Nonresponse Bias Measurement and Adjustment in the Follow-up Study of a National Cohort of Gulf War and Gulf War Era Veterans (Wave 3)

Erin K. Dursa, Ph.D., MPH[1], Aaron Schneiderman, Ph.D., MPH[1], Heather Hammer, Ph.D.[2], Stanislav Kolenikov, Ph.D.[2]
[1]Post Deployment Health, Office of Public Health, Department of Veterans Affairs, Washington DC, 20420
[2] Abt SRBI, Inc., Silver Spring, MD 20910, USA

Abstract

The Follow-up Study of a National Cohort of Gulf War and Gulf War Era Veterans is a multimode web, mail, and CATI survey. The original cohort for this longitudinal survey was comprised of 15,000 deployed Gulf War veterans and 15,000 non-deployed Gulf War Era veterans surveyed in 1995-1997 and then again in 2005. For the Wave 3 follow-up survey conducted in 2013, we used response propensity modeling (multiple logistic regression analysis) to examine the nonresponse (attrition) mechanism. Our assessment shows that the nonresponse weight adjustment via calibration successfully removed about 80% of the nonresponse bias in the test variable - marital status in 1991 as recorded in the frame data. It also significantly reduced the nonresponse bias initially observed in three of the five key Wave 3 survey variables examined: Chronic Multisymptom Illness (CMI), smoking, and alcohol use. However, as evidenced by significant correlations between the examined outcomes and response propensities, we were unable to reduce the nonresponse bias for two of the five key variables: PTSD symptoms and general health. The negative correlation of PTSD symptoms with response propensity indicates that respondents tend to have lower PTSD symptom screening scores (i.e., screen negative for PTSD) compared with nonrespondents. The positive correlation of response propensity with general health indicates that respondents tend to have higher self-reported general health compared with nonrespondents. This paper describes the methodology used in the nonresponse bias analysis and discusses the performance of the nonresponse correction and its meaning for the results.

Key Words: non-response, attrition, propensity, weight calibration, veteran health, military health, sequential multimode survey design

1. Background

The Follow-up Study of a National Cohort of Gulf War and Gulf War Era Veterans is the third wave of the National Health Survey of Gulf War Era Veterans and Their Families. Initiated by the U.S. Department of Veterans Affairs (VA) in 1995, the survey was designed as a retrospective cohort study that compared health indicators for a population-based sample of 15,000 troops deployed to the Persian Gulf area (Gulf War veterans) with 15,000 troops not deployed to the Persian Gulf Area (Gulf War Era veterans) (Kang et al. 2009). The cohort was surveyed again in 2005 (Wave 2), and the Follow-up Study

of a National Cohort of Gulf War and Gulf War Era veterans (Wave 3) was fielded between May 21 and September 2, 2013. All non-deceased, non-incarcerated members of the original cohort were eligible to participate in Wave 3 regardless of their participation in the prior two waves.

The Wave 3 survey data were collected with a sequential multimode design that started with a mailed invitation to complete the survey by Web, followed by a mail version of the questionnaire, followed by telephone. The original sample provided by the VA was comprised of 15,000 deployed Gulf War veterans and 15,000 non-deployed Gulf War Era veterans stratified by branch of service (Army, Air Force, Navy, and Marines) and component type (Active Duty, Reserve, or National Guard). Women were oversampled to comprise 20% of the original sample. After deleting volunteer study participants, the corrected original sample had 29,993 veterans.

Wave 3 refusal, cooperation and response rates (AAPOR 2015) show that even though finding and contacting cohort members was challenging (AAPOR CON3=54%), 93% of those contacted were willing to participate (AAPOR COOP1) and only 4% refused (AAPOR REF3). While the Wave 3 response rate was 50% (AAPOR RR5, not adjusted for prior waves), there was substantial variation in the response propensities of different subgroups of veterans, and this differential attrition suggests the potential for nonresponse bias. Of the 14,252 completed interviews, 9,643 (67.7%) were completed by mail, 3,808 (26.7%) on the Web, and 801 (5.6%) by telephone (computer-assisted telephone interview, CATI).

We used response propensity modeling (multiple logistic regression analysis) to examine the Wave 3 attrition mechanism. Our assessment shows that the nonresponse weight adjustment via calibration successfully removed about 80% of the nonresponse bias in the test variable (marital status in 1991 as recorded in the frame data. We also provide indirect evidence that the nonresponse adjustment significantly reduced the nonresponse bias initially observed in three of the five key Wave 3 estimates examined: Multisymptom Illness (CMI), smoking, and alcohol use. However, as evidenced by significant correlations between the examined outcomes and response propensities, we were unable to reduce the nonresponse bias for two of the five key estimates: positive screen for PTSD symptoms, and self-reported general health. This paper describes our nonresponse bias analysis and results.

2. Methods

The Follow-up Study of a National Cohort of Gulf War and Gulf War Era veterans covered a wide range of domains including: general health, health risk behaviors, medical health conditions, life experiences and daily activities, and socio-demographic characteristics. The Pilot Study was conducted with a sample of 500 veterans selected from the original cohort to identify necessary modifications to the questionnaire and to assess the mailing process and compatibility of the mail and Web versions. A total of 220 Pilot Study interviews (44% of 500) were completed; 62 (22.8%) by Web and 158 (78.2%) by mail. The final mail questionnaire was 16 pages and included 306 items. The invitation letter was mailed to all but the 1,126 members of the original cohort who were determined to be deceased prior to the start of data collection. The carefully timed contact protocol sequence for the Main Study included mailing an advance letter, a reminder letter, and as many as three scannable paper versions of the questionnaire and three thank

you/reminder postcards as needed prior to initiating the CATI nonresponse follow-up. In the nonresponse follow-up, non-refusing veterans who did not complete a Web or mail questionnaire were contacted by telephone to complete the survey using Computer Assisted Telephone Interviewing (CATI) administered by a trained interviewer upon obtaining oral consent. Veterans who completed the Web survey were mailed a $10 incentive check. Those who did not complete by Web received a $10 pre-paid incentive check with the first mailing of the questionnaire, and those who completed by CATI were mailed an additional $10 incentive check. Telephone follow-up conducted with the 14,139 nonrespondents to the Web and mail surveys yielded 801 CATI interviews. After removing the 194 duplicates who completed the survey in more than one mode, the final sample size was 14,252 including 220 surveys completed in the Pilot Study and 14,032 completed in the Main Study (Table 1).

Table 1, Sample Frame Characteristics by Wave 3 Survey Completion.

		Sample		Completed Interviews	
Category		Count	Column %	Count	RR1
Gender	Male	23,984	80.0%	11,358	47.4%
	Female	6,009	20.0%	2,894	48.2%
Deployed	Yes	15,225	50.8%	8,104	53.2%
	No	14,768	49.2%	6,148	41.6%
Branch	Air Force	3,468	11.6%	1,819	52.5%
	Army	19,214	64.1%	9,170	47.7%
	Marine	3,364	11.2%	1,438	42.7%
	Navy	3,947	13.2%	1,825	46.2%
Component	Active Duty	11,996	40.0%	5,567	46.4%
	Reserve	9,997	33.3%	4,792	47.9%
	National Guard	8,000	26.7%	3,893	48.7%
Total		29,993	100%	14,252	47.5%

2.1 Weighting and Variance Estimation

The Wave 3 weights were constructed to: (1) reflect the stratified sampling design used to draw the original sample in Wave 1, (2) adjust for nonresponse encountered in the field, and (3) correct known inconsistencies in the original frame information. The same set of weighting steps and procedures was applied to define the weights for each of Waves 1 and 3, as well as the panel weights needed to compare and measure change between Waves 1 and 3. The weighting proceeded in four steps: constructing probability weights; correcting misclassified deployment status in the frame count; weight calibration; and creating bootstrap replicate weights for variance estimation.

First, the probability weights were computed as the ratio of the frame counts (provided in Lee, Mahan, and Kang 2002) and sample counts, using the original, uncorrected frame information. The probability weights, defined for all of the 29,993 valid cases in the sample, have mean 49.87 and coefficient of variation 0.853. These weights reflect the balancing of deployed and non-deployed veterans, proportional representation of the branches and types of service, and oversampling of female veterans through the full factorial cross-classification of the 40 strata used in the original sampling design.

Whereas the probability weights reflect deployment status when the original sample was drawn, updated information indicated that some of the sampled veterans were misclassified in the frame. In the second step, the original frame counts were corrected for the veterans found to be misclassified in the sample.

In the third step of the weight adjustment that implicitly corrects for nonresponse, the weights were calibrated so that the weighted totals of the 40 stratification cells with deployment corrected, as well as rank, race, and age group summed up to the population totals (Kolenikov 2014). The frame-based calibration variables with totals estimated from the full sample were age group, rank and race. These variables were selected for calibration because they were highly significant predictors in the propensity model for the probability of completing a survey in Wave 3. While the response propensity models included other statistically significant predictors such as the interaction of service type with deployment status and branch, we followed Eltinge's (2002) advice to use the simpler weighting method for reasons of estimator stability and a general preference for parsimony. Doing so also allowed us to use marital status, a frame variable available for the full sample, to indicate the magnitude of nonresponse bias in the responding sample, and assess the extent to which the nonresponse bias was reduced by weight calibration.

To account for the sample design that included deep stratification, as well as nonresponse adjustments through weight calibration, in the fourth and final step, bootstrap replicate weights were created to facilitate correct variance estimation (Kolenikov 2010). To reduce the risk of encountering empty cells and perfect prediction in logistic regression modeling with replicate weights that may arise due to zero replicate weights, the bootstrap weights were averaged across three subsequent replicates (Yung 1997). The sampling variability in the estimated control totals was incorporated into replicate variance estimation. For each bootstrap replicate, the control totals were re-estimated based on the bootstrap sample drawn, and these re-estimated control totals were used in weight calibration, the third stage of weight adjustment. Compared to the bootstrap standard errors, the standard errors based only on the main weights and sampling strata were, on average, biased upwards by 28% because they did not account for the efficiency gains afforded by weight calibration.

Table 2. Marital Status in 1991: Point Estimates and 95% Confidence Intervals for the Frame, and Wave 3 Completes with Probability vs. Nonresponse Adjusted Weights

Type of Estimate	Married	Single	Other + Unknown
Frame	0.5412 (0.5340, 0.5485)	0.4228 (0.4156, 0.4300)	0.036 (0.0336, 0.0386)
Wave 3 Completes, Probability Weights	0.5971 (0.5866, 0.6076)	0.3673 (0.3571, 0.3777)	0.0355 (0.0322, 0.0393)
Wave 3 Completes, Nonresponse Adjusted Weights	0.5525 (0.5433, 0.5616)	0.4159 (0.4068, 0.4251)	0.0316 (0.0285, 0.0350)

As previously noted, we used frame marital status to indicate the magnitude of nonresponse bias and assess the extent to which weight calibration reduced the bias. Table 2 compares the weighted marital status estimates with the frame data. The table shows that married veterans were over-represented among the Wave 3 respondents. Applying the nonresponse adjusted weights removed about 80% of bias in the marital

status estimate, reducing the reported percent of the 1991-married Wave 3 respondents from 59.71% using probability weights only (5.59% difference from the frame value, significant at 5% level) to 55.25% using the nonresponse adjusted weights (1.13% difference from the frame value; (5.59-1.13)/5.59=79.8%). The nonresponse adjusted weights also worked well for the single (never married) category estimate. While it moved in the wrong direction for the Wave 3 estimate of veterans with other and unknown marital status, the confidence intervals for the frame, probability weighted, and nonresponse adjusted estimates all overlap meaning that the performance difference is not significant. This is likely explained by the weak (or lack of) association between other and unknown marital status and the frame variables.

3. Nonresponse Analysis

Nonresponse bias is a function of both the nonresponse rate and the difference between respondents and nonrespondents on the statistic of interest (National Research Council 2013). Several studies have shown that a low response rate may not yield high nonresponse bias if the difference between respondents and nonrespondents on the statistic of interest is small or ignorable in a statistical sense (Curtin, Presser, and Singer 2000; Keeter et al. 2000). Nevertheless, in the context of steadily declining response rates, differential nonresponse can be a problem (Link and Burks 2013), particularly among subgroups that tend to be underrepresented in general population surveys such as single-parent households, families with young children, and Latinos (Groves 2006; National Research Council 2013). This pattern suggests that the potential for nonresponse bias should be examined and addressed if there is any concern about the representativeness of the sample respondents.

Nonresponse bias is specific to a particular estimate (or model) and not to the survey in general. Whereas nonresponse adjusted weights can reduce or ameliorate the effects of attrition bias, the effectiveness of this approach can vary between surveys (Cellini et al. 2008) and among different estimates within a single survey. Dillman et al. (2014) explain that weighting on a characteristic can only ensure that the sample is representative with respect to that characteristic and other variables strongly correlated with it. Weighting will not ensure the representativeness of variables not correlated with the characteristic used for weighting. As a result, within the same survey, weighting can improve some estimates, have no effect on others, and potentially even harm others.

Although response and attrition rates are commonly used measures of data quality in panel surveys, neither is directly linked to bias and both can be poor predictors of it. While there is no comprehensive statistical theory of nonresponse bias, there is a large and growing body of research on nonresponse bias and how to mitigate its consequences. Statistical methods for dealing with nonresponse bias represent several approaches including: alternative nonresponse weighting adjustments (e.g., propensity models; selection models; calibration, including post-stratification and raking; pattern mixture models) (National Research Council 2013), balance indicators (B-indicators) (Särndal 2011), and representativity indicators (R-indicators) (Cobben and Schouten 2007).

For this study, we developed a response propensity model to examine the attrition mechanism between Waves 1 and 3, and to calculate the nonresponse adjustment. The advantage of using the propensity model approach is its ability to accommodate a large number of variables and the simplicity of it application (Hazelwood et al. 2007). We also

implemented a new test (Hammer et al. 2013) of the difference of two weighted estimates that are based on two different weights applied to the same variable. Finally, we assessed the effectiveness of the nonresponse adjustment by examining the success of the bias correction for marital status in 1991 (as shown above) and five key Wave 3 estimates: CMI, PTSD symptom screen, smoking, and alcohol use. In this assessment, we determined where the adjustment succeeded or failed to reduce a statistically significant bias to a non-significant difference. Note that the question used to measure CMI was added after the Pilot Study. In the nonresponse analysis, Pilot Study respondents with missing data on this variable were treated as "missing by design."

3.1 Nonresponse Bias Diagnostics and Mitigation

Prior to conducting the response propensity analysis, we wanted to determine if the sample distribution of Wave 3 respondents adequately matched the original sample distribution on the available frame variables. We did this by tabulating the frame variables with the probability sampling weights applied to the entire original sample (N=29,993) and the Wave 3 completes (N=14,252). In each comparison, we computed the standard error of the difference between the estimates, and used a z-test (equivalent to the test of the differences between respondents and nonrespondents) to determine if the difference was statistically significant at $p<.05$ two-sided. In this analysis, statistically significant differences indicate differential nonresponse between population groups and the potential for nonresponse bias in the Wave 3 estimates.

Our comparison of the frame variable distributions is presented in four blocks of columns in Table 3. The leftmost block identifies the tabulations of the frame variables for the entire original sample (N=29,993). The zero standard errors on branch and service type reflect the fact that these are stratification variables, and there is no sampling uncertainty about their population counts. While gender and deployment status are also stratification variables, they exhibit non-zero standard errors due to misclassification. The other frame variables with non-trivial standard errors, including rank, race, age, and marital status, reflect our lack of knowledge of their population counts.

The second block in the middle of Table 3 provides the tabulation of the frame variables on the subsample of 14,252 Wave 3 respondents. Because the Wave 3 respondents are a pseudo-random subsample of the entire sample, all of the Wave 3 estimates have non-zero standard errors. As indicated by the statistically significant differences between the frame and Wave 3 proportions, this block shows that the following groups were over-represented among Wave 3 respondents: Air Force and Army vs. Marines and Navy; Deployed vs. Non-deployed; Older (age 33+ in 1991, i.e., age 54+ in 2012) vs. Younger; Officer vs. Enlisted; White vs. Black; and Married in 1991 vs. Single in 1991. If the health behaviors and outcomes of interest differ between these categories, then analyzing the data with just the probability weights is likely to lead to nonresponse biases. The differences in service type and sex were not statistically significant.

The next column gives the p-value of the z-test for differences between the Wave 3 respondents and nonrespondents. While a direct comparison of respondents and nonrespondents to establish the risks and magnitudes of nonresponse biases is only possible on the frame variables, an indirect comparison is possible when the health behaviors and outcomes of interest are presented with probability weights and nonresponse adjusted weights. We present the indirect comparison analysis in two parts.

Table 3. Distributions of the Sample and Respondents on Selected Frame Variables

Variable	Category	Estimate using probability weights, entire sample			Estimate using probability weights, Wave 3 completes			p-value of difference	Estimate using NRA weights	
		Estimate	Std. error	Count	Estimate	Std. error	Count		Estimate	Std. error
Branch	Air Force	11.74%	0.00%	3,468	12.93%	0.26%	1,819	< 0.001	11.74%	0.00%
	Army	51.82%	0.00%	19,214	52.66%	0.41%	9,170	0.041	51.82%	0.00%
	Marine	14.85%	0.00%	3,364	13.75%	0.30%	1,438	< 0.001	14.85%	0.00%
	Navy	21.59%	0.00%	3,947	20.66%	0.35%	1,825	0.009	21.59%	0.00%
Type of Service	Active	77.47%	0.00%	11,996	77.57%	0.23%	5,567	0.644	77.47%	0.00%
	Reserve	14.42%	0.00%	9,997	14.21%	0.18%	4,792	0.253	14.42%	0.00%
	Guard	8.12%	0.00%	8,000	8.21%	0.11%	3,893	0.393	8.12%	0.00%
Female (updated)		10.21%	0.03%	6,009	10.30%	0.19%	2,894	0.637	10.16%	0.00%
Deployed (updated)		47.58%	0.18%	15,225	51.87%	0.45%	8,104	< 0.001	46.39%	0.00%
Age Group in 1991	17-25	43.34%	0.37%	11,977	36.22%	0.53%	4,680	< 0.001	43.09%	0.41%
	26-32	28.48%	0.34%	8,227	28.83%	0.51%	3,855	0.341	28.73%	0.39%
	33-39	16.80%	0.28%	4,859	20.14%	0.45%	2,684	< 0.001	16.80%	0.31%
	40+	11.39%	0.21%	4,903	14.82%	0.35%	3,019	< 0.001	11.38%	0.21%
Rank in 1991	Enlisted	86.49%	0.25%	25,949	82.49%	0.42%	11,832	< 0.001	86.49%	0.23%
	Officer	12.24%	0.24%	3,696	15.67%	0.40%	2,194	< 0.001	12.24%	0.23%
	Warrant	1.27%	0.09%	348	1.83%	0.15%	226	< 0.001	1.27%	0.08%
Race	Black	22.27%	0.31%	6,804	18.32%	0.42%	2,702	< 0.001	22.27%	0.30%
	Hispanic	4.71%	0.16%	1,372	4.43%	0.23%	594	0.101	4.71%	0.18%
	Other	3.95%	0.15%	1,007	3.80%	0.22%	440	0.355	3.98%	0.17%
	White	68.97%	0.35%	20,760	73.37%	0.49%	10,489	< 0.001	68.97%	0.34%
Marital status in 1991	Married	54.12%	0.37%	15,147	59.71%	0.54%	7,986	< 0.001	55.25%	0.46%
	Other + unknown	3.60%	0.13%	1,487	3.55%	0.18%	766	0.749	3.16%	0.16%
	Single	42.28%	0.37%	13,359	36.73%	0.53%	5,500	< 0.001	41.59%	0.46%
Age in 1991	Treated as continuous	28.866	0.053	29,966	30.185	0.083	14,238	< 0.001	28.9	0.053

First, the last two columns of Table 3 tabulate frame variables for the Wave 3 respondents using the nonresponse adjusted weights. Although these nonresponse adjusted estimates are defined for the same sample as the estimates for the 14,252 Wave 3 respondents in the middle of the table, ideally, they should yield results that are identical or close to the estimates for the entire sample in the leftmost part of the table.

The zero standard errors on branch, service type, sex, and deployment status indicate that the weights were calibrated to match the known population totals for these variables. Age group (albeit defined somewhat differently with six percentile categories vs. the four analytical categories used in Table 3), rank, and race were also used as calibration variables. This explains why their proportions match the entire sample proportions in the leftmost block.

Weight calibration generally has several effects on the estimates. First, nonresponse biases may be reduced as shown in marital status, where calibration weighting removed 80% of nonresponse bias in the married category. Specifically, the over-representation of veterans who were married in 1991 among Wave 3 respondents was reduced from 59.7% with probability weights to 55.2% with nonresponse adjusted weights compared to the population target of 54.1%. The second effect of calibration (and, for that matter, of many other weight adjustments) is an increase in the variability of weights that leads to an increase in the standard errors. The third effect of calibration weighting, and its original motivation, is *reducing* standard errors (Deville and Särndal, 1992). As shown in Table 3, all of the standard errors in the third block on the right side of the table are uniformly smaller than those in the middle block with probability weights. To capture the efficiency gains of calibration, one needs to use either specialized estimation procedures that explicitly take calibration into account (currently available in R and SUDAAN), or implicitly account for that through replicate weights (available more widely in SAS, R, Stata, SUDAAN and WesVar). Our implementation of variance estimation with calibrated weights, done in Stata, relies on bootstrap replicate weights (Kolenikov 2010).

The second part of our qualitative analysis of the differences between probability weights and nonresponse adjusted weights is provided in Table 4. The variables included in Table 4 are: three self-reported health outcomes (general health, PTSD symptom screen, and CMI), self-reported demographic information (education and income) and two self-reported health-related behaviors (alcohol use and smoking). While no statistical tests are available to compare the estimates, all of the nonresponse adjustments are in the expected direction with respect to increasing the representation of less educated and lower income veterans by the nonresponse adjusted weights.

Table 4. Selected Health Outcomes and Other Survey Variables Estimated with Probability and Nonresponse Adjusted Weights

	Probability Weights		NRA Weights	
Variables	Proportion	std. error	Proportion	std. error
General Health				
1 Poor	5.52%	0.25%	5.38%	0.26%
2 Fair	21.63%	0.46%	21.69%	0.50%
3 Good	36.55%	0.54%	36.88%	0.55%
4 Very Good	26.23%	0.49%	26.37%	0.55%
5 Excellent	10.07%	0.33%	9.68%	0.33%
Positive PTSD Symptom Screen	15.32%	0.40%	15.86%	0.41%
Chronic Multisymptom Illness	31.52%	0.51%	31.18%	0.51%
Education				
1 HS or below	16.89%	0.41%	17.87%	0.43%
2 Some college or associate	44.76%	0.55%	46.19%	0.52%
3 Bachelor's degree	19.76%	0.43%	19.56%	0.43%
4 Graduate or professional degree	18.60%	0.43%	16.37%	0.36%
Income				
Less than $20,000	7.26%	0.28%	7.92%	0.30%
$20,000 - $34,999	11.01%	0.35%	11.25%	0.34%
$35,000 - $49,999	13.78%	0.38%	14.09%	0.40%
$50,000 - $74,999	22.69%	0.47%	22.82%	0.44%
$75,000 - $99,999	16.13%	0.41%	16.20%	0.43%
$100,000 or more	29.14%	0.51%	27.72%	0.50%
Alcohol				
Never drink	18.77%	0.43%	18.82%	0.45%
0 - 4 drinks per week	50.73%	0.56%	50.83%	0.57%
5 - 10 drinks per week	16.25%	0.42%	15.91%	0.41%
11or more drinks per week	10.26%	0.34%	10.30%	0.34%
Smoking				
Never smoked	45.48%	0.56%	46.07%	0.57%
Not in past 12 months	33.30%	0.53%	32.32%	0.52%
0 - 9 cigarettes per day	6.01%	0.27%	6.42%	0.29%
10 - 19 cigarettes per day	6.93%	0.29%	7.17%	0.30%
20 or more cigarettes per day	8.27%	0.32%	8.03%	0.33%

3.2 Response Propensity Modeling

As the preliminary step in analyzing the historic differences in the response propensities of the different groups of veterans, we defined three historic response propensity groups based on the response patterns in the first two waves conducted in 1995 and 2005 respectively. The lowest response propensity category is comprised of veterans who did not complete the survey in either Wave 1 or Wave 2; the middle category includes those who responded only once (in either Wave 1 or Wave 2 but not both); and the highest response propensity category is comprised of the veterans who responded in both Waves 1 and 2. The decision not to differentiate between veterans who responded in Wave 1 but not Wave 2 vs. those who responded in Wave 2 but not Wave 1 was based on the results of exploratory analyses indicating that these two groups of veterans had similar frame, health, and demographic profiles as well as similar response propensities.

The following groups of veterans were found to be more likely to respond to prior waves: Air Force; Guard; deployed; age 33 or older in 1991; Officer or Warrant rank in 1991; White; and married in 1991. Those less likely to respond in prior waves were: Navy; not deployed; age 17-25 in 1991; enlisted in 1991; Black; Hispanic; and single in 1991. These findings are very similar to those reported in Table 3 with respect to Wave 3 participation. The similarity suggests that the response patterns exhibited in the two prior waves were repeated in Wave 3.

The full response propensity model was obtained by searching for the best complex-design corrected AIC (Lumley and Scott 2015), and is reported in Table 5. This model is based on frame variables only, and the complex design features accounted for are limited to stratification and unequal probabilities of selection. While age, race, rank in 1991, marital status in 1991, and sex all have significant coefficients, the attrition groups are the strongest predictors. The model fits the data well as evidenced by the non-significant Archer-Lemeshow goodness-of-fit test.

Table 5. Final Response Propensity Model for Wave 3

Frame Variables	Categories	Odds Ratio	Standard Error
Race	Black (vs. White)	0.6456	0.0255**
	Hispanic (vs. White)	0.8478	0.0633*
	Other/unknown (vs. White)	0.8039	0.0668**
Gender	Male (vs. female)	0.7597	0.0461**
Age Group in 1991	26-32 (vs. 17-25)	1.0881	0.0889
	33-39 (vs. 17-25)	1.5172	0.1580**
	40+ (vs. 17-25)	2.0269	0.2625**
Age x Gender Interaction	26-32 (vs. 17-25), male	1.2293	0.1113*
	33-39 (vs. 17-25), male	1.2398	0.1405
	40+ (vs. 17-25), male	1.0679	0.1454
Rank in 1991	Officer (vs. enlisted)	1.3454	0.0675**
	Warrant (vs. enlisted)	1.6604	0.2506**
Marital Status in 1991	Other + unknown (vs. married)	0.7391	0.0584**
	Single (vs. married)	0.8913	0.0342**
Deployment Status	Not deployed (vs. deployed)	0.7026	0.0275**
Branch	Air Force (vs. Army)	0.9543	0.0593
	Marine (vs. Army)	0.8848	0.0515*
	Navy (vs. Army)	0.8189	0.0397**
Type of Service	Reserve (vs. Active Duty)	1.0689	0.0522
	Guard (vs. Active Duty)	1.0400	0.0519
Branch x Type of Service Interaction	Air Force, Reserve	1.1188	0.1165
	Air Force, Guard	1.0488	0.0965
	Marine, Reserve	1.0444	0.0876
	Navy, Reserve	1.2074	0.1046*
Deployment Status x Type of Service Interaction	Non-deployed, Reserve	0.7488	0.0427**
	Non-deployed, Guard	0.7261	0.0447**
	Archer-Lemeshow fit p-value	0.9105	

Notes: * $p<0.05$; ** $p<0.01$. The pool of variables also included interaction of deployment with age and with branch. Updated versions of deployment status and gender used in this analysis.

One well-respected methodology for quantifying nonresponse bias risk comes from the correlation analysis of response propensities and outcomes. For instance, when an outcome is positively correlated with response propensity, then the higher levels of the outcome will be overrepresented in the responding sample, and the sample estimates of the mean outcome will be biased upwards. Following Witt (2010), we analyzed both Pearson moment correlations (with the interpretation offered above) and semi-partial correlations of the outcome and response propensity residuals, where the latter characterize the remaining unmodeled issues in survey response.

Table 6 reports the results for the Pearson correlations for the frame variables. The semi-partial correlations are not reported because all of the frame variables are used in the final propensity model, and as a result, their semi-partial correlations with response propensity (i.e., correlations between variables and propensity residuals) are zero. For respondents, the propensity residuals are always equal to 1 (response propensity), therefore, the semi-partial correlation is simply the negative of the Pearson correlation. Although the significant Pearson correlations of the frame variables with response propensity in Table 6 indicate the potential for nonresponse bias, with the exception of marital status, all of these variables were used in the nonresponse weight adjustment via calibration; therefore, the nonresponse biases in these variables were removed.

Table 6. Pearson Correlations for Response Propensity with the Frame Variables Included in the Full Response Propensity Model

Frame Variable and Category	Pearson Correlation	Std. error	z-statistic	p-value
Branch = Air Force	0.156	0.034	4.537	0.0000
Branch = Army	0.070	0.034	2.041	0.0413
Branch = Marines	-0.129	0.031	-4.113	0.0000
Branch = Navy	-0.095	0.035	-2.758	0.0058
Type = Active	0.010	0.021	0.499	0.6175
Type = Reserve	-0.024	0.021	-1.155	0.2481
Type = Guard	0.015	0.016	0.939	0.3477
Gender = Male	-0.012	0.027	-0.435	0.6634
Gender = Female	0.012	0.027	0.435	0.6634
Deployed	0.362	0.034	10.513	0.0000
Not Deployed	-0.362	0.034	-10.513	0.0000
Marital status in 1991= married	0.472	0.030	15.567	0.0000
Marital status in 1991= other/unknown	-0.010	0.033	-0.302	0.7628
Marital status in 1991= single	-0.472	0.030	-15.964	0.0000
Race = Black	-0.400	0.029	-13.789	0.0000
Race = Hispanic	-0.056	0.033	-1.727	0.0841
Race = Other/unknown	-0.037	0.037	-1.010	0.3123
Race = White	0.401	0.031	12.872	0.0000
Age in 1991 (continuous variable)	0.680	0.021	32.952	0.0000

Note: Semi-partial correlations of propensity residuals are zero as the propensity model includes all frame variables. Updated versions of deployment status and gender used in this analysis.

Table 7 reports the magnitude and significance of the correlations of response propensities with outcomes. This is the key determinant of the nonresponse bias in Bethlehem's (2002) stochastic response model. The table shows that nonresponse bias

has been significantly reduced for CMI, smoking, and alcohol use. However, as evidenced by significant correlations of outcomes and response propensities, we were unable to reduce the nonresponse bias in the PTSD symptom screen and general health estimates. The negative correlation of the PTSD screening outcome with response propensity indicates that Wave 3 respondents were less likely to screen positive for PTSD symptoms compared with nonrespondents. The positive correlation of response propensity with general health indicates that Wave 3 respondents tended to have higher self-reported general health compared to nonrespondents.

Table 7. Pearson Correlations for Response Propensity with the Health Outcome and Related Survey Variables Included in the Full Response Propensity Model

Health Outcomes and Related Variables	Pearson Correlation	Std. error of correlation	z-statistic	p-value
Positive screen for PTSD symptoms	-0.076	0.013	-5.709	0.000
Chronic Multisymptom Illness	-0.008	0.017	-0.506	0.613
General health	0.044	0.017	2.562	0.010
Smoking status	-0.010	0.015	-0.681	0.496
Alcohol use	-0.003	0.011	-0.234	0.815
Education	0.278	0.024	11.741	0.000
Income	0.168	0.018	9.440	0.000

Note: Semi-partial correlations for respondents are the reverse sign of the Pearson correlations and have identical standard errors/significance.

3.3 Magnitude of Nonresponse Biases

The degree of nonresponse bias can be quantified in terms of the relationship between correlation and regression. If x is the explanatory variable in a simple bivariate regression and y is the dependent variable, then the correlation coefficient ρ and regression coefficient β are related as

$$\hat{\beta} = \hat{\rho}\frac{S_y}{S_x}$$

where S_x and S_y are the standard deviations of the respective quantities. Considering unit response as the explanatory variable, the weighted proportion of Wave 3 respondents is equal to 0.4622, leading to $S_x = 0.4986$. Similarly, the weighted standard deviation of screening positive for PTSD symptoms among Wave 3 respondents is equal to 0.3653, and 1.0322 for general health. Hence the estimated regression coefficients in the regression of these outcomes on response are -0.0547 and 0.0956.

These regression coefficients reflect the change in the probability of screening positive for PTSD symptoms, and the average value on the general health scale, respectively, when the explanatory variable, unit response, changes from 0 to 1 (i.e., from a unit nonrespondent to a completed survey). Thus we can estimate the influence of nonresponse bias and multiply by the nonresponse rate to remove the influence of the bias in the positive PTSD symptom screen prevalence estimate and the self-reported general health mean. Compared to the average positive PTSD symptom screen prevalence of 15.86% (with a standard error of 0.41%), the nonrespondents may have prevalence rates that are higher by about 5.5%. This leads to the nonresponse bias

corrected estimate of PTSD prevalence equal to 18.8%. Also, the general health mean for nonrespondents is about 0.096 lower when compared to the estimated general health overall mean of 3.133 (with a standard error of 0.012), leading to the nonresponse bias corrected estimate of 3.081.

The implication of these detected nonresponse biases is that the survey data on these two variables should be used with caution for analyses that are not limited to comparisons of the levels of prevalence across groups with similar demographic compositions and similar response rates where the biases can be expected to cancel each other out. Estimates of the overall prevalence of a positive PTSD symptom screen in any particular group are at risk of nonresponse biases unless the group exhibits negligible variability of response propensity, or has a negligible nonresponse rate as one special case of the no variability condition. In contrast, CMI is not significantly correlated with unit response propensity in Table 9. Therefore, analyses of this health outcome are not subject to limitations related to unit nonresponse biases.

Discussion

The Follow-up Study of a National Cohort of Gulf War and Gulf War Era Veterans has several strengths. To date, it is the largest and longest running prospective cohort of Gulf War and Gulf War Era (comparison population) veterans. The study has produced numerous publications, and is the source of much of what is known about the health conditions affecting Gulf War veterans (Kang et al. 2002; Wallin et al. 2009; Coughlin et al. 2011; Toomey et al. 2009; Kang et al. 2000). The sample was carefully selected to reflect the population of veterans who served in the military during the Gulf War with respect to demographic and military characteristics (Kang et al. 2000). The longitudinal design and favorable response after 20 years allows for the study of disease development over time and the role of deployment related exposures (collected at baseline) in disease development. The non-deployed Gulf War Era veterans provide an appropriate comparison group for determining the extent to which Gulf War deployment impacts the long term health of those who served beyond what is expected due to aging. All outcomes were self-reported which can introduce bias; however, medical records validation of a sample of respondents found 86% agreement between conditions that the veteran reported in the survey and what was documented in the veteran's medical record.

Few longitudinal studies of veterans have published robust nonresponse analyses. Investigators from the Millennium Cohort Study, the largest longitudinal population based study of active duty service members and veterans, used propensity models derived from their baseline data to determine predictors of response to the first follow up (Littman et al. 2010). Consistent with the results of our study, older age, female sex, officer rank, and ever married, were associated with higher likelihood of response at the first follow-up. Participants with a history of smoking, chronic alcohol consumption, major depressive disorder (no association observed for history of PTSD) and those who had separated from service between the baseline assessment and the first follow up were less likely to respond. In contrast to our nonresponse analysis, the Millennium Cohort Study propensity scores and nonresponse weighting suggest that nonresponse does not considerably influence the health outcome estimates. The authors also found that weighting for nonresponse had little impact on the distribution of self-reported health at follow-up, implying that self-reported health did not predict response.

Our paper describes a statistical analysis of nonresponse and nonresponse bias in the Follow-Up Study of a National Cohort of Gulf War and Gulf War Era veterans, the third wave of a large population based longitudinal health. To assess the extent of nonresponse bias and develop nonresponse bias adjustments to mitigate it, we used frame marital status in 1991 (Wave 1) as a test variable to determine the extent of nonresponse bias in the Wave 3 estimates. The nonresponse adjusted weights eliminated 80% of the bias in this estimate. We applied this same method to five important outcomes of interest: CMI, smoking status, alcohol use, PTSD symptom screen, and self-reported general health. We found that nonresponse bias was significantly reduced for CMI, smoking status, and alcohol use. We were not able to achieve this for the PTSD symptom screen or self-reported health status estimates, as there were significant correlations between these outcomes and response propensities.

References

Bethlehem, J. G. 2002. Weighting Nonresponse Adjustments Based on Auxiliary Information. Ch. 18, pp. 275–288, in: *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. Eltinge and R. Little. John Wiley and Sons, New York..

Cellini, S. R., S. M. Mckernan and C. Ratcliffe. 2008. The dynamics of poverty in the United States: A review of data, methods and findings. *Journal of Policy Analysis and Management* **27**: 577-605.

Cobben, F. and B. Schouten. 2007. An Empirical Evaluation of R-Indicators. *Discussion Paper* 08006. Voorburg/Heerlen: Statistics Netherlands. Available: http://www.risq-project.eu/papers/ cobben-schouten-2008-a.pdf [Retrieved January 2013].

Coughlin, S.S., H.K. Kang, C.M. Mahan. 2011. Alcohol use and selected heath conditions of 1991 Gulf War Veterans: survey results, 2003-2005. *Prev Chronic Dis*. **8**(3):A52.

Curtin, R., S. Presser and E. Singer. 2000. The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly* **64**: 413–428.

Deville, J. C. and C. E. Sarndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** (418): 376-382.

Dillman, D. A., Smyth, J. D. and L. M. Christian. 2014. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.

Eltinge, J. 2002. Diagnostics for the Practical Effects of Nonresponse Adjustment Methods. Ch. 29 in: *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. Eltinge and R. Little. John Wiley and Sons, New York.

Groves, R. M. 2006. Nonresponse rates and nonresponse error in household surveys. *Public Opinion Quarterly* **70**: 646-675.

Hammer, H., S. Kolenikov, R. Martonik, C. Sabina and C. A. Cuevas. 2013. Adjusting for attrition bias in a national longitudinal survey of dating violence among Latino youth. *Paper presented at the 69th Annual Conference of the American Association of Public Opinion Research (AAPOR)*, Anaheim, CA. May 15.

Hazelwood, L. N., T. I. Mach and J. D. Wolken. 2007. Alternative Methods of Unit Nonresponse Weighting Adjustments: An Application from the 2003 Survey of Small Business Finances. *Finance and Economics Discussion Series* 2001-10. Washington, DC: Federal Reserve Board.

Kang, H. K., C. M Mahan, K. Y Lee, C. A Magee and F. M. Murphy. 2000. Illnesses among United States veterans of the Gulf War: a population-based survey of 30,000 veterans. *Journal of Occupational and Environmental Medicine* **42**(5):491-501.

Kang, H. K., C. M. Mahan, K. Y. Lee, F. M. Murphy, S. J. Simmens, H. A. Young and P. H. Levine. 2002. Evidence for a deployment related Gulf War Syndrome by factor analysis. *Archives of Environmental Health* **57**(1):61-68.

Kang, H. K., B. L. MA, C. M. Mahan, S. A. Eisen and C. C. Engel. 2009. Health of US veterans of 1991 Gulf War: A follow-up survey in 10 years. *Journal of Environmental Medicine* **51** (4): 1-10.

Keeter, S., C. Miller, A, Kohut, R. M. Groves and S. Presser. 2000. Consequences of reducing nonresponse in a telephone survey. *Public Opinion Quarterly* **64**: 125-48.

Kolenikov, S. 2010. Resampling variance estimation for complex survey data. *The Stata Journal* **10**(2): 165-199.

Kolenikov, S. 2014. Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal* **14**(1): 22–59.

Lee, K. Y., C. M. Mahan and H. K. Kang. 2002. Sampling and Non Response in the National Health Survey of Gulf War-Era Veterans and Their Families. Department of Veterans Affairs, Washington, DC.

Link, M.W. and A. T. Burks. 2013. Leveraging auxiliary data, differential incentives, and survey mode to target hard-to-reach groups in an address-based sampling design. *Public Opinion Quarterly* **77**(3): 696-713.

Littman, A. J., E. J. Boyko, I. G. Jacobson, J. Horton, G. D Gackstatter, B. Smith, T. Hooper, T. S. Wells, P. J. Amoroso and T. C. Smith. 2010. Assessing non-response bias at follow-up in a large prospective cohort of relatively young and mobile military service members. *BMC Research Methodology* **10**:99.

Lumley, T., and A. Scott. 2015. AIC and BIC for Modeling Complex Survey Data. *Journal of Survey Statistics and Methodology*, **3** (1), 1–18.

National Research Council. 2013. Tourangeau, R. and T. J. Plewes (Editors). *Nonresponse in Social Science Surveys: A Research Agenda*. Panel on a Research Agenda for the Future of Social Science Data Collection, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Särndal, C. 2011. The 2010 Morris Hansen Lecture dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics* **27**:1-21.

The American Association for Public Opinion Research. 2015. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 8th edition. AAPOR.

Toomey, R., R. Alpern, J. J. Vasterling, D. G. Baker, D. J. Reda, M. J. Lyons, W. G. Henderson, H. K. Kang, S. A. Eisen and F. M. Murphy. 2009. Neurological functioning of U.S. Gulf War veterans 10 years after the war. *Journal of the International Neuropsychological Society* **15**(5):717-729.

Wallin, M. T., J. Wilken, M. H. Alfaro, C. Rogers, C. M. Mahan, J. C. Chapman, T. Fratto, C. Sullivan, H. Kang and R. Kane. 2009. Neuropsychological assessment of a population-based sample of Gulf War veterans. *Cognitive and Behavioral Neurology*. **22**(3):155-166.

Witt, M. B. 2010. Estimating the R-indicator, its standard error and other related statistics with SAS and SUDAAN. *Proceedings of Survey Research Methods Section*. Alexandria, VA: American Statistical Association https://www.amstat.org/sections/srms/Proceedings/y2010/Files/309481_61666.pdf

Yung, W. 1997. Variance estimation for public use files under confidentiality constraints. In *Proceedings of the Survey Research Methods Section*: 434-439. American Statistical Association, Alexandria, VA.