Statistical Journal of the IAOS 35 (2019) 289–297
DOI 10.3233/SJI-170406
IOS Press

Integrating the results of a nonresponse follow-up survey into the survey from which its items were selected

Phillip S. Kott
RTI International, 6010 Executive Blvd, Suite 902, Rockville, 20852, MD, USA
Tel.: +1 301 468 8281; E-mail: pkott@rti.org

Abstract. A nonresponse follow-up (NRFU) survey was conducted for the National Pilot of the Residential Energy Consumption Survey (RECS), an address-based sample survey of potential primary residences in the US enumerated by web and mail. Virtually all unit (i.e., whole-record) nonrespondents to the National Pilot were sent a short mail questionnaire containing 18 key items from the full survey. Here, we first compare two ways of adjusting variables collected on the NRFU for unit nonresponse. In one, only the weights for respondents to the full National Pilot survey were adjusted to compensate for nonresponse using a calibration weighting procedure that assumes response to be a logistic function of variables known for the entire sample (the NRFU sample was ignored). In the other, only the NRFU-survey respondents' weights were adjusted for nonresponse using an analogous calibration weighting scheme, while weights for the respondents to the full survey were not adjusted. The resulting two national estimates for many of the NRFU variables were then compared. When the two were significantly different, the latter estimate was treated as unbiased and added as a calibration variable when adjusting (a second time) for unit nonresponse to the full sample. When they were not significantly different, both were deemed unbiased, and the mean of the two added as a calibration variable when readjusting for nonresponse to the full sample. The theory behind this practice and its repercussions are discussed.

Keywords: Residential Energy Consumption Survey, calibration weighting, augmented sample, compositing factor

1. Introduction

This paper describes a sensible and statistically defensible method for integrating a short nonresponse follow-up (NRFU) survey of unit nonrespondents into a larger survey, where the items on the NRFU were a subset of the items on the larger survey. The concept of a NRFU is credited to Hansen and Hurwitz [1]. More elaboration can be found in Vandenplas et al. [2].

We will use as an illustrative example the (US) National Pilot of the Residential Energy Consumption Survey (RECS), a voluntary address-based probability sample of 9,650 housing units enumerated by web and mail, and its NRFU survey in which all nonrespondents to the National Pilot were recontacted by mail except hard refusals (sampled housing units that requested not to be recontacted).

The RECS National Pilot was recently investigated by this author in Kott and Liao [3].

Unlike many NRFUs in practice, the NRFU sample for the National Pilot was virtually a complete census of unit nonrespondents, not a probability subsample of nonrespondents as is usually the case (see McMillen et al. [4]). Moreover, the NRFU survey did not include questions about why the housing unit did not respond to the full National Pilot survey (as did Couper et al. [5]). What it included were items of critical interest to the Energy Information Administration (EIA), which funded the survey.

Section 2 describes the survey and sample design of the RECS National Pilot. Section 3 outlines how it was weighted for non-eligibility and nonresponse. Section 4 discusses the NRFU survey and how its results could be integrated into the full National Pilot estimates. Section 5 contains some concluding remarks including a look at how the integration affects standard errors.

2. The RECS National Pilot

The RECS National Pilot was an attempt to convert what historically has been an in-person interview survey into one conducted by web and mail. More information on this project can be found elsewhere (Berry and O'Brien [6]). For our purposes, the RECS National Pilot (hereafter the "National Pilot") used four randomly-assigned protocols and two randomly-assigned incentive levels in data collection from a stratified, two-stage sample drawn using an address-based sampling frame with mail invitation and up to six mailings.

The protocols were, 1, web only, 2, choice of web or mail, 3, choice of web or mail but with an added $10 incentive to respond via web, and, 4, web in the first mailing followed by a choice in subsequent mailings. The two incentive levels both provided the sampled housing unit (HU) $5 initially. One provided an extra $10 upon completion while the other provided an extra $20. There was a shortened mail follow-up survey (NRFU) for all but the hardest nonrespondents.

Two unusual issues faced in the enumerations of the National Pilot have an impact on the analysis to be described here. Not all HUs in the sampling frame were occupied, and some were occupied but not primary residents. Only data from primary residents were deemed in scope for the National-Pilot estimates.

A latent-variable model (Biemer et al. [7]) was used to estimate the probability that a sampled HU was occupied based on its frame characteristics, the disposition of the first three mailings, and whether it responded to the survey. Those estimates have been incorporated into the base weights. Also, incorporated into the base weights is the estimated probability of a non-vacant HU being a primary residence. Every responding primary residence had an estimated probability of 1, and all HU determined not to be primary residences a probability of 0. The rest have been assigned a probability of being a primary residence based on a logistic regression conducted among partially or fully responding HUs to either the National Pilot or the nonresponse follow-up (NRFU) survey for which primary residence status could be determined. A fuller discussion of these weighting steps is beyond the scope of this discussion.

3. Weights for the National Pilot

The base weight (BASE_WT) for an HU in the RECS National Pilot is the product of two components: its primary sampling unit weight and its conditional housing unit weight. An HU's primary sampling unit weight is the inverse of the selection probability of the primary sampling unit (PSU) containing it. A PSU is a county or group of contiguous counties randomly selected from one the 19 RECS geographical domains, the design strata, with Alaska and Hawaii each being its own domain. The 200 PSUs were selected with probabilities proportional to expected size. The sample PSUs were used for both the RECS National Pilot and the traditional multi-stage 2015 RECS (Energy Information Administration [8]) for subsequent analyses not described here.

The conditional housing unit weight is the inverse of the conditional selection probability of selecting a particular HU within a sampled PSU. The base weight of a selected housing unit is the product of its PSU and conditional housing weight. Both National Pilot respondents and nonrespondents have base weights.

Weight adjustment factors are often implemented in survey statistics to reduce the impact of nonresponse and coverage errors and to increase statistical efficiency (i.e., reduce standard errors). The first two weight adjustments to the National Pilot, the nonvacancy adjustment factor and the primary HU adjustment factor, were applied to the entire sample, both National Pilot respondents and nonrespondents, because some nonresponding sampled HUs were vacant, and among non-vacant HUs, some were not primary residences.

First, the probability that a HU was not vacant was estimated using latent class modeling described in Biemer et al. [7]. It is 1 for every responding HU, but can be less than 1 for nonresponding HUs. The nonvacancy adjustment factor for a nonresponding HU is the inverse of this estimated probability.

Many sampled HUs responded only partially to either the full Pilot survey or the NRFU survey. For such an HU, we can determine whether it is a primary residence. The estimated probability that a remaining sampled HU (i.e., one that does not even partially respond) is a primary residence was determined using an unweighted logistic regression model with an urbanicity indicator (described in some detail a footnote to Table 1), the address-based frame indicator of whether the HU was a single-family dwelling unit, and the fraction of owned HUs in the Census block group contain-

Table 1
Calibration variables for tentative nonresponse adjustment

Calibration variable	Some details
Modified RECS Domain	17 levels; Alaska added to the domain with Oregon and Washington, and Hawaii added to California
Urbanicity	2 levels (URBAN_1)
Protocol	4 levels
Incentive	2 levels
Housing unit type	Single or multiple family unit (variable on the frame)
CBG ownership rate	Percentage of owner HUs
CBG low income	Median income below $60,000 per year (yes or no)

URBAN_1 was defined at the Census tract level using USDA rural-urban continuum codes USDA rural-urban continuum codes (http://www.ers.usda.gov/data-products/rural-urban-continuum-codes/). CBG = Census Block Group using 2013 American-Community-Survey (ACS) 5-year averages. Note: We follow SUDAAN terminology and label a survey item that generates dummy variables a "categorical variable" that takes on "levels."

ing the HU as the explanatory variables. The primary housing unit factor is the inverse of this estimated probability. It is 1 for all sampled HUs determined to be primary residences and 0 for those determined not to be primary residences from survey responses.

The *eligibility-adjusted base weight* for an HU (ELIG_WT) is the product of an HU's base weight, non-vacancy factor and primary housing unit factor. It is used to estimate full-sample estimates for a set of characteristics. These estimated totals are the targets used in nonresponse adjustment for the full sample in a manner to be explained shortly.

Each respondent to the National Pilot survey received a *tentative nonresponse adjustment factor* (all other sampled HUs receive a TNR_FC of 0). Based on the characteristics of the HU, this factor is the inverse of an estimate of the probability that the HU responds when sampled. In other words, the implicitly estimated probability of response ("tentative" because there is a subsequent poststratification adjustment) is treated as an additional phase of probability of selection. In fact, the tentative nonresponse-adjusted weight TNR_WT of a HU responding to the RECS National Pilot is its base weight times its tentative nonresponse adjustment factor: TNR_WT = BASE_WT × TNR_FC.

The characteristics used in estimating the tentative nonresponse-adjusted weights are referred to as the tentative nonresponse "calibration variables" because the TNR_FC were chosen using the WTADJUST procedure in SUDAAN 11 (Research Triangle Institute [9]) so that the following calibration equation holds for every characteristic:

$$\sum_{HU \in Sample} \left\{ \text{BASE\_WT} \times \text{TNR\_FC} \times \begin{matrix}\text{Calibration}\\ \text{Variable}\end{matrix} \right\} = \sum_{HU \in Sample} \left\{ \text{ELIG\_WT} \times \begin{matrix}\text{Calibration}\\ \text{Variable}\end{matrix} \right\},$$

where both summations are over the full-survey sample. Recall that TNR_FC is zero for nonrespondents while ELIG_WT need not be. The summed values on the right-hand side of this equation are the calibration targets used in determining the TNR_FC. The mathematical details of calibration weighting can be found in the appendix. SUDAAN allows calibration targets to be random, as they are here, and their randomness is incorporated into the variance estimation.

Table 1 features the list of nonresponse calibration variables used in the above equation. Selecting calibration variables is analogous to choosing the variables for a logistic-regression model with response/nonresponse as the dependent variable. The list in Table 1 was culled from a larger list determined by expert opinion. The final set of variables was selected mostly through backward selection using unweighted logistic regression, after which the need to include interaction terms was investigated. The exact form of some of the calibration variable was based on the form producing the logistic regression with highest model F value. The final logistic response model was fit via calibration weighting as explained in the appendix. No TNR_FC was larger than 4.75, and the average value was roughly 2.47.

A shortened version of the 2015 RECS National Pilot survey, the nonresponse follow-up (NRFU) survey containing 18 items of special interest to EIA, was sent to all nonrespondents to the full National pilot survey except for hard refusals. The unweighted response rate for the full survey was 37.8%, which increased to 51.8% for NRFU-survey variables. The 18 survey items generated over 20 NRFU-survey variables. For example, the item, "What fuel does your main water heater use?" (FUELH20) generated three dummy variables: FUELH2O = Natural gas, yes (1) or no (0); FUELH2O = Electricity, yes (1) or no (0); and FUELH2O = Other, yes (1) or no (0).

In SUDAAN terminology, a survey item that generates dummies is a "categorical variable" with a "level" for each dummy. A variable that does not generate dummies, even when it takes on only a finite number of values, is labeled "continuous".

In the next section, we describe an experimental weighting regime that was not used in the National Pilot in part because the wording of certain key items was not the same in the full survey and the NRFU. The regime described here ignores any disparity in wording (which was minor in all cases) and integrates the NRFU responses into the estimates to demonstrate a method for doing so.

4. Full-survey weighting that incorporates NRFU-variable responses

The combination of the samples for the full and NRFU surveys is referred to as the "augmented sample." A second nonresponse-adjusted estimate was computed for an item on the NRFU (hereafter a "NRFU item") in addition to the one computed using the tentative nonresponse-adjusted weights described last section. In both estimates, imputed values were used for NRFU item nonresponse. See Energy Information Administration [8, pp. 12–13] for a description of RECS imputation. The method used for the second estimate employed the augmented sample but used NRFU survey respondents alone to compensate for nonresponse. It assumes nonrespondents are more like NRFU-survey respondents than full-survey sample respondents, even after adjusting for differences in their known characteristics, because NRFU respondents also failed to respond to the original National Pilot survey.

In the second estimation method, the augmented-sample nonresponse adjustment factor (ANR_FC) was set to 1 for all respondents to the full survey (and to 0 for all NRFU-survey nonrespondents including refusals from the full survey who were not sent a NRFU survey). The augmented-sample nonresponse adjustment factor for a respondent to the NRFU survey is then computed analogously to tentative nonresponse adjustments but with different targets.

The calibration equations (one for every calibration variable) used to determine the ANR_FC were

$$\sum_{HU\,\in\,\substack{augmented\\sample}} \left\{\text{BASE\_WT} \times \text{ANR\_FC} \times \begin{matrix}\text{Calibration}\\ \text{Variable}\end{matrix}\right\} = \sum_{HU\,\in\,\substack{augmented\\sample}} \left\{\text{ELIG\_WT} \times \begin{matrix}\text{Calibration}\\ \text{Variable}\end{matrix}\right\},$$

Table 2
Calibration variables for the augmented-sample nonresponse adjustment

Calibration variable	Some detail
Census Division plus 1	10 levels (Arizona, New Mexico, and Nevada form the 10th Division)
Urbanicity	2 levels (URBAN_1)
Protocol (original)	4 levels
NRFU added Incentive	2 levels
Housing Unit Type	Single or multiple family unit

where the summations are again over the augmented sample, but *the ANR_FC values are freely chosen (i.e., not set at 0 or 1) only for the NRFU respondents* (ANR_FC = 1 for full-survey respondents). Then the augmented sample weights are defined by

$$\text{ANR\_WT} = \text{BASE\_WT} \times \text{ANR\_FC}.$$

Again, see the appendix for more mathematical details.

Table 2 features the list of nonresponse calibration variables chosen after model selection (analogous to that for the variables in Table 1). Again, a logistic response model was fit via calibration weighting. This resulted in some ANR_FC values larger than 9. A truncated logistic response model was then fit instead, one that assumed no probability of response was less than 1/8. With it, no ANR_FC was greater than 7.1.

Although the augmented sample is larger than the full sample, the variability of the augmented-sample weights is such that the estimated variances of NRFU items from using ANR_WT is often higher than using TNR_WT with the full sample. Estimated variances were computed using the Taylor-series linearization routine in SUDAAN treating weights as inverse selection probabilities (including unit response).

Using a method described later in this section describing results in Table 4, our model fitting revealed that there were NRFU items with significantly different estimates when ANR_WT is used for weighting rather than TNR_WT (INTERNET, AIRCOND, and certain levels of FUELHEAT, FUELH2O, and EQUIPM). For categorical variables, only levels with augmented-sample estimates of 10% or higher have been tested for significant differences were then judged not significant. In all, 28 NRFU-related variables have been tested using a Bonferroni-Holm procedure set at the initial 0.1 level (i.e., the difference with the largest t-value in absolute value was deemed significant if it corresponded to a two-sided probability of less than 0.1/28, the second largest as well if it corresponded to a two-sided probability of less than 0.1/27, and so forth until a t-value was not deemed significant.

Table 3
Calibration variables for the final nonresponse adjustment

Calibration variable	Some details
Modified RECS domain	17 levels; Alaska added to the domain with Oregon and Washington and Hawaii is added to California
Urbanicity	2 levels (URBAN_1)
Protocol	4 levels
Incentive	2 levels
Housing unit type	Single or multiple family unit
CBG ownership rate	% of owner HUs
CBG low income	Median income below $60,000
TYPEHUQ (Housing Unit Type)	5 levels (mobile, detached single unit, attached single unit, apartment in building with less than five units, other)
HU OWNED	2 levels
YEARMADERANGE	Continuous* (by decade)
Number of HouSeHoLD MEMbers	Continuous
BEDROOMS	Continuous
NUMber of Smart PHONES	Continuous
DESKTOP	Continuous
Clothes WASHER	2 levels
Number of TVCOLOR	Continuous
DISHWASH	2 levels
NUMber of FREEZers	Continuous
NUMber of ReFRIGerators	Continuous
DRYER	2 levels
NUMber of LAPTOPs	Continuous
COOLTYPE	4 levels: Central, Window, Both; AIRCOND = 0 is treated as a level (Computed with ANR_WT):
FUELHEAT	3 levels: Electric, gas, other (Computed with ANR_WT)
FUELH2O (water heating)	3 levels (Computed with ANR_WT)
Heating EQUIPMent	4 levels (Computed with ANR_WT)
INTERNET	2 levels (Computed with ANR_WT)

*Using SUDAAN terminology (see the text).

At this point, we have two potential estimates for each variable derived from both the full and NRFU surveys (after imputing for item nonresponse): (1) an estimate based on the respondents to the full survey using the weights TNR_WT; and (2) an estimate based on respondents to either the full survey or the NRFU survey using the weights ANR_WT. The latter estimate is assumed to be unbiased. The former may or may not be biased.

For a NRFU variable deemed *not* to be biased when estimated with either set of weights (because its two estimates are not significantly different), one could, in principle, choose an NRFU compositing factor CNR_FC between 0 and 1 so that when the weights

$$\text{CNR\_WT} = \text{CNR\_FC} \times \text{ANR\_WT} + (1 - \text{CNR\_FC}) \times \text{TNR\_WT}$$

(setting TNR_WT to 0 for NRFU respondents) are applied, the variance of the resulting estimate would be minimized. In practice, the best we can do is minimize the estimated variance, which may not be the same thing.

For a NRFU variable deemed to have a bias when estimated using the TNR_WT weights, one sets CNR_FC = 1. For consistency when one level of a variable (like FUELHEAT) is estimated using CNR_FC = 1 (so that CNR_WT = ANR_WT), then all the levels are so estimated.

For those NRFU variables whose estimates are deemed not to be biased when the TNR_WT are used, the variance-minimizing CNR_FC varies by variable. Setting CNR_FC at 1/2 turned out to be a reasonable choice for all variables where using the TNR_WT was deemed *not* to produce biased estimates.

The interim weights TNR_WT, ANR_WT, and CNR_WT are all means to an end – improved control totals that take advantage of NRFU data where it makes sense to do so. The control totals are used to generate the final adjustment factors (unless the TNR_WT are selected as the final nonresponse weights).

Returning to the full National Pilot sample, we can now recompute the nonresponse-adjusted weights to try to remove the biases observed when using TNR_WT. Through this step, we can also decrease the variance of full-sample estimates of NRFU variables that are *not* biased when TNR_WT is used. This is done by adding the totals for the variables in Ta-

Table 4
NRFU variable estimates computed with different weights

Variable (estimated number per HU)	p-value of difference between using the two sets of weights	Tentative nonresponse-adjusted estimate and its standard error		Augmented-sample nonresponse-adjusted estimate and its standard error		Composite nonresponse-adjusted estimate when CNR_FC = 1/2 and its standard error	
Detached HU	0.247	0.633	0.0144	0.620	0.0155	0.627	0.0140
Attached HU	0.137	0.098	0.0076	0.108	0.0087	0.103	0.0075
In building with 5 or more units	0.382	0.154	0.0128	0.148	0.0128	0.151	0.0122
Owned HU	0.377	0.680	0.0119	0.672	0.0121	0.676	0.0110
YEARMADERANGE	0.093	4.251	0.0730	4.177	0.0680	4.214	0.0667
NHSLDMEM	0.192	2.537	0.0288	2.501	0.0292	2.519	0.0251
BEDROOMS	0.999	2.841	0.0357	2.841	0.0317	2.841	0.0313
DESKTOP	0.084	0.540	0.0135	0.563	0.0160	0.552	0.0130
NUMTABLET	0.312	0.978	0.0239	1.001	0.0260	0.990	0.0222
NUMSMPHONE	0.986	1.639	0.0295	1.640	0.0295	1.639	0.0263
CWASHER	0.953	0.850	0.0101	0.849	0.0104	0.850	0.0095
TVCOLOR	0.118	2.329	0.0319	2.370	0.0313	2.350	0.0287
DISHWASH	0.166	0.731	0.0132	0.720	0.0131	0.725	0.0124
NUMFREEZ	0.073	0.371	0.0139	0.394	0.0144	0.383	0.0128
NUMFRIG	0.966	1.394	0.0148	1.395	0.0133	1.394	0.0127
DRYER	0.987	0.834	0.0108	0.834	0.0107	0.834	0.0099
NUMLAPTOP	0.014	1.068	0.0273	1.015	0.0262	1.041	0.0243
INTERNET	0.000	0.872	0.0071	0.836	0.0087	0.854	0.0070
AIRCOND	0.000	0.867	0.0106	0.896	0.0085	0.881	0.0091
Central Air Cond. Only	0.333	0.626	0.0134	0.634	0.0127	0.630	0.0122
Window Air Cond. Only	0.055	0.192	0.0091	0.205	0.0095	0.198	0.0084
FUELHEAT = Natural gas	0.068	0.476	0.0150	0.457	0.0150	0.467	0.0141
FUELHEAT = Electricity	0.000	0.366	0.0136	0.410	0.0136	0.388	0.0125
FUELH2O = Natural gas	0.004	0.487	0.0158	0.451	0.0164	0.469	0.0151
FUELH2O = Electricity	0.046	0.452	0.0141	0.475	0.0150	0.464	0.0135
Central furnace	0.477	0.608	0.0111	0.602	0.0124	0.605	0.0108
Built-in electric unit in walls, floors, etc.	0.000	0.073	0.0053	0.102	0.0070	0.088	0.0052
Heat pump	0.457	0.116	0.0076	0.121	0.0087	0.118	0.0073

Note: Red denotes the estimates used as target variables for the final nonresponse-adjusted weights. Standard errors and p-values were computed ignoring any contribution to standard-error reduction from the nonresponse adjustments.

ble 4-2 computed from the augmented sample to the calibration equations from Section 3.2 used to implicitly determine the final nonresponse-adjustment factor, FNC_FC. The added calibration equations have the form:

$$\sum_{HU\in Sample}\left\{\text{BASE\_WT} \times \text{FNR\_FC} \times \begin{matrix}\text{Calibration}\\ \text{Variable}\end{matrix}\right\} = \sum_{HU\in Sample}\left\{\text{CNR\_WT} \times \begin{matrix}\text{Calibration}\\ \text{Variable}\end{matrix}\right\},$$

where the summations are over the full-survey sample (including nonrespondents).

Despite the large number of calibration variables in Table 3, all targets were met, even when we set the floor for the probability of response at 1/6. In fact, no FNR_FC was larger than 5.6 with that setting.

The final nonresponse-adjusted weights were then

$$\text{FNR\_WT} = \text{BASE\_WT} \times \text{FNR\_FC}.$$

For nonrespondents, FNR_FC is 0. These adjustments, FNR_FC, adjust base weights for eligibility and nonresponse in a single step, now that we have improved control totals.

Applying the final nonresponse-adjusted weights to full-survey respondents would ideally ensure the equality of the estimated NRFU variable in Table 3 with estimates computed from the augmented sample using the CNR_WT at the national level, but not necessarily within subpopulations (like a division or a housing type). Even at the national level, the ideal equality may be lost when imputation is finalized using the final nonresponse-adjusted weights.

Table 4 contains a display of the alternative estimates for NRFU variables. The estimated means are computed with one of the sets of weights described in the text. For a proportion, like DWASHER, the estimated number per HU is the estimated proportion of HUs with that item. Then for a multilevel variable such

Table 5
NRFU-variable coefficients of variation (CVs) when computed with different nonresponse weights

Variable (estimated number per HU)	CVs computed with TNR_WT and its BRR replicates	CVs computed with FNR_WT and its BRR replicates	Log (Col 2/Col 1)
Detached HU	0.019435	0.018124	−0.06984
Attached HU	0.086315	0.077540	−0.10721
Apartment in Bld with 5 or more units	0.071195	0.072748	0.02158
Owned HU	0.013178	0.012075	−0.08745
YEARMADERANGE	0.014648	0.014417	−0.01585
NHSLDMEM	0.010740	0.009350	−0.13858
BEDROOMS	0.010149	0.008602	−0.16543
DESKTOP	0.023399	0.022455	−0.04116
NUMTABLET	0.020917	0.019220	−0.08462
NUMSMPHONE	0.013268	0.012647	−0.04791
CWASHER	0.011502	0.009104	−0.23381
TVCOLOR	0.013612	0.012333	−0.09869
DISHWASH	0.016696	0.016248	−0.02722
NUMFREEZ	0.032998	0.028757	−0.13756
NUMFRIG	0.009766	0.008381	−0.15287
DRYER	0.011815	0.009938	−0.17296
NUMLAPTOP	0.018629	0.018953	0.01728
INTERNET	0.006454	0.009758	0.41341
AIRCOND	0.012095	0.010000	−0.19022
Central air conditioning only	0.022150	0.021296	−0.03932
Window air conditioning only	0.043813	0.043575	−0.00546
FUELHEAT = Natural gas	0.026686	0.027746	0.03898
FUELHEAT = Electricity	0.029910	0.029587	−0.01086
FUELH2O = Natural gas	0.027816	0.028139	0.01155
FUELH2O = Electricity	0.028147	0.025511	−0.09831
Central furnace	0.016253	0.018405	0.12436
Built-in electric unit in walls, floors, etc.	0.072399	0.067865	−0.06467
Heat pump	0.056098	0.062300	0.10487

as TYPEHUQ (housing type) the value is the estimated proportion at a particular level, for example, TYPEHUQ = 2 (detached single-family HU).

The standard errors and p-values in Table 4 have been computed using PROC DESCRIPT in SUDAAN, ignoring any contribution to standard-error reduction from the tentative or augmented-sample nonresponse adjustment for relative simplicity. Both samples were treated like stratified multistage samples assuming with replacement sampling of PSUs. The self-selection into the NRFU subsample was treated as independent across housing units. The assumption of with-replacement selection of PSUs is not strictly true, but commonly made with complex survey data so that weighted PSU-level aggregates can treated as independent with a common mean within each (variance) stratum (when a variance stratum combines design strata, variance estimates can be conservative). This facilitates linearized variance estimation (Research Triangle Institute [9, pp. 60–64].

To generate p-values for differences between two estimates of the same proportion computed with the same observations but with different weights, the two estimates were treated as means of different domains. Each sampled HU was repeated in the data set, one version had the TNR_WT weights and was assigned to Domain A, while the other had the ANR_WT weight and was assigned to domain B. Both were contained in the same PSU. A CONTRAST statement was employed to test the difference between the two "domain" means. This methodology treated the PSU-level aggregates of the estimated difference as independent, while capturing the correlation within a PSU of the same housing unit being in both domains.

5. Some concluding remarks

The main goal of this paper was to demonstrate a reasonable method for integrating the results of a nonresponse follow-up (NRFU) survey with a limited number of items into a sample survey, called the "full survey." For some, items, the NRFU results revealed biases in estimates produced by the full-survey without additional nonresponse adjustment. For others, the NRFU-collected information served as additional sampled data thereby potentially reducing standard errors.

The method involves these steps:

Step 1. Determine tentative nonresponse-adjusted (TNR) weights for respondents to the main survey using calibration weighting as if there were no NRFU.

Step 2. Determine augmented-sample nonresponse-adjusted (ANR) weights for survey items on the NRFU using calibration weighting adjusting only for NRFU nonresponse (i.e. only elements responding to the NRFU but not the main survey are weight-adjusted for nonresponse).

Step 3. Add full-population estimates for the NRFU survey items to the calibration variables used to create final weights for all items to the main survey. These NRFU item estimates are either the estimates computed using the ANR weights or a composite of the estimated using the ANR and TNR weights, the former being used with the two estimates (one computed with TNR weights and the other with ANR weights) are significantly different, the latter otherwise.

Table 5 displays coefficients of variation for NRFU-variable estimates computed using only the original full sample and its nonresponse adjustment weights (TNR_WT) and then reweighting that sample using the NRFU-survey results to form additional calibration targets (FNR_WT). Standard errors at this stage (i.e., after the weights have been determined) were computed using Fay's BRR technique (Judkins [10]).

The last column is a symmetric measure of the percent difference between the CVs. Observe that $\log(\text{Col } 1/\text{Col } 2) = -\log(\text{Col } 2/\text{Col } 1)$.

Not surprisingly for the first 17 variables, the ones for which there was deemed to be no bias in the estimates from the full survey, the CVs tend to be lower when computed using FNR_WT (9% lower, on average). We would expect the similar results from variables correlated with one or more of these 17. For the remaining NRFU variables (starting with INTERNET), the CVs are sometimes lower and sometimes higher using FNR_WT (averaging 2% higher), but are likely less biased.

References

[1] Hansen M, and Hurwitz W. The problem of nonresponse in sample surveys. J Amer Stat Assoc. 1946; 41, 517-529.

[2] Vandenplas C, Dominique J, Staehli M, and Alexandre P. Identifying pertinent variables for nonresponse follow-up surveys: Lessons learned from four cases in Switzerland. Surv Res Meth. 2015; 9, 141-158.

[3] Kott P, and Liao D. Calibration weighting for nonresponse with proxy frame variables (so that unit nonresponse can be not missing at random). J Off Stat. 2018; 34, 107-120.

[4] McMillen M, Harris-Kojetin B, Miller R, and Ware-Martin A. Nonresponse in Measuring and Reporting Sources of Error in Surveys, Statistical Policy Working Paper 31. In: Subcommittee on Measuring and Reporting the Quality of Survey Data, Federal Committee on Statistical Methodology, Daniel Kasprzyk, Chair 2001. pp. 4-11.

[5] Couper M, Peytchev A, Strecher V, Rothert K, and Anderson J. Following up nonrespondents to an online weight management intervention: Randomized trial comparing mail versus telephone. Journal of Medical Internet Research. 2007; 9(2), e16.

[6] Berry C, and O'Brien E. Managing the fast-track transformation of a 35-year old federal survey, Presented at the 2016 FedCASIC Workshop, Washington DC., https://www.census.gov/fedcasic/fc2016/ppt/2_2_Speed.pdf.

[7] Biemer P, Kott P, and Murphy J. Estimating mail or web survey eligibility for undeliverable addresses: a latent class analysis approach, ProcASA Surv Res Meth Sec. 2016, 1166-72. https://ww2.amstat.org/MembersOnly/proceedings/2016/data/assets/pdf/389587.pdf.

[8] Energy Information Administration, Residential Energy Consumption Survey (RECS) 2015 Technical Documentation Summary, 2017, https://www.eia.gov/consumption/residential/reports/2015/methodology/pdf/methodology_report.pdf.

[9] Research Triangle Institute, SUDAAN Language Manual, Volumes 1 and 2, Release 11 2012. Research Triangle Park, NC: Research Triangle Institute.

[10] Judkins D. Fay's method for variance estimation. J Off Stat. 1990; 6, 223-239.

[11] Kott P, and Liao D. Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. Surv Res Meth. 2012; 6, 105-11.

[12] Folsom R, and Singh A. The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification, Proc ASA Surv Res Meth Sec. 2000, 598-603.

[13] Kim J, and Riddles M. Some theory for propensity scoring adjustment estimator. Surv Meth. 2012; 38, 157-65.

Appendix

Calibration weighting in the RECS National Pilot has the form (using generic notation)

$$w_k = d_k a_k,$$

where the (vector) calibration equation $\sum\limits_{k \in S} w_k \mathbf{z}_k = \widehat{\mathrm{T}}_{\mathbf{z}}$, is satisfied, d_k is the eligibility-adjusted base weight of HU k before the calibration-weight adjustment, w_k is its weight after the calibration-weight adjustment, a_k is its weight-adjustment factor described below, S is the HU sample, $\boldsymbol{z}_k$ is a vector of calibration variables including a constant or the equivalent, and $\widehat{\mathrm{T}}_{\mathbf{z}}$ is an estimated total for the vector of calibration variables. For tentative nonresponse adjustments it is $\widehat{\mathrm{T}}_{\mathbf{z}} = \sum_S d_k \boldsymbol{z}_k$.

The adjustment factor for a_k is restricted to 0 for nonrespondents in nonresponse adjustment (and restricted to 1 for full-survey respondents in augmented-sample nonresponse adjustment). Otherwise it has this form of the generalized exponential model (See Kott and Liao [11]; Folsom and Singh [12] coined the term "generalized exponential model"):

$$a_k = \frac{L + \exp\left(\mathbf{g}^T \mathbf{z}_k\right)}{1 + \frac{\exp(\mathbf{g}^T \mathbf{z}_k)}{U}},$$

where $0 \leqslant L \leqslant U \leqslant \infty$, and the vector $\mathbf{g}$ is chosen (using Newton's method) so that the calibration equation holds, if possible. By restricting the L and U as the above equation does, it opens the possibility that no $\mathbf{g}$ exists that satisfies it.

Observe that restricting L to be no smaller than 1 (when possible) ensures that the weight-adjustment factor must be at least 1. When $L = 1$ and $U = \infty$, this form of calibration weighting for nonresponse adjustment treats response as a logistic function of the vector $\mathbf{z}_k$ (Kim and Riddles [13] show that calibration weighting is superior to employing a maximum-likelihood-based technique when adjusting for survey nonresponse). For other settings of L and U, nonresponse is equivalent to a truncated logistic function of $\mathbf{z}_k$, where the probability of response is restricted to the range (1/U, 1/L). We can employ a set of restrictions to ensure that no weight is too high or too low. For example, when we set $U = 6$ (i.e., 1/U = 1/6), as we did in the final nonresponse adjustment no adjusted weight is more than 65 times its initial weight. Some sets make satisfying the calibration equation impossible (e.g., we could not have set $U = 5$ for the final nonresponse adjustment). For the tentative nonresponse adjustment, no bound was set on U.

For all the nonresponse adjustments, L was set at 1, which means that the estimated probability or response was bounded above by 1.

Satisfying the calibration equation may not be possible even when there are no restrictions on L and U because of the number of components in the vector $\mathbf{z}_k$ (but that never happened with the RECS National Pilot data).