Composite Response Rates for Surveys with Nonresponse Follow-up

A.C. Singh, J.A. Dever, and V.G. Iannacchione
RTI International

Abstract
The double-sampling estimator of response rates for surveys with a nonresponse follow-up, although unbiased, may be unstable if the follow-up survey is based on a small sample size. In this paper, we show how the dual-frame calibration idea of Singh and Wu (1996, 2003) can be used to obtain a more precise composite response rate (CRR). The CRR combines the unstable but unbiased double-sampling estimator of response rate with the stable but biased estimator of response rate from the main survey after being corrected for under-reporting of response in the main survey. The latter estimator is biased due to under-reporting as it does not use the nonrespondents who could have been converted to respondents if the follow-up survey were administered to them. To this end, the framework of a population response model to define appropriate finite population parameters corresponding to estimated response rates is invoked. The data from the 1991 Gulf War Veterans Survey is used to illustrate the proposed method of CRR.

Key Words: Response rate as a finite population parameter; Response under-reporting; Follow-up surveys; Dual-frame calibration

1. INTRODUCTION

For surveys with nonresponse follow-up, the estimated response rate (unweighted or weighted) based on the combined survey is, by definition, greater than the corresponding estimated response rate from the main survey. This observation leads to some interesting questions: What are the corresponding parameters? Do they bear the same relationship as the estimates? These questions can be addressed by considering the difference between the corresponding finite population parameters under a population response model. Under this model, specific to the survey instrument and conditions, each unit in the population is assigned a random response indicator of 1 or 0 before actual sampling takes place; see e.g., Fay (1991). Denoting the main survey by A and the combined survey by B, and the two finite population total parameters as T_r^* and T_r where r denotes the response indicator, we have

T_r^* = Total number of individuals in the population (typically corresponding to some domain of interest) who would respond if the main survey were administered to the whole population (1.1a)

T_r = Total number of individuals who would respond to the main survey plus the total number who would respond to the follow-up among the nonrespondents to the main survey. (1.1b)

It follows that the response rates satisfy the relation (where N denotes the population size),

$$N^{-1}T_r^* \le N^{-1}T_r \qquad (1.2)$$

Let the two samples be denoted respectively by s_A consisting of selected units from the main survey and s_B consisting of respondents from the main survey and the selected units from the nonrespondents in the follow-up. Note that the sample size for s_B is smaller than that for s_A. Also let d_{kA} and d_{kB} denote respectively the design weights for the two samples, where

$$d_{kB} = \begin{cases} d_{kA} & \text{if } r_A=1 \\ d_{kA}d_{kF} & \text{if } r_B=1 \end{cases}$$

The last component (d_{kF}) denotes the weight due only to the follow-up sample. Now the weighted response totals $\hat{T}^*_{r(A)}$ and $\hat{T}_{r(B)}$ (a double-sampling estimator) provide design-unbiased estimates of the corresponding parameters T_r^* and T_r. They are defined as

$$\hat{T}^*_{r(A)} = \sum_{s_A} d_{kA} r_{kA}, \quad \hat{T}_{r(B)} = \sum_{s_B} d_{kB} r_{kB} \qquad (1.3)$$

If the finite population parameters defined by (1.2) are of interest, then we need weighted response rates for unbiased estimation. If the design is ignorable or, if the response rate for the given sample is of interest, then the unweighted response rate will suffice.

Suppose the parameter of interest is T_r and not T_r^* because the parameter T_r^* is believed to suffer from under-reporting in the main survey. In other words, some nonrespondents in the main survey could be converted to respondents if they were followed-up. Now for estimating T_r, the unbiased estimator $\hat{T}_{r(B)}$ is typically not precise enough because of the small sample size of the follow-up. Therefore, we turn to $\hat{T}^*_{r(A)}$ for extra information which is somewhat similar to an interesting application of the Dual-Frame Calibration (DFC) idea of Singh and Wu

(1996, 2003) and by Singh, Iannacchione, and Dever (2003) in the context of efficient estimation for surveys with a nonresponse follow-up. The estimator $\hat{T}^*_{r(A)}$ is stable but known to be biased for T_r. We need to correct it for bias due to under-reporting , and then combine it with $\hat{T}_{r(B)}$ to get a composite response rate (CRR) estimator, denoted by $N^{-1}\hat{T}_{r(A\|B)}$. To achieve this goal, we first adjust the estimator $\hat{T}^*_{r(A)}$ by modeling the under-reporting bias via a weight adjustment factor, and then propose a method for computing the CRR estimator as an "optimal" combination of the two estimators such that poststratfication controls to various subpopulation counts and set of zero controls representing new auxiliary information in terms of selected study variables are satisfied. The optimality here does not refer to the usual one in regression estimation. Instead it refers to the optimal choice of a parameter representing relative effective sample size in the calibration approach.

Section 2 provides a motivation of the proposed method of CRR estimation. **Section 3** presents a description of the CRR method. **Section 4** presents an application to the 1991 Gulf War Veterans Survey. Finally, concluding remarks are given in **Section 5**.

2. MOTIVATION FOR THE COMPOSITE RESPONSE RATE ESTIMATOR

In the application of dual-frame estimation to estimate response rates for surveys with nonresponse follow-up, the two frames are identical and the corresponding estimated response rate totals $\hat{T}^*_{r(A)}$ and $\hat{T}_{r(B)}$ are dependent. Note that $\hat{T}^*_{r(A)}$ is a biased estimate for the parameter T_r of interest due to under-reporting, a new complication in the context of the usual problem of dual-frame estimation. However, there are some simplifying features of the current problem as well. For instance, since the only study variable of interest in this case corresponds to the response indicator (r) within various domains, there is no need for any adjustment for nonresponse bias. Moreover, our main interest is to obtain efficient estimates of response rates for various domains, and not to produce a set of final calibrated weights that could be used to calculate response rates for arbitrary domains. Therefore, we optimize the combining factor separately for each domain of interest.

To correct first for the under-reporting bias from sample *A*, we propose to use the generalized exponential model (GEM) of Folsom and Singh (2000) for poststratification of all selected units except that bounds are set differently. However, for sample *B*, the usual poststratification is first performed before combining it with the sample *A* estimator. More specifically, for sample *A* respondents, we needed to adjust the initial weights upwards to account for the under-reporting bias and so we set l_{kA} =1 and u_{kA} =10 (say). We chose the value for the center parameter c_{kA} as the ratio of the two overall estimated response rates, $(N^{-1}\hat{T}_{r(B)})\big/(N^{-1}\hat{T}^*_{r(A)})$ which represents a global adjustment to reach an unbiased estimator $\hat{T}_{r(B)}$. For the sample *A* non-respondents, we needed to adjust the weights downwards and, therefore, we set l_{kA} =0.1 (say) and u_{kA}=1 while c_{kA} was set to $(1-N^{-1}\hat{T}_{r(B)})\big/(1-N^{-1}\hat{T}^*_{r(A)})$, the ratio of the two overall estimated nonresponse rates. For sample *B*, we set bounds as in the case of usual post-stratification, i.e., l_{kA} =0.1 (say), u_{kA} =10 (say), and c_{kA}= 1.

Now to combine the two estimates (one is bias adjusted and the other is poststratified) while accounting for possible different designs, we perform a DFC-type poststratification by concatenating the two samples (see also Singh, Iannacchione, and Dever (2003) using GEM. In DFC with GEM with pre-specified parameters:

$$\ell_{kA} < c_{kA} < u_{kA} \quad \text{and} \quad \ell_{kB} < c_{kB} < u_{kB},$$

the adjustment factors are modeled as

$$a_{kA} = \frac{\ell_{kA}(u_{kA}-c_{kA}) + u_{kA}(c_{kA}-\ell_{kA})\exp_A}{(u_{kA}-c_{kA}) + (c_{kA}-\ell_{kA})\exp_A},$$
$$a_{kB} = \frac{\ell_{kB}(u_{kB}-c_{kB}) + u_{kB}(c_{kB}-\ell_{kB})\exp_B}{(u_{kB}-c_{kB}) + (c_{kB}-\ell_{kB})\exp_B}$$

where

$$\exp_A = \exp[\eta_A^{-1} A_{kA}(\mathbf{x}'_{kA}\boldsymbol{\lambda}_{xA} + r'_{kA}\boldsymbol{\lambda}_z)],$$
$$A_{kA} = \frac{(u_{kA}-\ell_{kA})}{(u_{kA}-c_{kA})(c_{kA}-\ell_{kA})}$$

The corresponding expressions for sample *B* are defined analogously. The parameters η_A, representing the effective sample size of design *A* with respect to design *B*, and η_B (=1-η_A) are assumed to be given a priori based on historical data. In practice, they may be specified by a grid search such that the variance of the estimator for a given reporting domain (r-variable) is minimized. Note that the specification of this parameter takes account of the dependence between the two estimators if any. In addition to the usual calibration equations corresponding to subpopulation counts for the auxiliary x-variables, new calibration equations in the form of zero controls for the r-variables are solved to estimate the GEM model parameters (λ's). The zero control simply forces the difference between the two total estimators for a number of selected r-variables to zero after calibration. This is the key idea underlying the DFC methodology of Singh and Wu (1996, 2003).

It may be noted that for the key study variables (r) used in the zero controls, both samples give rise to identical estimates, and the composite estimator remains the same for arbitrary combining factors ζ_A and ζ_B (= 1 - ζ_A) for estimators from the two frames. That is

$$\hat{T}^{DFC}_{r(A\|B)} = \hat{T}^{DFC}_{r(A)} = \hat{T}^{DFC}_{r(B)}$$

where

$$\hat{T}^{DFC}_{r(A\|B)} = \zeta_A \hat{T}^{DFC}_{r(A)} + (1-\zeta_A)\hat{T}^{DFC}_{r(B)}, \qquad (2.2)$$

$$\hat{T}^{DFC}_{r(A)} = \sum_{s_a} r_{kA} w_{kA}, \text{ and } \hat{T}^{DFC}_{r(B)} = \sum_{s_B} r_{kB} w_{kB}. \qquad (2.3)$$

The weights (w) denote the final calibrated weights for the two samples s_A and s_B.

For other arbitrary study variables (r) which are not part of the zero controls, a suitable combining factor common for all variables can be obtained from considerations based on minimizing the variance approximated under a model as in Kish's use of unequal weighting effect. We use the combining factor given by (2.4)

$$\zeta_A = \left(\sum_{s_B} w_{kB}^2 - \sum_{s_A} w_{kA} w_{kB} \right) \Bigg/ \left(\sum_{s_A} w_{kA}^2 + \sum_{s_B} w_{kB}^2 - 2\sum_{s_A} w_{kA} w_{kB} \right)$$

3. COMPOSITE RESPONSE RATE ESTIMATION

3.1 Point Estimation

First the relative effective sample size parameters η_A and η_B (=1 – η_A) are specified in the interval (0,1) for a grid search. Then the CRR estimator is obtained in three steps.

Step I: This is somewhat similar to the post-stratification adjustment with respect to covariates known for all units in the sampling frame, and is in the same spirit as the Hajek-adjustment to reduce the impact of extreme weight values except that for sample *A* it is different from the usual post-stratification because bounds for respondents and nonrespondents in sample *A* are chosen such that adjustment for under-reporting bias is simultaneously performed while satisfying the usual controls. This step is performed for each sample via GEM using bounds as given in Section 2.

Step II: This is also somewhat like post-stratification in that zero control totals correspond to the selected set of *r*-variables, and the subpopulation counts correspond to *x*-variables which subsume the covariates used in Step I above.. (Here the two samples A and B are concatenated first and then their weights are calibrated together unlike the previous step.)

Step III: For each value of η_A chosen on the grid (0,1) in increments of 0.05, the CRR estimate for each *r*-variable is obtained. Next, specific to each *r*-variable, an optimal η_A is selected such that the variance is minimized.

3.2 Variance Estimation

As described in Singh, Iannacchione, and Dever (2003), variance of $\hat{T}_{r(A\|B)}$ can be estimated by Taylor linearization. Since both $\hat{T}_{r(A)}$ and $\hat{T}_{r(B)}$ are identical and equal to $\hat{T}_{r(A\|B)}$, their variances are also identical. Alternatively, jackknifing can be used to estimate the variance by repeated use of DFC for pseudo-replicates. In computing the jackknife variance of $\hat{T}_{r(A\|B)}$, we need a combining parameter ζ_A to compute the composite estimate within each stratum/ replicate combination because the two estimates ($\hat{T}_{r(A)}$ and $\hat{T}_{r(B)}$) are not necessarily equal at the stratum level by construction. They are however equal at the overall level by construction. In contrast, for $\hat{T}_{r(A)}$ and $\hat{T}_{r(B)}$, we do not need the combining parameter ζ_A in using jackknife separately. This suggests that the jackknife variance of $\hat{T}_{r(A\|B)}$ should be invariant to the choice of ζ_A which, however, needs further investigation.

4. APPLICATION

The Tenth Anniversary Gulf War Veterans Health Survey (GWHS) is a national probability-based survey of men and women who served in the 1991 Persian Gulf War within all branches of the U.S. Armed Forces. The primary objectives of the study are (1) to provide national estimates of Gulf War veterans who report significant health concerns and (2) to model the key correlates of those health concerns. Other objectives include comparisons between active-duty military and reservists, and the development of separate explanatory models for the occurrence of health concerns in male and female veterans. The objective of the sample design for this study was the selection of a probability sample of veterans from the target population of sufficient size to support these analytic objectives.

The *target population* for the GWHS is the more than 685 thousand men and women who served in the 1991 Persian Gulf War with all branches of the U.S. Armed Forces. We selected a stratified systematic sample of 10,301 veterans from the sampling frame maintained by Defense Manpower Data Center. We defined four primary strata by subdividing active-duty military and reservists by gender. Within each primary stratum, veterans who had registered with Department of Defense's Gulf War Comprehensive Clinical Evaluation Program (CCEP) and received a medical diagnosis based on International Classification of Diseases, 9^{th} Revision were over-sampled to obtain a sufficient number of veterans reporting significant health concerns. Additionally, the frame was sorted by race/ethnicity to ensure a representative sample.

The survey originally was implemented as a mail survey in 2001. An overall response rate of 54.4 percent (RR3 definition from AAPOR 2004) was achieved after three mailings of the instrument, as well as a reminder post card, and a reminder telephone call. Response rates to the mail survey were highest among females, reservists, and those who had been evaluated by the CCEP.

The response rate to the mail survey was 20 percentage points lower than expected. In an effort to reduce the potential bias associated with nonresponse to the mail survey, the project team decided to conduct a telephone follow-up of a sub-sample of nonrespondents to the mail survey. We based the follow-up sub-sample size of 1,000 mail nonrespondents (about one-fifth of all mail nonrespondents) on funding available to the study.

We allocated the follow-up sample inversely proportional to the mail response rates of each stratum. Prior to selection, each mail nonrespondent was classified as probable 'easy to contact' or 'difficult to contact' based on whether an interviewer had made contact with someone in the veteran's household during calls made to prompt the return of the mail survey. Mail nonrespondents classified as 'easy to contact' were over-sampled to increase the expected effective sample size of the follow-up. To decrease response burden, the telephone follow-up obtained information on 69 of the 151 questions included in the mail survey.

As **Table 1** illustrates, the unweighted (i.e. RR3) response rates to the Gulf War survey do not account for differential selection rates that were used to select the sample. These unweighted response rates sometimes can lead to biased conclusions about the response propensity of the surveyed population. For example, the 54.4 percent response rate (using the AAPOR RR3) to the mail portion of the survey was almost seven percentage points higher than the corresponding design-weighted response rate (WRR) of 47.6 percent. The disparity was the result of over-sampling females and reservists, two groups who responded at noticeably higher rates than males and active-duty personnel. For the telephone follow-up, we achieved a 55.1 percent response rate (AAPOR RR3) which was noticeably lower than the WRR of 50.0 percent because we over-sampled mail nonrespondents with updated contact information. The response patterns for the follow-up were similar to the mail survey although the largest increase in response rate occurred among active-duty males not evaluated by the CCEP.

The WRR estimates the response rate that would be obtained if everyone in the target population were selected for interview. In addition, treating the response rate as a population parameter enables us to calculate the sampling variance of the WRR in a design-consistent fashion. Associating a measure of precision with response rates provides important information about their reliability and provides a statistical basis for comparing response rates among important subgroups.

A total of 5,709 eligible sample members responded to either the mail survey or the telephone follow-up, another 441 were found to be ineligible (i.e., non Gulf War veterans); and the remaining 4,151 were nonrespondents. If the RR3 formulation is used, the overall response rate for the survey is 59.7 percent. However, the RR3 formulation is misleading because only a subsample of initial nonrespondents was selected for follow-up. In this situation, the WRR in the form of a double-sampling estimator (DSE) provided a design-consistent estimator of the combined response to the initial survey and the follow-up.

While the usual double-sampling estimator provides an unbiased estimate of the combined response rate, it may be unstable especially if the follow-up sub-sample is small. Again using the survey of Gulf War veterans as an example, the double-sampling estimator of the combined WRR among female reservists was 81.2 percent compared to a combined WRR of 74.4 percent among females on active duty. However, the 6.8 percentage point difference was not statistically significant because of the large sampling error associated with the DSE estimators.

We used dual-frame calibration (DFC) to develop a composite estimator of the overall response rate for surveys with a nonresponse follow-up. In our case, the composite estimator is a weighted combination of two correlated estimators. The first estimator is a weighted mean of a zero/one initial response indicator assigned to all sample members. This estimator is expected to be stable but biased because it ignores the responses obtained from the follow-up. The second estimator is the typical double-sampling estimator that combines the initial responses with those of the follow-up. This estimator is expected to be relatively unbiased but unstable especially if the follow-up sample is small. We used the steps described in **Appendix A** to develop a composite response rate (CRR) that strikes a balance between variance and bias.

We used a grid search of the relative effective sample size parameters η_A and η_B (=1–η_A) in the interval (0,1) to determine the optimal (i.e., minimum variance) CRR. A comparison of weighted response rates using the design weights and the post-stratified weights is shown along with the CRRs for various study domains in **Table 2**. In general, the CI half-widths associated with the CRRs are smaller than those associated with the post-stratified DSEs reflecting the improved efficiency of dual-frame calibration.

5. SUMMARY AND DISCUSSION

For surveys with a nonresponse follow-up, a composite response rate (CRR) estimator (a type of dual-frame calibration estimator) was proposed to improve the efficiency of the usual double-sampling estimator which is a weighted response rate. Using the population response model, we demonstrate that the weighted response rate provides an unbiased estimate for non-ignorable designs unlike the unweighted response rate obtained using the AAPOR RR3 Definition.

An important underlying notion associated with calculating a CRR for a survey with a nonresponse follow-up is that response is under-reported in the initial or first phase of the survey. For the Gulf War survey, we assumed that under-reporting occurred because at least some of the sample members who did not respond by mail were likely to respond to the telephone follow-up. This notion of under-reported response can be extended to mixed-mode surveys. For example, assume that, instead of selecting a sub-sample of 1,000 initial (mail) nonrespondents to the Gulf War Survey, we select an *independent* sample of 1,000 Gulf War sample members for telephone interview. In this case, the overall response propensity is still under-reported in the initial (i.e., mail) sample because (by the same model) there are sample members who are unlikely to respond by mail but will respond by phone. However, under-reporting of response also can be assumed in the sample selected for telephone interview. That is, there are at least some sample members who are unlikely to respond by telephone but will respond to the mail. Here the issue of households with regular mailing address but without landline telephone or households with telephones but without regular mailing addresses can be dealt with by assuming that each household can be interviewed in either mode, at least conceptually.

This under-reporting of response in both samples motivates the use of weight adjustment factors that are constrained to be greater than or equal to one for respondents (because they are under-represented in each sample), and less than one for nonrespondents. An interesting implication of this constraint is that the CRR could be greater than either of the individual response rates. For example, if the response rate to the mail portion was 50 percent and the response rate to the telephone portion was 35 percent, the CRR could exceed 50 percent. (The actual magnitude of the CRR would depend on what adjustment factors are needed to satisfy the specified control totals.) Ultimately, a CRR that exceeds the RR of either sample may be reasonable, especially if one is willing to accept the notion that the overall response propensity is under-reported in each sample.

6. REFERENCES

American Association for Public Opinion Research (2004), *Standard Definitions: Final Dispositons of Case Codes and Outcome Rates for Surveys. 3rd edition*. Lenexa, Kansas: AAPOR.

Fay, R.E. (1991), "A Design-Based Perspective On Missing Data Variance," *Proceedings of the Annual Research Conference, US Bureau of the Census*, pp. 429-440.

Folsom Jr., R.E., and Singh, A.C. (2000), "A Generalized Exponential Model for Sampling Weight Calibration for a Unified Approach to Nonresponse, Post-stratification, and Extreme Weight Adjustments," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 598-603.

Singh, A.C., and Wu, S. (2003), "An Extension of Generalized Regression Estimator to Dual-Frame Surveys," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 3911-3918.

Singh, A.C., and Wu, S. (1996), "Estimation for Multiframe Complex Surveys by Modified Regression," *Proceedings of the Statistical Society of Canada, Survey Methods Section*, pp. 69-77.

Singh, A.C., Iannacchione, V.G., and Dever, J.A. (2003), "Efficient Estimation for Surveys with Nonresponse Follow-Up Using Dual-Frame Calibration," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 3919-3930.

Table 1. Comparison of Weighted and Unweighted Response Rates

	Initial Survey			Follow-Up Survey		
Study Domain	**Sample Size**	**RR3 (%)**	**WRR*(%)**	**Sample Size**	**RR3 (%)**	**WRR*(%)**
Overall	10,301	54.4	47.6 ± 2.9	1,000	55.1	50.0 ± 3.5
Active Males	4,489	50.4	45.9 ± 3.6	472	52.3	47.7 ± 5.2
Active Females	2,311	54.7	53.5 ± 6.0	218	54.6	47.6 ± 7.8
Reserve Males	2,310	58.4	53.3 ± 4.0	208	58.7	53.9 ± 7.9
Reserve Females	1,191	60.9	60.0 ± 5.3	102	61.8	58.1 ± 8.9
Males	6,799	53.1	47.0 ± 3.1	680	54.3	49.6 ± 4.4
Females	3,502	56.8	55.4 ± 4.5	320	56.9	50.8 ± 6.1
Active Duty	6,800	51.8	46.4 ± 3.4	690	53.0	47.7 ± 4.3
Reserves	3,501	59.2	54.1 ± 3.5	310	59.7	55.2 ± 6.0
On CCEP	4,274	59.9	58.2 ± 1.7	372	57.8	53.5 ± 5.4
Not on CCEP	6,027	50.5	47.0 ± 3.0	628	53.5	48.0 ± 4.6
NH Black	3,062	42.8	31.6 ± 3.6	364	51.1	46.4 ± 5.5
NH White	6,104	60.0	53.2 ± 3.0	512	55.3	50.3 ± 4.9
Other Race	1,135	55.1	46.4 ± 4.9	124	66.1	60.1 ± 9.4

*Weighted response rates and 95% confidence intervals calculate using design weights.

Table 2. Comparison of Weighted Response Rates*

	Design Weights		Post-Stratified Weights*		DFC Weights	
Study Domain	**Initial Survey**	**Initial and FU Surveys**	**Initial Survey**	**Initial and FU Surveys**	**CRR**	η_A
Overall	47.6 ± 2.9	72.6 ± 3.5	71.7 ± 2.4	72.6 ± 3.5	71.9 ± 2.4	0.80
Active Males	45.9 ± 3.6	71.4 ± 4.3	70.0 ± 3.1	71.4 ± 4.3	70.3 ± 2.9	0.80
Active Females	53.5 ± 6.0	74.4 ± 6.4	77.3 ± 4.3	74.4 ± 6.4	77.1 ± 4.4	0.95
Reserve Males	53.3 ± 4.0	77.4 ± 5.3	77.2 ± 2.9	77.7 ± 5.2	77.3 ± 2.9	0.90
Reserve Females	60.0 ± 5.3	81.2 ± 6.6	82.6 ± 3.4	80.3 ± 7.0	82.4 ± 3.1	0.90
Males	47.0 ± 3.1	72.3 ± 3.8	71.1 ± 2.6	72.3 ± 3.8	71.4 ± 2.6	0.80
Females	55.4 ± 4.5	76.4 ± 4.9	78.8 ± 3.2	76.1 ± 5.0	78.7 ± 3.3	0.95
Active Duty	46.4 ± 3.4	71.6 ± 4.1	70.5 ± 2.9	71.6 ± 4.1	70.7 ± 2.8	0.80
Reserves	54.1 ± 3.5	77.9 ± 4.7	77.9 ± 2.5	78.1 ± 4.6	78.0 ± 2.6	0.90
On CCEP	58.2 ± 1.7	79.1 ± 3.4	82.1 ± 1.0	79.0 ± 3.4	81.9 ± 1.1	0.95
Not on CCEP	47.0 ± 3.0	72.3 ± 3.7	71.1 ± 2.6	72.3 ± 3.7	71.4 ± 2.5	0.80
NH Black	31.6 ± 3.6	60.7 ± 7.4	53.2 ± 4.2	60.6 ± 7.4	54.7 ± 4.3	0.80
NH White	53.2 ± 3.0	76.3 ± 4.0	77.3 ± 2.2	76.3 ± 4.0	77.2 ± 2.2	0.90
Other Race	46.4 ± 4.9	75.0 ± 8.5	75.9 ± 3.7	74.9 ± 8.6	75.8 ± 3.7	0.90

*Post-stratified weights for the initial survey include an adjustment for under-reporting.

Appendix A.

Calculation of Composite Response Rates

We applied the following methodology to compute composite response rates for the Gulf War Veterans Health Survey:

1. **Define two over-lapping samples.**
   s_A = 10,301 veterans initially selected for the survey; and,
   s_B = 5,599 mail respondents plus the 1,000 mail nonrespondents selected for follow-up.
   Note that s_B is a proper subset of s_A.

2. **Assign the design weights.**
   d_{kA} = the inverse of the selection probability assigned to the kth sample member.
   d_{kB} = d_{kA}, if the kth sample member responded to the mail survey. For the 1,000 mail nonrespondents selected for the follow-up, d_{kB} equals d_{kA} times the inverse of the follow-up selection probability.
   Note that $\sum_{s_A} d_{kA} = \sum_{s_B} d_{kB}$ = 685,074 veterans on the sampling frame.

3. **Construct variance replicates.**
   We created 294 variance replicates (a.k.a. random groups) that enable us to combine the data obtained from the mail survey with that obtained from the telephone survey and then use the jackknife method to estimate the variances of survey outcomes in a design-consistent fashion. Within each of the eight first-phase strata, we randomly assigned 35 sample members to each replicate with the requirement that each replicate have approximately equal numbers of mail respondents and at least one follow-up respondent.

4. **Obtain post-stratification totals.**
   A total of 20 control totals T_x were created using combinations of the following variables:

   - Component (2) = Active, Reserve
   - Gender (2) = Male, Female
   - Race/Ethnicity (3) = NH Black, NH White, Other
   - CCEP (2) = Present on CCEP, Absent from CCEP

   Note that the post-stratification totals sum to 689,183 veterans and reflect slightly more complete totals that those for the sampling frame.

5. **Adjust for Under-reporting.**
   We used GEM to calculate Hajek-type adjustment factors $a_{kA,PS}$ and $a_{kB,PS}$ for all the selected units that were applied to the design weights to force them to sum to the 20 control totals. The resulting weights are written as

   $$w_{kA,PS} = d_{kA}\ a_{kA,PS}$$
   $$w_{kB,PS} = d_{kB}\ a_{kB,PS}.$$

   Because the response rate is under-reported for s_A, we constrained the adjustment factor ($a_{kA,PS}$) to be greater than one for the mail respondents. Conversely, the adjustment factors for the s_A nonrespondents were naturally constrained to be less than 1.0 to obtain the necessary control totals. The adjustment factors for s_B remained unconstrained. Using the standard GEM notation, we have

   (1) l_{kA} (=1) < c_{kA} < u_{kA} (= maximum) for the s_A respondents;
   (2) l_{kA} (=0) < c_{kA} < u_{kA} (= 1) for the s_A nonrespondents; and
   (3) l_{kB} (=0) < c_{kB} < u_{kB} (=maximum) for all the s_B cases.

   Additionally, we set the centering factors (the desired mean of the distribution of the adjustment factors) to the following:

   (1) c_{kA} = [WRR for s_B / WRR for s_A] for the s_A respondents;
   (2) c_{kA} = [1 – WRR for s_B] / [1 – WRR for s_A] for the s_A nonrespondents; and
   (3) c_{kB} = 1 (standard post-stratification adjustment) for all the s_B cases.

   Note:
   $\sum_{s_A} w_{kA,PS} = \sum_{sB} w_{kB,PS}$ = 689,183 veterans.

6. **Obtain Response Domains.**
   We created a total of 14 sample-specific response indicators for the following reporting domains. These are referred to as the zero controls in the text.
   - Overall (1)
   - Component by Gender (4)
   - Component (2)
   - Gender (2)
   - CCEP (2)
   - Race/Ethnicity (3)

7. Calculate DFC adjustment factors.

We used GEM to calculate DFC adjustment factors $a_{kA,DFC}$ and $a_{kA,DFC}$ that were applied to the adjusted design weights so that the difference between the set of weighted response rates for s_A and s_B was zero (14 zero controls) while maintaining the 20 control totals. The resulting weights are written as:

$$w_{kA,DFC} = d_{kA}\ a_{kA,PS}\ a_{kA,DFC}$$

$$w_{kB,DFC} = d_{kB}\ a_{kB,PS}\ a_{kB,DFC}.$$

All the DFC adjustment factors were constrained to be positive; the centering factors were set to one.

Note: $\sum_{s_A} w_{kA,DFC} = \sum_{s_B} w_{kB,DFC}$.

Using the η grid, we determined that a scaling constant (η_A) of 0.80 and higher minimized the variances for all the 14 response domains.

8. Construct replicate sets of DFC weights.

We constructed 294 replicated sets of DFC weights for use in the jackknife method of variance estimation. Each set of weights was constructed by first excluding one variance replicate per stratum, and then adjusting the weights of the remaining replicates within the same stratum for the subsampling, and (3) then applying the DFC methodology (Step 7) to the remaining respondents

This process was repeated until each of the 294 variance replicates had been excluded from one replicated set of weights. For each replicated set of DFC weights, we used GEM to calculate jackknife adjustments factors $a_{kA,AJDFC}$ and $a_{kB,AJDFC}$ that were applied to the design weights from Step 7 so that the DFC constraints were maintained.

$$w_{kA,AJDFC} = d_{kA}\ a_{kA,PS}\ a_{kA,DFC}\ a_{kA,AJDFC}$$

$$w_{kB,AJDFC} = d_{kB}\ a_{kB,PS}\ a_{kB,DFC}\ a_{kB,AJDFC}.$$

9. Determine optimal CRRs.

We calculated CRRs using the DFC adjustment weights ($w_{kA,AJDFC}$, $w_{kB,AJDFC}$). The half-width confidence intervals were calculated for a range of η_A using the Jackknife weights in SUDAAN. Note that the Jackknife point estimates are not equivalent to the DFC point estimate because the calibration adjustment factor is non-linear.