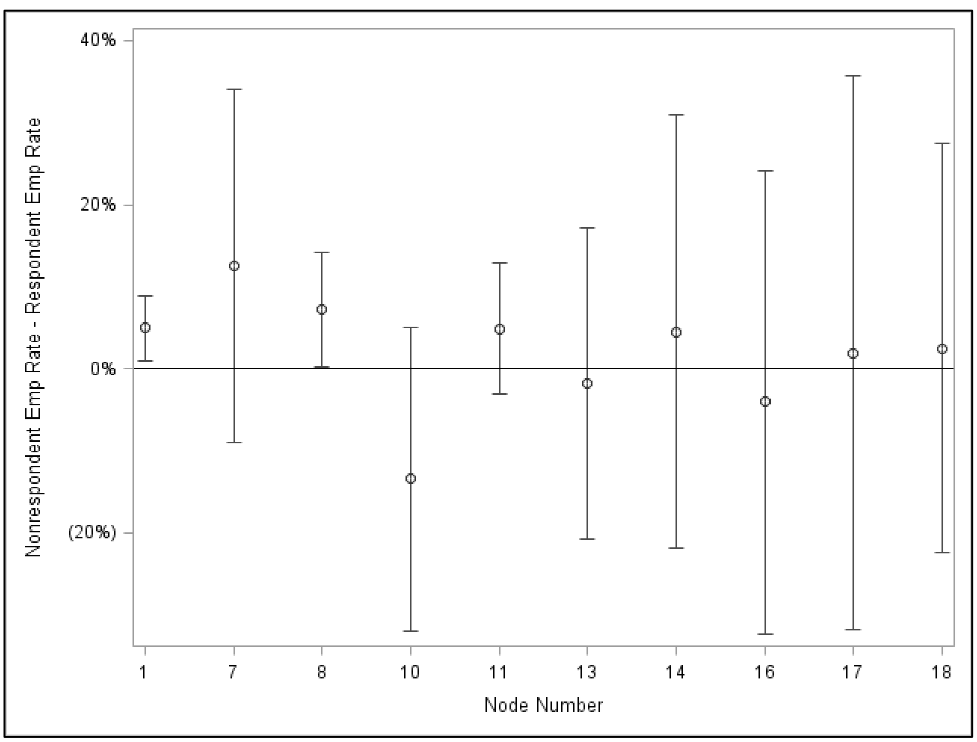

Figure 3: CPS Employment Status Rate Difference 95% Confidence Intervals for CPS-VS Nonrespondents vs. Respondents by Node

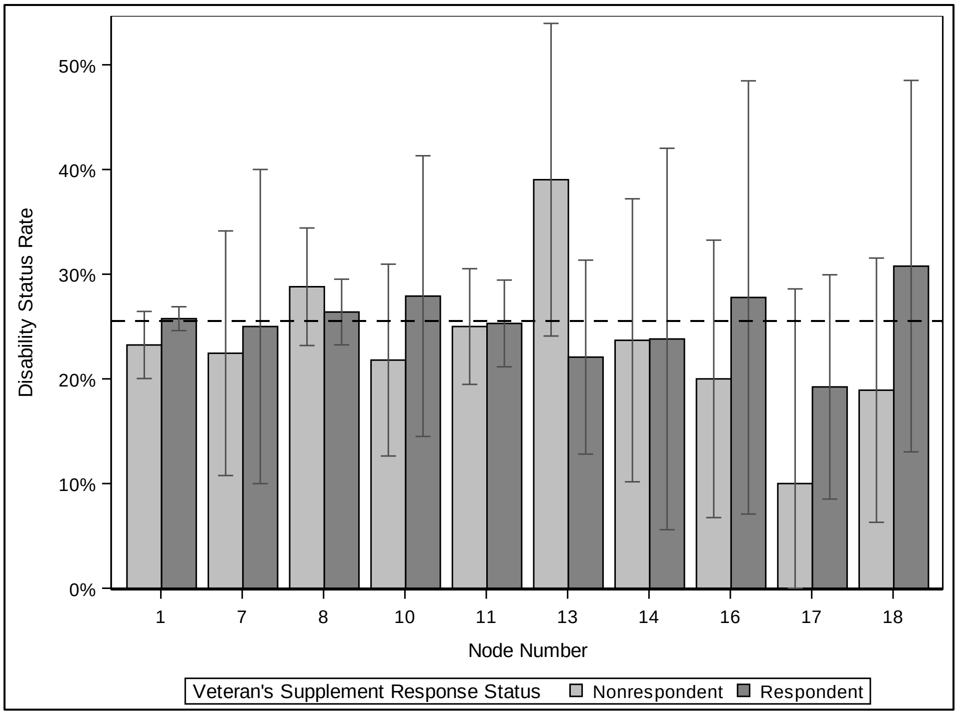

Figure 4: CPS Disability Status Rate by Node for CPS-VS Respondents vs. Nonrespondents

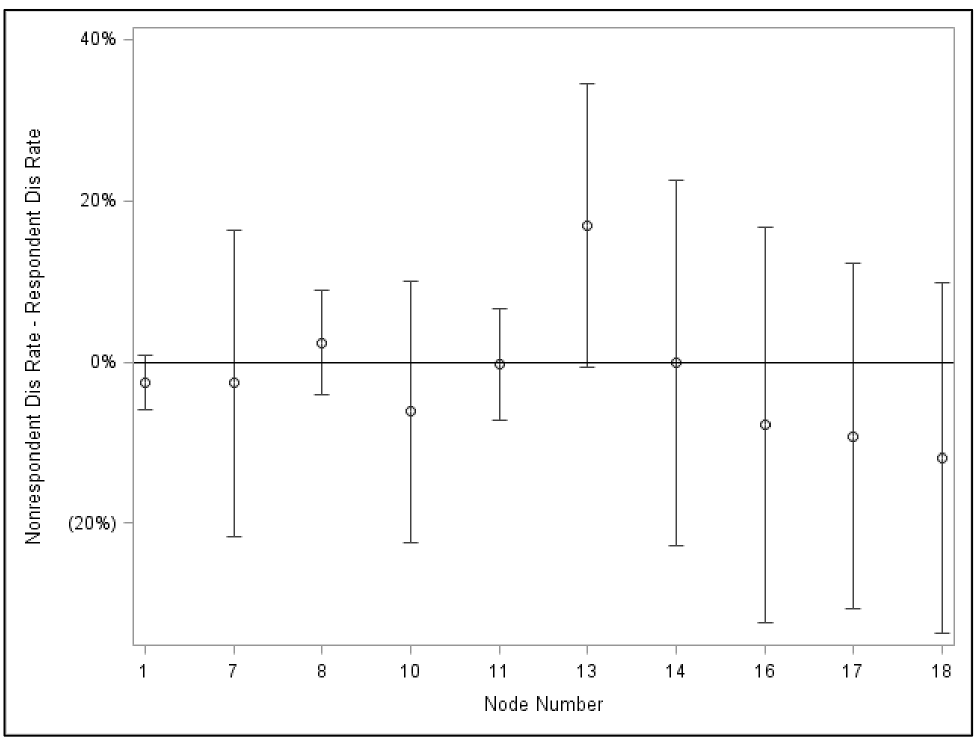

Figure 5: CPS Disability Status Rate Difference 95% Confidence Intervals for CPS-VS Nonrespondents vs. Respondents by Node

5. References

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences (revised ed.),* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, R., and Rodriguez, R. N. (2013) "High-performance statistical modeling." In *Proceedings of the SAS Global Forum 2013 Conference, Cary, NC: SAS Institute Inc.*, 401-2013.

deVille, B. (2006). *Decision Tress for Business Intelligence and Data Mining using SAS Enterprise Miner*. Carey, NC: SAS Institute, Inc.

SAS. "SAS/STAT 14.1 User's Guide The HPSPLIT Procedure." *Cary, NC: SAS Institute Inc.* 2015. https://support.sas.com/documentation/onlinedoc/stat/141/hpsplit.pdf

U.S. Bureau of Labor Statistics. "Current Population Survey, August 2016 Veterans Supplement Technical Documentation." (2016). https://www2.census.gov/programs-surveys/cps/techdocs/cpsaug16.pdf

U.S. Census Bureau. "Current Population Survey Supplement." DataFerrett, 2017. https://thedataweb.rm.census.gov/ftp/cps_ftp.html#cpssupps