Weighting for Nonresponse in the Overseas Citizen Population Survey: The Effects of Sample Size and Adjustment Method on Accuracy of Estimates

Jonathan Mendelson* Pengyu Huang †

Abstract

Survey nonresponse can be accounted for during weighting by using methods such as response propensity adjustments and calibration adjustments. These methods typically rely on the availability of adjustment variables related to an individual's response propensity and to key survey variables. However, complicated patterns of nonresponse can prevent the ability to control for all levels of relevant variables simultaneously, forcing compromises, such as using a simpler response propensity model or coarsening the level of data used in forming adjustment cells. These compromises may lessen the effectiveness of adjustments at reducing nonresponse bias. Further, sample size limitations could lead to increased design effects, given that small adjustment cells could lead to increased weight variation.

Using resampling methods, we explore the effects of sample size and nonresponse adjustment method on estimates in a post-election survey of overseas U.S. citizens following the 2014 General Election. In this survey, response rates were heavily impacted by voter participation history, state of registration, and country of residence. We found that holding weighting scheme constant, sample size reductions led to increased design effects that likely resulted from smaller adjustment cells. We also found that weighting schemes with lower complexity yielded a larger squared bias component of the mean squared error, particularly at larger sample sizes, although the simulated bias was fairly small and, therefore, not of practical importance.

Key Words: nonresponse, propensity, voting, weighting, adjustment, resampling, calibration

1. Introduction

Nearly every major sample survey suffers from the problem of unit nonresponse, wherein some sampled units do not reply to the survey. Nonresponse is typically accounted for in the estimation strategy, using methods such as response propensity adjustments and calibration, the latter of which includes poststratification, raking, and linear calibration as special cases. The effectiveness of these methods at reducing nonresponse bias can depend on the adequacy of the underlying model for nonresponse. Effectively reducing bias may require the ability to accurately estimate individual-level response propensities for the full population and/or forming response adjustment cells that are homogenous with respect to the relevant survey measure. Accounting for nonresponse may also involve the use of nonlinear estimators that are asymptotically unbiased, yet which may have biases for small samples.

In practice, it can be challenging to properly account for nonresponse. The true response propensity is not known. It may not be possible to control for all levels of relevant adjustment variables simultaneously. The theoretical unbiasedness of estimators often depends on large sample properties, but in practice, there are limited sample sizes, which may force tradeoffs. For example, to take advantage of large sample properties, it might be necessary to coarsen the adjustment categories, which could reduce the accuracy of estimated response propensities and/or violate the assumption of homogeneity within adjustment cells.

*Corresponding author: jmendelson@forsmarshgroup.com. Fors Marsh Group, 1010 N. Glebe Rd., Suite 510, Arlington, VA 22201; and Joint Program in Survey Methodology, University of Maryland.

†Fors Marsh Group, 1010 N. Glebe Rd., Suite 510, Arlington, VA 22201.

In this paper, we use resampling methods to assess the impact of nonresponse adjustment scheme and sample size on the accuracy of estimates in a study with complicated nonresponse patterns. We assess the joint effects of nonresponse adjustment method (post-stratification vs. raking), complexity of adjustment method (low, medium, and high), and sample size.

2. Background

2.1 Weighting Methods for Nonresponse

Classical survey sampling theory is built around the idea that if every member of the population has a positive and known probability of selection, it is possible to create unbiased estimators of population characteristics (i.e., the expected value of the estimator is equal to the true population value); further, with positive joint inclusion probabilities, it is also possible to unbiasedly estimate confidence intervals for many commonly estimated parameters, such as totals, proportions, and means. However, nearly every major sample survey suffers from the problem of unit nonresponse, wherein some sampled units do not reply to the survey. Thus, it is necessary to account for nonresponse in the estimation strategy.

2.1.1 Horvitz–Thompson Estimator

With complete survey response, the population total $Y = \sum_{i=1}^{N} Y_i$ can be estimated via the Horvitz–Thompson (π) estimator, $\hat{Y}_\pi = \sum_{i=1}^{N} \frac{\delta_i Y_i}{\pi_i} = \sum_{i \in S} \frac{Y_i}{\pi_i}$, where δ_i is an indicator variable that is equal to 1 if a given unit i is in the sample S and 0 otherwise, Y_i is the population value of a variable of interest for unit i, π_i is unit i's probability of selection, and N is the population size. Similarly, the population mean $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ can be estimated via $\hat{\bar{Y}}_\pi = \frac{\hat{Y}_\pi}{N}$. The π estimator can easily be shown to be unbiased with respect to the sample design, given that $\mathrm{E}(\delta_i) = \pi_i$; similarly, $\hat{N}_\pi = \sum_{i \in S} \frac{1}{\pi_i}$ is an unbiased estimator of the population size.

2.1.2 Response Propensity Adjustments

The two main paradigms for viewing unit nonresponse in surveys are the deterministic and stochastic frameworks. Under the deterministic framework, nonresponse is viewed as a fixed characteristic of individuals; however, survey statisticians generally find this paradigm to be of limited utility. Instead, under the stochastic framework, each population member $i \in U$ is assumed to have a response propensity $0 < \phi_i \leq 1$, which indicates the individual's probability of replying to the survey. Under nonresponse, $\hat{N}_\pi$ can no longer be used as an unbiased estimator for the population size, because of the loss of part of the sample. That is, by letting R denote the responding subset of the original sample and letting $\bar{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i$ denote the average response propensity over the population, we have that $\mathrm{E}\left(\sum_{i \in R} \frac{1}{\pi_i}\right) = N\bar{\phi} \neq N$. A naive method to account for nonresponse is, thus, to apply a ratio adjustment for the estimated population size under nonresponse, yielding $\hat{\bar{Y}}_0 = \frac{\sum_{i \in R} Y_i/\pi_i}{\sum_{i \in R} 1/\pi_i}$. However, Kalton and Maligalig (1991) showed that this estimator has an approximate bias of $\mathrm{Bias}\left(\hat{\bar{Y}}_0\right) = \mathrm{E}\left(\hat{\bar{Y}}_0 - \bar{Y}\right) \approx \frac{1}{N\bar{\phi}} \sum_{i \in U} \left(Y_i - \bar{Y}\right)\left(\phi_i - \bar{\phi}\right)$.

Thus, the stochastic framework for nonresponse motivates the use of propensity score adjustments, which involves estimating sample members' response propensities, $\{\hat{\phi}_i\}$, and applying an adjustment equal to their multiplicative inverse, such that our estimated population total is now $\hat{Y}_{NR} = \sum_{i \in R} \frac{d_i}{\hat{\phi}_i} Y_i$, where $d_i = \frac{1}{\pi_i}$ is the design weight for sample

member $i \in S$ (Little, 1986). Here, nonresponse is treated as an additional phase of sampling, with the response propensities being estimated in some fashion. If the estimated response propensities are accurate, such an adjustment can be effective at reducing nonresponse bias. For computational convenience, we define the nonresponse-adjusted weight as $w_i^{NR} = \frac{d_i}{\hat{\phi}_i}$ for $i \in R$, so that our weighted total can be computed as $\hat{Y}_{NR} = \sum_{i \in R} w_i^{NR} Y_i$.

The response propensities $\{\hat{\phi}_i\}$ are often estimated via logistic regression and the inverse may be applied directly for adjustment purposes, or the estimated response propensities may be used in forming classes for cell-based adjustments (e.g., Brick, 2008; Valliant, Dever, & Kreuter, 2013). For some surveys, multiple stages of response propensity adjustments may be applied; for example, it is common to account first for unknown survey eligibility and then to account for survey nonresponse among those known to be eligible. When an unequal probability sampling design is used, there is some disagreement regarding whether the model used to estimate response propensities should be weighted by the design weights or unweighted (e.g., Flores Cervantes & Brick, 2016; Little & Vartivarian, 2003). Although an unweighted model can lead to more efficient weights (i.e., lower weight variation), the weighted model can provide design-based protection against model misspecification, allowing for estimated model parameters that are unbiased with respect to the sampling design for estimating the population-level model parameters.

2.1.3 Calibration

Calibration estimators are of the form $\hat{\mathbf{X}}_{cal} = \sum_S w_i \mathbf{X}_i = \mathbf{X}$, where $\mathbf{X} = \sum_U \mathbf{X}_i$ is a vector of population totals over the universe U and w_i is the weight for sample unit $i \in S$ (Deville & Särndal 1992; Kott 2009). Special cases of calibration are poststratification (Holt & Smith, 1979), raking (Deming & Stephan, 1940), and the generalized regression (GREG) estimator under a linear model. Although some definitions of calibration estimators assume complete response, the estimators are often adapted for situations with incomplete response.

Poststratification involves dividing the sample into G groups, termed *poststrata*, and using these poststrata as classes when applying a weighting class adjustment, such that the resulting weights are calibrated with respect to the poststrata. For example, the nonresponse-adjusted weights could be poststratified by computing the poststratified weight for individual i within group g as $w_{gi}^{PS} = w_{gi}^{NR} \frac{T_g}{\hat{T}_g}$, where w_{gi}^{NR} represents that individual's nonresponse-adjusted weight, T_g denotes the total number of individuals in group g in the population, $\hat{T}_g = \sum_{i=1}^{n_g} w_{gi}^{NR}$ is the sample-based estimate for the number of individuals in this group (using the previous, nonresponse-adjusted weights), and n_g indicates the number of individuals with weights in group g.

Raking is an iterative procedure that calibrates the weights on several sets of raking dimensions by virtue of repeatedly applying a series of poststratification adjustments until the weights converge. For example, a set of raked weights could be computed as $w_{rdi}^{R} = w_{0i} \prod_{r=1}^{R} \prod_{d=1}^{D} adj_{rdi}$, where w_{0i} is the weight before raking for individual i and adj_{rdi} is the weighting adjustment received in round r and raking dimension d for individual i obtained by poststratifying the previous set of interim weights. For a given round, the weights are sequentially poststratified to the different sets of raking dimensions (i.e., the weights are poststratified to the first set of raking dimensions, then poststratified to the second set, and so forth, through the Dth set); rounds of adjustments are repeatedly applied, until the weights converge in the Rth round. Note that raking and linear calibration often yield similar results (Rizzo et al., 1996; Brick & Jones, 2008), and thus, linear calibration is not explored in this paper.

2.1.4 Discussion

In practice, the particular method of adjustment (e.g., response propensity adjustments, poststratification, raking) can sometimes matter less than whether the form takes advantage of key auxiliary variables and appropriately reflects the patterns of nonresponse (Brick, 2008; Brick & Jones, 2008; Flores Cervantes & Brick, 2016). However, it is worth noting that many commonly used adjustment methods, including response propensity adjustments and calibration adjustments, commonly result in non-linear estimators, which then might rely on the use of theory involving asymptotic results for samples of increasing size. In such a case, an estimator that is biased for small samples might have a bias that approaches zero as the sample size approaches infinity. Thus, the estimator can be used under the justification that it is approximately unbiased for large samples. For a basic example, the ratio estimator $\hat{Y}_r = \frac{\hat{Y}_\pi}{\hat{X}_\pi} X$ is well known as being biased for small samples in most scenarios, although it is approximately unbiased for large samples. Weighting adjustment methods often include the application of a series of ratio adjustments; for example, poststratification involves applying a ratio adjustment for each poststratum, and raking involves a series of poststratification adjustments.

Although many estimators incorporating weighting adjustments are approximately unbiased for large samples, the small sample results might be unclear. Further, when encountering complicated nonresponse patterns, sample size constraints might force the adjustment model to be simplified in a manner that could lead to biased parameter estimates. Sample size constraints might also force adjustment categories to be coarsened in a way that better meets the assumption of a large sample size, but this coarsening of adjustment categories could reduce the effectiveness of adjustments by virtue of reducing the homogeneity of adjustment cells.

2.2 Overseas Citizen Population Survey

The 2015 Overseas Citizen Population Survey (OCPS) was the first-ever survey of U.S. civilian voters overseas, conducted by the Federal Voting Assistance Program (FVAP) to help FVAP better understand this population, given FVAP's goal of assisting U.S. citizens overseas in exercising their right to vote. This section summarizes key aspects of the 2015 OCPS methodology that pertain to the analysis at hand. Further details are available in the survey's methodological report (FVAP, 2016).

2.2.1 Sampling Frame

The survey's target population was U.S. citizens who were registered to vote and residing overseas at the time of the 2014 General Election, had requested an absentee ballot, and were not considered Uniformed Service voters.[1] Given that U.S. election law varies by state, U.S. voter data is maintained by states, counties, and/or towns. Although there is not a federal database of registered voters, some commercial vendors acquire state and local voter files, standardize them, and combine them into a single, national file. However, these commercial voter files tend to focus on domestic voters, and in some circumstances, they lack information about absentee ballot requests and/or overseas status.

Therefore, a sampling frame was created on FVAP's behalf, using two types of data sources: first, records of *confirmed absentee ballot requesters* ($N = 100162$), and second, records of *unconfirmed requesters* based on voter file records only ($N = 79700$). The lists

[1] Uniformed service voters comprise active duty members of the Uniformed Services, their spouses, and dependents.

of *confirmed absentee ballot requesters* were the primary source of records and were obtained by contacting the relevant officials in each state, as well as local election officials, as necessary. The resulting records were then matched to a commercial vendor's national voter file to append auxiliary variables, including individual-level history of voting in recent elections. The lists of *unconfirmed requesters* were for those states, counties, and/or towns that did not provide an absentee file. For these states and localities, the commercial firm searched existing voter records using custom database queries to locate voters with international addresses. To create the final sampling frame, the data files from these two types of sources were combined and cleaned of cases outside of the target population, that could not be contacted, or that were duplicate records.

Note that the rest of this paper focuses on the *confirmed absentee ballot requesters* portion of the sampling frame, as most of the sample was allocated to this portion of the frame. This allocation decision had been made since this portion of the frame was thought to have preferable coverage properties.

2.2.2 Sampling Design

Among confirmed absentee ballot requesters, a probability sample of size $n = 36000$ was drawn from a frame[2] of size $N = 100162$. As it pertains to this analysis, the most important feature of the sample design was the use of unequal probability sampling: specifically, the use of oversampling for requesters in countries or world regions with fewer members (including the selection with certainty for a subset), as well as selection with certainty for requesters registered in states with rare ballot policies. Sample stratification was also used to improve the balance of the sample with respect to key characteristics that were expected to relate to survey measures.

More specifically, the sample was drawn using Chromy's method of sequential random sampling (Chromy, 1979; Williams and Chromy, 1980), which is a probability-proportional-to-size selection method that implicitly stratifies based on a sorted list. The use of a probability-proportional-to-size selection method allowed for selection probabilities that varied by country, without the need to specify a large number of explicit strata.

To compute a country-level compromise allocation that would balance between domain and population estimation requirements, a compromise allocation was computed for 179 mutually exclusive groups (one group for voters in states with rare ballot policies, with the remaining 178 groups determined by country). This compromise allocation was designed to result in selection with certainty for voters in the smallest groups, equal allocation for medium-sized groups, and proportional allocation for the largest groups. For group g, the compromise sampling rate was computed as:

$$r_g = \begin{cases} 1, & \text{if } N'_g \leq 400 \\ \frac{400}{N'_g}, & \text{if } 400 < N'_g \leq 1860 \\ .20718\frac{N_g}{N'_g} \approx .215, & \text{if } 1860 < N'_g \end{cases}$$

where N_g is the total number of cases in the frame for group g, $\sum_{g=1}^{G} N_g = 100162$, and N'_g is the number of frame cases available for sampling for group g after excluding records that had been selected for the pilot survey, where $\sum_{g=1}^{G} N'_g = 95171$ records were available for sampling. As such, the compromise allocation rate was approximately monotonically nonincreasing with increasing group size; the cutoff point of 1860 and associated sampling

[2]For simplicity in reporting, we treat this as the full sampling frame (i.e., excluding the 79,700 records of unconfirmed requesters, for whom sampling was conducted independently).

rate of approximately $\frac{400}{1860} = .215$ for those meeting this minimum number of available frame cases had been determined implicitly via an iterative procedure.

Next, the frame was divided into eight mutually exclusive explicit strata, with one stratum for voters in states with rare ballot policies and all other voters stratified by world region.[3] The measure of size (MOS) for record i within group g was specified as $m_{gi} \approx r_g$, for $g = 1, ..., 179$, $i = 1, ..., N'_g$.[4] Using these explicit strata and measures of size, implicit stratification was achieved by applying Chromy's sequential algorithm to a list that had been sorted by voter participation history,[5] World Governance Indicators (WGI) index score, and domestic ZIP code, with imputation applied for missing data for sorting purposes. By specifying the stratum sample size as the total MOS within stratum, the probability of selection for a given record was equal to its MOS.

As a result of the unequal selection probabilities, the full-sample design effect from weighting (Kish, 1992; i.e., the unequal weighting effect), before nonresponse and calibration adjustments, was equal to $1 + L = 1 + \frac{\sum_{i \in S} (d_i - \bar{d})^2}{n(\bar{d})^2} = 1.36$, where $\bar{d} = \frac{\sum_{i \in S} d_i}{n}$ is the average design weight for the full sample, regardless of whether the sample members ultimately responded.

2.2.3 Data Collection

The survey was conducted using a push-to-web methodology. Sample members were invited to participate in the survey through mail invitations (supplemented by email invitations for sample members whose email addresses were available). Each sample member was contacted four times via mail (and up to four times via email, as applicable). The initial mail invitation and first follow-up letter each invited the sample member to reply online; a second follow-up letter was sent to initial nonrespondents, with a paper questionnaire and return postage and an envelope; and as a final reminder, a postcard was sent to the sample member's international and domestic addresses. This resulted in a final design-weighted response rate of 26% (calculated via AAPOR Response Rate 3; American Association for Public Opinion Research, 2015).

2.2.4 Nonresponse Patterns

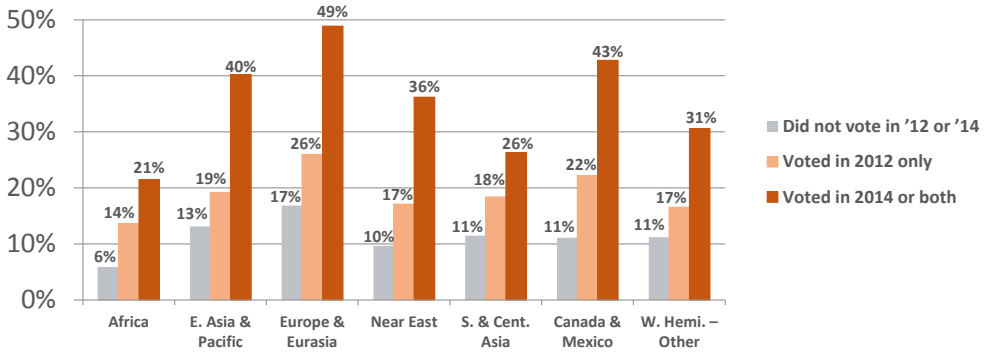
There are numerous factors that could potentially affect survey response in an international survey of overseas U.S. voters. Civic engagement has been shown to correlate with survey nonresponse in several contexts; the relationship between electoral and survey participation could cause overrepresentation of politically engaged voters in political surveys and might be a factor in turnout bias in post-election surveys (Barber et al., 2014; Brehm, 1993; Burden, 2000; Sciarini & Goldberg, 2015; Sciarini & Goldberg, 2016). Countries have different levels of infrastructure, with differing quality of roads, quality of postal systems, and/or internet availability, and thus, a sample member's country or world region could be associated with contactability. States and localities have different voter file practices, which

[3]The seven regions were Africa, East Asia and Pacific, Europe and Eurasia, Near East, South and Central Asia, Western Hemisphere – Canada and Mexico, and Western Hemisphere – Other.

[4]A minor adjustment had been applied to ensure that the total MOS in a given stratum was an integer. Specifically, for non-certainty cases, the MOS incorporated a flat multiplicative adjustment by stratum to ensure that $\sum_{i=1}^{N'_h} m_{hi} = round\left(\sum_{i=1}^{N'_h} r_{hi}\right)$, where m_{hi} and r_{hi} refer to the MOS and compromise allocation for record i within stratum h, for $h = 1, ..., 8$. Note also that due to domain precision requirements, cases in three strata were designated to be sampled with certainty, regardless of group size: Africa (non-rare policies); South and Central Asia (non-rare policies); and those in states with rare ballot policies.

[5]The three voter participation categories were based on having voted in the 2012 and/or 2014 general elections: (1) voted in both, (2) voted in only 2012 or 2014, and (3) voted in neither.

Figure 1: Response Rate by Region and Voter Participation History



could also affect record quality (Ansolabehere & Hersh, 2010; Berent, Krosnick, & Lupia, 2011) and could, thus, affect contactability. Civic engagement of voters could vary by state, given the differences in state and local requirements and procedures for registering to vote, requesting an absentee ballot, and/or remaining a registered voter. By affecting the composition of the frame population, these state and local factors could affect survey cooperation rates.

In this survey, unit nonresponse was strongly associated with voter participation, country of residence, and state of registration. Of particular note were differences in response rates by region and voter participation history (Figure 1), although there were also differences by state, WGI index score (which varies by country), and age. Given that these variables could be expected to covary with survey measures, an appropriately constructed weighting methodology was important for mitigating the risk of nonresponse bias.

2.2.5 Weighting Methods

The official survey weights were computed in five steps (FVAP, 2016) via methods analagous to those outlined by Valliant, Dever, and Kreuter (2013; Ch. 13–14):

1. Disposition codes were assigned to classify all sample members based on survey eligibility and completion. This resulted in classifying each sample member as an eligible respondent (ER), eligible nonrespondent (ENR), ineligible sample member (IN), or sample member with unknown eligibility (UNK).

2. Design weights were computed as the inverse of selection probabilities. Given that pilot sample members had been excluded from the main survey selection process, the main survey sample was treated as the second phase of a two-phase sample. For frame member $i \in U$, the unconditional probability of selection was $\pi_i = \left(1 - \pi_i^1\right)\left(\pi_i^{2|\neg 1}\right)$, where π_i^1 was the pilot probability of selection and $\pi_i^{2|\neg 1}$ was the probability of selection in the main sampling phase conditional on not being sampled in the pilot phase. Thus, the design weight for sample member $i \in S$ was equal to $d_i = \frac{1}{\pi_i} = \frac{1}{\left(1-\pi_i^1\right)\left(\pi_i^{2|\neg 1}\right)}$.

3. A response propensity adjustment was applied to account for sample members with unknown eligibility. More specifically, a logistic regression model, weighted by the design weights, $\{d_i\}$, was estimated to predict known eligibility (i.e., that the sample member was an ER, ENR, or IN). The predictors were voter participation history,

world region at time of mailing, age, age squared, WGI index score, and state.[6] These variables were selected based on their estimated relationships with response propensity and key survey variables. Known eligibility–adjusted weights were computed as:

$$w_i^K = \begin{cases} d_i/\hat{\phi}_i^K, & \text{for ERs, ENRs, INs} \\ 0, & \text{for UNKs} \end{cases}$$

where $\hat{\phi}_i^K$ was the model-estimated probability of having known eligibility.[7]

4. A response propensity adjustment was applied to account for nonresponse among sample members known to be eligible. A logistic regression model, weighted by the known eligibility–adjusted weights, $\{w_i^K\}$, was estimated predicting survey completion (ERs) among eligible sample members (ERs and ENRs). The predictors were voter participation history, age, age squared, and WGI mean. The completion-adjusted weights were computed as:

$$w_i^C = \begin{cases} w_i^K/\hat{\phi}_i^C, & \text{for ERs} \\ w_i^K, & \text{for INs} \\ 0, & \text{for ENRs} \end{cases}$$

where $\hat{\phi}_i^C$ was the model-estimated probability of completion among eligibles.

5. The weights were raked to control totals, which were population counts or estimated population counts from the sampling frame.[8] The calibration process included ERs and INs, since the control totals included both eligibles and ineligibles.[9] Each raking dimension incorporated a cross-classification with voter participation history (i.e., did not vote in the 2012 or 2014 general elections; voted in the 2012 General Election only; voted in the 2014 General Election or both; or missing voter participation data). Categories were collapsed in certain circumstances due to cell sparseness. The four raking dimensions were:

   (a) *Voter participation history* by *country*. Countries with fewer than 350 sample members were combined by world region before cross-classifying with voter participation history. Sample members with missing voter participation data were cross-classified by world region rather than by country due to cell sparseness. For six countries (China, Mexico, Singapore, South Korea, Sweden, and the United Arab Emirates), the category of individuals voting in the 2012 General Election only was combined with the corresponding category of those voting in neither general election, to avoid extreme adjustments.

[6] Voter participation history indicated whether the individual voted in the 2012 and/or 2014 general elections. States with fewer than 250 sample members were combined into a single group. Missing data for age was imputed to the mean, and indicator variables were included, as necessary, to reflect all missing data patterns for age and voter participation history.

[7] Cases that had been identified as ineligible at the full-sample level (i.e., due to having military addresses, U.S. addresses, or out of scope country addresses) were excluded from the known eligibility model and received an adjustment factor of 1.

[8] The control totals were population counts for raking dimensions (b) and (d). For raking dimension (a), a minor adjustment to the population counts had been applied to categories with cases from Dominica or Dominican Republic to correct for some initial misclassification. For raking dimension (c), the population counts incorporated the imputed values for sex to ensure internally consistent control totals and improve convergence.

[9] Before weighting, an additional round of frame cleaning was applied to identify any remaining individuals in the frame with military addresses that had not been previously identified. This additional frame cleaning allowed for these cases to be excluded from control totals and from entering the calibration process. This resulted in a final frame population of $N = 99750$, for purposes of calibration.

(b) *Voter participation history* by *state*. States with fewer than 250 sample members were combined into a single category before cross-classifying with voter participation history. Due to cell sparseness, individuals with missing voter participation history were combined across several states. Voter participation categories were combined into a two-way categorization within a limited number of states to avoid extreme adjustments.

(c) *Voter participation history* by *sex*. Imputation was applied for a small proportion of the frame (3.2%) with unknown sex, with imputed values determined primarily based on first name, middle name, and birthdate.

(d) *Voter participation history* by *age*. Due to cell sparseness, individuals who did not vote in either general election and who had missing age were combined with those who voted in the 2012 General Election only and who had missing age.

After calibration, the weights of eligible respondents and ineligibles conformed to control totals from the sampling frame; subsequently, ineligibles were excluded from survey estimates, because they were outside of the target population (therefore, eligible respondents were implicitly treated as a subpopulation of the frame population).

3. Simulation

3.1 Design

Using resampling methods, a Monte Carlo simulation was conducted to simulate various estimation designs. There were 18 estimation designs tested, with varied replicate sample sizes, calibration method, and level of weighting complexity (Table 1). For a given replicate r of replicate sample size t, for $t = 2250, 4500, 9000, 18000, 36000$, and $r = 1, 2, ..., R_t$, where $R_t = 5000 \cdot (36000/t)$ is the number of replicates of sample size t, a simple random sample with replacement (SRSWR) of size t was drawn from the initial sample of size $n = 36000$, and up to six weighting methods were applied, varying by complexity (low, medium, high) and calibration type (poststratification [PS], raking). Disposition codes and survey responses for a given sample member were treated as fixed, so that the simulation would reflect actual response patterns. With the smaller replicate sample sizes, only the simpler levels of complexity were used, as would be necessary in practice.

Table 1: Treatments Tested–Calibration Methods by Complexity and Replicate Size

	Replicate Sample Size				
	2,250	4,500	9,000	18,000	36,000
Low complexity	PS, raking	PS, raking	PS, raking	PS, raking	PS, raking
Medium complexity			PS, raking	PS, raking	PS, raking
High complexity					PS, raking

Before conducting the simulation, imputation methods were applied for all population members with item-missing data for voter participation history, sex, and/or age, primarily using hot deck imputation, with replacement, with donor cells formed in a manner that attempted to preserve key relationships.[10] To reduce computational complexity and remove a source of variable error, the imputed values were assumed to be true values for purposes of

[10]The one exception to the use of hot deck imputation was in imputing sex. First, sex was predicted based on first name and year of birth (where available). For those with unknown sex, these predictions were assumed to

the simulation. A subsequent robustness check indicated that this assumption had minimal effect on the resulting point estimates.

3.2 Weighting Conditions

The weighting conditions were analagous to the previously described official survey weighting procedures, with an adjustment to account for the replication subsampling stage and some simplifications applied in creating alternate adjustment schemes. That is, for replicate r of size t, individual i from the original sample S received an adjusted design weight equal to $d'_{(r,t)i} = d_i \delta_{(r,t)i} \frac{36000}{t}$, where $\delta_{(r,t)i}$ is the number of times i was selected for the given replicate, such that $\sum_{i \in S} \delta_{(r,t)i} = t$. The *high complexity raking* estimation scheme was roughly equivalent to the weighting scheme used for official survey estimates: a robustness check indicated that key sample estimates were nearly identical to the offical survey estimates. The *medium* and *low* levels of complexity involved simplifications to the nonresponse adjustment models and adjustment categories, as would be necessary in practice at lower sample sizes. Given that regional estimates were a primary focus of the survey and regional variation was a key aspect of the nonresponse patterns, the poststratification adjustment schemes were formed by simply poststratifying the nonresponse-adjusted weights based on the first raking dimension (vote history by country or region).

More specifically, weighting methods were as follows:

- *High complexity – raking*: The weights for individual i of replicate r and replicate sample size t were computed as $w_{(r,t)i} = d'_{(r,t)i} adj^K_{(r,t)i} adj^C_{(r,t)i} adj^R_{(r,t)i}$, where the adjusted design weight was multiplied by adjustment factors for known eligibility ($adj^K_{(r,t)i}$), survey completion ($adj^C_{(r,t)i}$), and raking ($adj^R_{(r,t)i}$), computed in a manner comparable to that previously described (albeit without the categories for missing data, given that missing data were imputed at the population level, with the imputed values assumed to be true for simulation purposes).

- *Medium complexity – raking*: The methods were the same as those in *high complexity – raking*, except with modifications to raking dimensions 1 and 2 (RD1 and RD2), such that the minimum sample size thresholds for country and state were increased to 1,200 (from 350) and 1,000 (from 250), respectively. That is, for RD1 (voter participation history by country), countries with fewer than 1,200 sample members in the original sample (of size $n = 36,000$) were combined into an *other* category for the given world region, before cross-classification with voter participation history; for RD2 (voter participation history by state), states with fewer than 1,000 members of the original sample were combined into an *other* category, before cross-classification with voter participation history.

- *Low complexity – raking*: The methods were the same as those in *medium complexity – raking*, except with simplifications to the first nonresponse model and three of the raking dimensions. For the first nonresponse model (i.e., for known eligibility), in which the predictors included state, the minimum sample size threshold for a state to be left as its own category (rather than being combined into an *other* group) was increased from 250 to 1,000. For calibration, RD1 was based on region rather than country; further, the regions of Africa, Near East, and South and Central Asia were treated as a single, combined region. For RD2, the minimum sample size threshold

be true, given that these methods correctly classified 97% of those with non-missing sex. Hot deck imputation was then applied for those who could not be classified using such methods, with cells based on predicted gender of middle name (when available) or voter participation history.

for state was increased from 1,000 to 4,000. Finally, RD4 was based on a four-way classification of age (rather than six-way).

- *High complexity – poststratification*: Same as *high complexity – raking*, except without raking dimensions 2–4. In other words, calibration solely consisted of poststratification to RD1 (voter participation history by country).
- *Medium complexity – poststratification*: Same as *medium complexity – raking*, except without raking dimensions 2–4; that is, calibration was poststratification to RD1.
- *Low complexity – poststratification*: Same as *low complexity – raking*, except without raking dimensions 2–4; that is, calibration was poststratification to RD1.

3.3 Analysis Methods

To assess the effect of the different weighting schemes, we examined the simulated design effects, mean squared error (MSE), variance, and bias across several survey analysis variables and domains. We selected survey analysis variables that might be prone to nonresponse bias in unadjusted estimates, and therefore, for which an effective weighting scheme was necessary for mitigating the risk of bias. Specifically, this included self-reported voter participation status in the 2010, 2012, and 2014 general elections (percent definitely voted), as well as self-reported reliability of country's postal system (percent reliable or very reliable) and quality of roads (percent high quality or very high quality). The domains assessed included the full population, world regions (seven categories), and voter participation status in the 2012 and 2014 general elections, based on administrative data (three categories: voted in neither; voted in 2012 only; voted in 2014 or both). This resulted in a total of 990 estimators of proportions (5 survey measures $\times$ 11 domains $\times$ 18 combinations of sample size by complexity by calibration method).

In simulating the bias for an estimator of a given population proportion P for some characteristic and domain (e.g., the rate of self-reported voting in the 2014 General Election among absentee ballot requesters in Africa), the full-sample weighted estimate via the *high complexity raking* method, $\hat{p}$, was treated as the true value.[11] Thus, for a given weighting scheme s and sample size t, the bias was simulated as $\mathrm{Bias}(\hat{p}_{(s,t)}) = \hat{\bar{p}}_{(s,t)} - \hat{p}$, where $\hat{\bar{p}}_{(s,t)} = \frac{1}{R}\sum_{r=1}^{R}\hat{p}_{(s,t,r)}$ and is the average estimate over the R replicates for the given design. Similarly, variance was simulated as $\mathrm{Var}(\hat{p}_{(s,t)}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{p}_{(s,t,r)} - \hat{\bar{p}}_{(s,t)}\right)^2$ and MSE was simulated as $\mathrm{MSE}(\hat{p}_{(s,t)}) = \mathrm{Var}(\hat{p}_{(s,t)}) + \left(\mathrm{Bias}(\hat{p}_{(s,t)})\right)^2$.

In assessing the resulting estimators, two types of design effects were computed. First, the design effect from weighting for a given set of weights was computed as 1 plus the relative variance of the final weights of eligible respondents.[12] Second, the design effect for a

[11] More specifically, P would be estimated via $\hat{p} = \frac{\sum_{i\in ER} w_i D_i I_i}{\sum_{i\in ER} w_i D_i}$, where w_i denotes the final weight of eligible respondent $i \in ER$, D_i is an indicator variable of domain membership that is 1 if $i \in D$ and 0 otherwise, and I_i is an indicator variable that is 1 if i has the population characteristic (e.g., self-reported voter in the 2014 General Election) and 0 otherwise.

[12] For weighting scheme s, replicate size t, and replicate r, let $R_{(str)i}$ equal 1 if the ith draw was an eligible respondent and 0 otherwise; let $\delta_{(str)i}$ denote the number of times the record associated with the ith draw was resampled. Then, $\mathrm{DEFF}_{w(str)} = \frac{\sum_{i=1}^{t} R_{(str)i}\left(w'_{(str)i} - \bar{w}'_{(str)}\right)^2}{e_{(str)}\bar{w}'^2_{(str)}}$, where $e_{(str)} = \sum_{i=1}^{t} R_{(str)i}$ is the replicate's number of eligible respondents, $w'_{(str)i} = \frac{w_{(str)i}}{\delta_{(str)i}}$ is the weight associated with a given draw, and $\bar{w}'_{(str)} = \frac{\sum_{i=1}^{t} R_{(str)i} w'_{(str)i}}{e_{(str)}}$ is the average weight of eligible respondents in that replicate, across draws. Note that this adjustment to weights for computing DEFF_w was necessary since a record with design weight d_i sampled δ_i times in a replicate resulted in one case with adjusted design weight $d_i\delta_i\frac{36000}{t}$, rather than δ_i cases with adjusted design weight $d_i\frac{36000}{t}$.

given population value Y being estimated was computed as $\mathrm{DEFF}(\hat{y}_{(s,t)}) = \frac{\mathrm{Var}\left(\hat{y}_{(s,t)}\right)}{\mathrm{Var}_{\mathrm{SRS}}\left(\hat{y}_{(s,t)}\right)}$, where the numerator was the simulated variance and the denominator was the variance that would have been obtained under simple random sampling.[13]

3.4 Hypotheses

We had two hypotheses:

H1. *Reducing the sample size will increase the design effects.* Under the traditional Hansen–Hurwitz estimator for sampling with replacement, $\hat{y}_{WR} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$, where p_i is the single-draw selection probability and we are averaging over n draws, and the sampling variance $\mathrm{Var}(\hat{y}_{WR}) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2$ is inversely proportional to the number of draws. However, given the use of nonresponse and calibration adjustments, which may be more variable at smaller sample sizes, we anticipated higher design effects for lower replicate sample sizes.

H2. *Coarsening the adjustment categories should increase the bias of estimators.* In our study, the *high complexity* weighting methods should best reflect the complicated nonresponse patterns, assuming the sample size is sufficient to allow them to be employed. We expected that coarsening the adjustment categories would reduce the homogeneity of cells, which could reduce the effectiveness of adjustments for reducing nonresponse bias. This may particularly be an issue for smaller domains.

4. Findings

4.1 H1 Results

In support of H1, the design effect from weighting increased with decreasing sample size, across weighting methods (Figure 2). The proportional increase in the design effect from weighting resulting from halving the sample size was larger at smaller sample sizes.

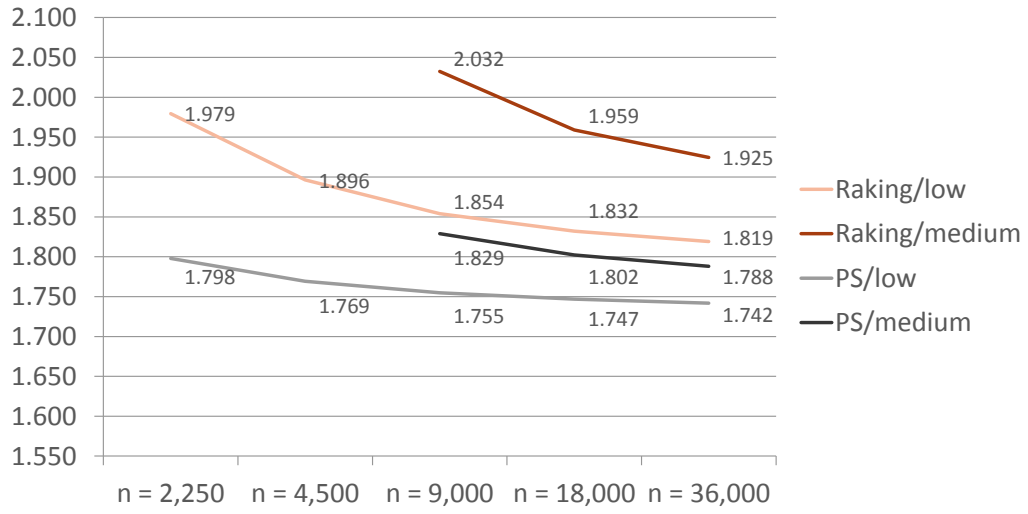
Similar findings were obtained for the design effects of the specific estimators tested. That is, lower sample sizes tended to be associated with higher design effects when holding constant the survey measure, domain, and weighting method. This effect was most pronounced at the smallest sample sizes—for replicates of size 2250 or 4500—whereas at larger sample sizes, the design effects by sample size were in some cases nearly indistinguishable. This is illustrated in Figure 3, which displays the average design effect across subpopulations for a given weighting method and replicate sample size. Design effects by subpopulation for self-reported vote status in 2014 are provided in the appendix; similar patterns were obtained for the other four survey measures tested.

4.2 H2 Results

In assessing H2, we first examined the simulated bias for estimates among the full population, by weighting method and survey variable, for replicates with sample size $t = 36000$, given that this was the only sample size for which we tested the high-complexity methods. None of the estimates had much bias, although the high complexity methods outperformed the medium and low complexity methods for estimating the proportion of respondents who

[13] $\mathrm{Var}_{\mathrm{SRS}}\left(\hat{y}_{(s,t)}\right)$ was estimated as $\frac{\mathrm{var}(\hat{y})}{\mathrm{deff}(\hat{y})} \cdot \frac{36000}{t}$, where $\mathrm{var}(\hat{y})$ and $\mathrm{deff}(\hat{y})$ were estimated in Stata 12 as the full-sample variance and design effect for the high complexity raking methodology, without subsampling, assuming an infinite population (given that replicate sampling was conducted with replacement), and taking into account the random nature of the number of subpopulation members sampled.

Figure 2: Mean Design Effect From Weighting $(1 + L)$ Across Replicates by Sample Size and Weighting Method



reported that their country's roads were of high quality (Table 2). Overall, these results suggest that the six weighting schemes provide fairly comparable results for full-population estimates.

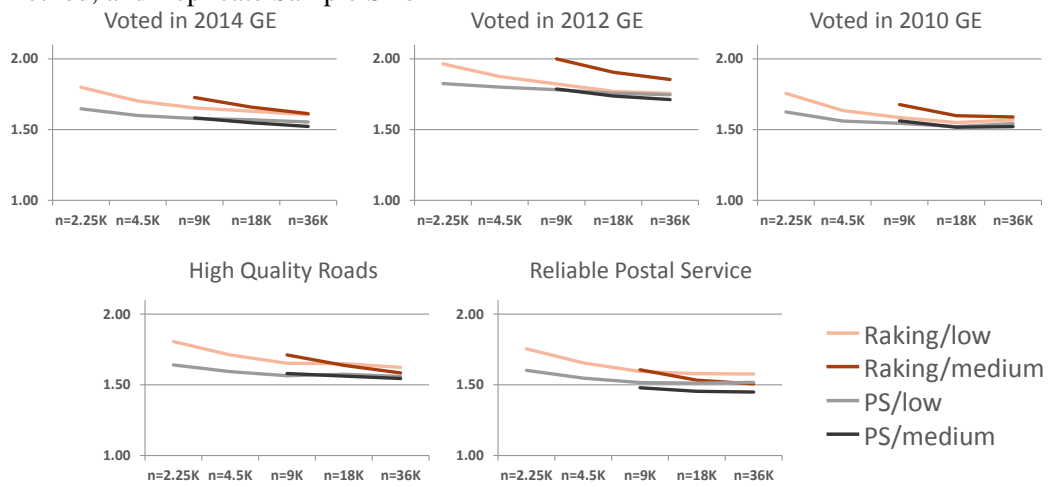
Table 2: Observed Bias by Question and Weighting Method (%; $n = 36000$)

		Question				
Complexity	Calibration Method	Reliable Mail	Quality Roads	Voted 2010 GE	Voted 2012 GE	Voted 2014 GE
Low	PS	0.2	-0.5	-0.1	0.2	0.2
Low	Raking	0.2	-0.6	-0.1	0.1	0.1
Medium	PS	-0.1	-0.6	0.1	0.2	0.1
Medium	Raking	0.0	-0.5	0.0	-0.1	-0.2
High	PS	-0.2	-0.1	0.2	0.4	0.3
High	Raking	0.0	0.0	0.0	0.0	0.0

Next, we examined the proportion of MSE attributable to variance; that is, $\mathrm{Var}(\hat{y})/\mathrm{MSE}(\hat{y})$. If bias was an unimportant part of the error, then this should have been close to 100%. Averaging again over replicates and then averaging over all survey measures and domains, we found that for all weighting methods, most of the error was attributable to variance rather than to the squared bias (Table 3).

Coarsening the adjustment categories increased the portion of the MSE attributable to the squared bias component, though the impact was smaller at lower sample sizes; that is, at smaller sample sizes, variance was a relatively larger component of the MSE. Further, nearly all of the error was attributable to variance when using the most complex adjustment method available for the given sample size (i.e., low complexity at sample sizes of 2,250 and 4,500; medium complexity for sample sizes of 9,000 and 18,000; and high complexity for a sample size of 36,000).

Figure 3: Mean Design Effect Across Subpopulations by Survey Measure, Weighting Method, and Replicate Sample Size



Note: Y-axis represents the average design effect, across domains, for a given survey measure (i.e., $\frac{1}{D}\sum_{d=1}^{D}\text{DEFF}\left(\hat{y}_d\right)$ for survey variable y and domain d). *GE* denotes general election.

Table 3: Average Proportion of MSE Attributable to Variance Across Measures and Domains by Replicate Sample Size and Weighting Method

	Low Complexity		Medium Complexity		High Complexity	
Sample Size	PS	Raking	PS	Raking	PS	Raking
2,250	98.2%	98.2%				
4,500	97.0%	97.0%				
9,000	94.3%	94.3%	95.5%	96.6%		
18,000	90.0%	90.0%	91.6%	93.8%		
36,000	83.6%	83.9%	86.3%	90.0%	91.9%	99.9%

5. Discussion

5.1 Conclusions

We found clear support for H1. For smaller replicate sample sizes, we found a larger design effect from weighting across the four weighting methods that were tested at multiple sample sizes. Similarly, the design effects for specific measures were higher at smaller sample sizes, when holding constant the survey measure, subpopulation, and weighting method.

We also found some weak support for H2, in that the low and medium complexity estimators exhibited a slight simulated bias for one of the five survey questions and the lower levels of weighting complexity were associated with a higher portion of MSE attributable to the squared bias for larger sample sizes. However, the level of bias was very small across estimators, as to not be practically meaningful, particularly given that the larger sample sizes allowed for a higher level of complexity of the adjustment scheme. Due to computational constraints, our H2 analysis focused on overall estimates and larger domains. Given that there may be larger differences at lower levels of aggregation (e.g., country), further work is needed to assess whether the bias component may be meaningful for smaller domains.

5.2 Future Directions

In a study with complex nonresponse patterns, we found that sample size reductions led to increases in the design effect for estimators that incorporated nonresponse and calibration adjustments. This effect was likely due to the size of the adjustment cells, given that smaller cells could lead to increased weight variability. Work should be done to replicate this finding in other survey contexts and to better understand the conditions under which sample size reductions lead to meaningful increases in the design effects. If these findings persist in other survey settings, then this would have direct implications for sample size and sample allocation calculations when incorporating weighting adjustments for nonresponse.

We also note that the strong relationship between response rates and voter participation in our study could have major implications for election-related studies, if replicable in other surveys. For post-election surveys, unadjusted estimates of voter turnout could be subject to large nonresponse biases. This implies the utility of using administrative data on electoral participation for adjustment purposes. For political polls, estimators that do not explicitly account for a relationship between response propensity and voting propensity may be biased. For such studies, this implies the utility of using a voter-file-based sampling frame, to allow the researcher to explicitly model and account for this relationship.

6. Authors' Note

This research was based on data from the Federal Voting Assistance Program (FVAP). The views, opinions, and findings contained in this paper are solely those of the authors and should not be construed as an official U.S. Department of Defense position, policy, or decision, unless so designated by other documentation.

REFERENCES

The American Association for Public Opinion Research (2015), "Standard definitions: final dispositions of case codes and outcome rates for surveys," 8th edition, Oakbrook Terrace, IL: AAPOR.

Ansolabehere, S., & Hersh, E. (2010), "The quality of voter registration records: a state-by-state analysis," report, *Caltech/MIT Voting Technology Project*.

Barber, M. J., Mann, C. B., Monson, J. Q., & Patterson, K. D. (2014), "Online polls and registration-based sampling: a new method for pre-election polling," *Political Analysis*, 22, 321–335.

Berent, M. K., Krosnick, J. A., & Lupia, A. (2011), "The quality of government records and over-estimation of registration and turnout in surveys: lessons from the 2008 ANES Panel Study's registration and turnout validation exercises," *American National Election Studies*, working paper nes012554.

Brehm, J. (1993), "The phantom respondents: opinion surveys and political representation," *University of Michigan Press*.

Brick, J.M. (2008), "Unit nonresponse and weighting adjustments: a critical review," *Journal of Official Statistics*, 29, 329–353.

Brick, J.M. & Jones, M. (2008), "Propensity to respond and nonresponse bias," *Metron - International Journal of Statistics*, LXVI, 51–73.

Burden, B. C. (2000), "Voter turnout and the national election studies," *Political Analysis*, 8, 389–398.

Chromy, J. R. (1979), "Sequential sample selection methods," in *Proceedings of the American Statistical Association*, Survey Research Methods Section.

Deming, W. E., & Stephan, F. F. (1940), "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *The Annals of Mathematical Statistics*, 11, 427–444.

Deville, J. C., & Särndal, C. E. (1992), "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, 87, 376–382.

Federal Voting Assistance Program (2016), "Overseas citizen population analysis: volume 2: survey administration and methodology," technical report, `https://www.fvap.gov/uploads/FVAP/Reports/FVAP-OCPA_201609_final.pdf`.

Flores Cervantes, I., & Brick, J.M. (2016), "Nonresponse adjustments with misspecified models in stratified designs," *Survey Methodology*, 42, 161–177.

Holt, D., & Smith, T.M.F. (1979), "Post stratification," *Journal of the Royal Statistical Society, Series A*, 33–46.

Little, R.J. (1986), "Survey nonresponse adjustments," *International Statistical Review*, 54, 139–157.

Little, R.J. & Vartivarian, S. (2003), "On weighting the rates in non-response weights," *Statistics in Medicine*, 22, 1589–1599.

Kott, P.S. (2009), "Calibration weighting: combining probability samples and linear prediction models," in *Handbook of Statistics 29B*, 55–82.

Kalton, G. & Maligalig, D. (1991), "A comparison of methods of weighting adjustment for nonresponse," in *Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference*, 409–428.

Kish, L. (1992), "Weighting for unequal Pi," *Journal of Official Statistics*, 8, 183–200.

Rizzo, L., Kalton, G., & Brick, J.M. (1996), "A comparison of some weighting adjustment methods for panel nonresponse," *Survey Methodology*, 22, 43–53.

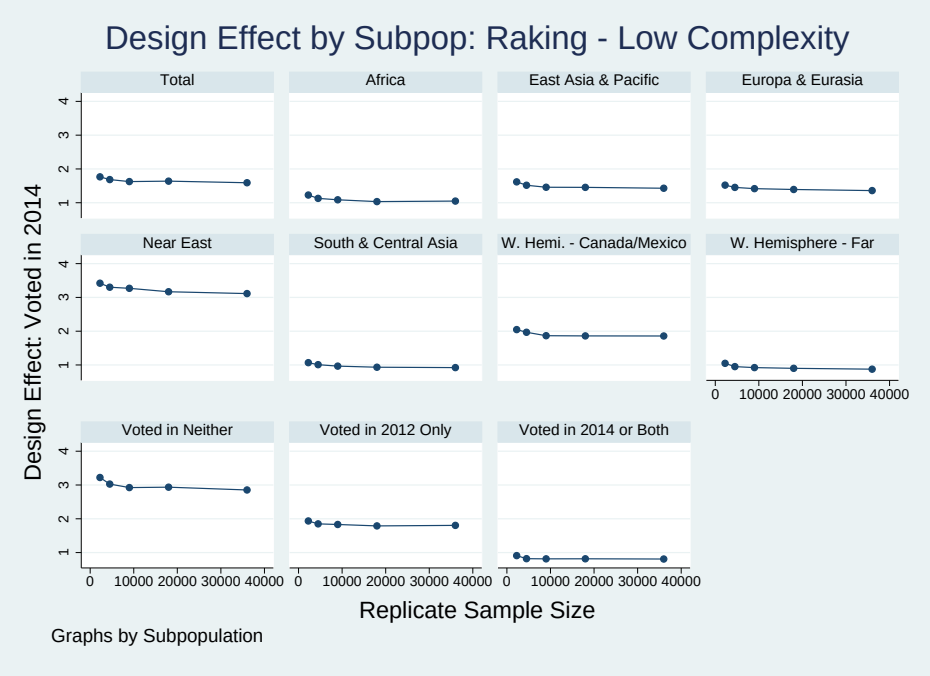
Sciarini, P., & Goldberg, A. C. (2015), "Lost on the way: nonresponse and its influence on turnout bias in postelection surveys," *International Journal of Public Opinion Research*.

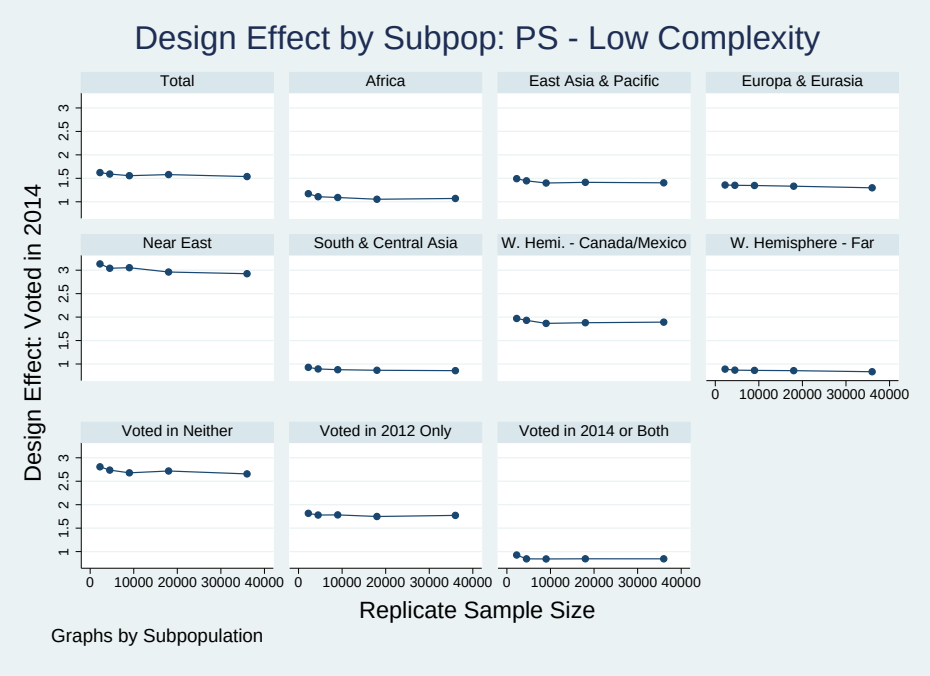
Sciarini, P., & Goldberg, A. C. (2016), "Turnout bias in postelection surveys: political involvement, survey participation, and vote overreporting," *Journal of Survey Statistics and Methodology*, 4, 110–137.

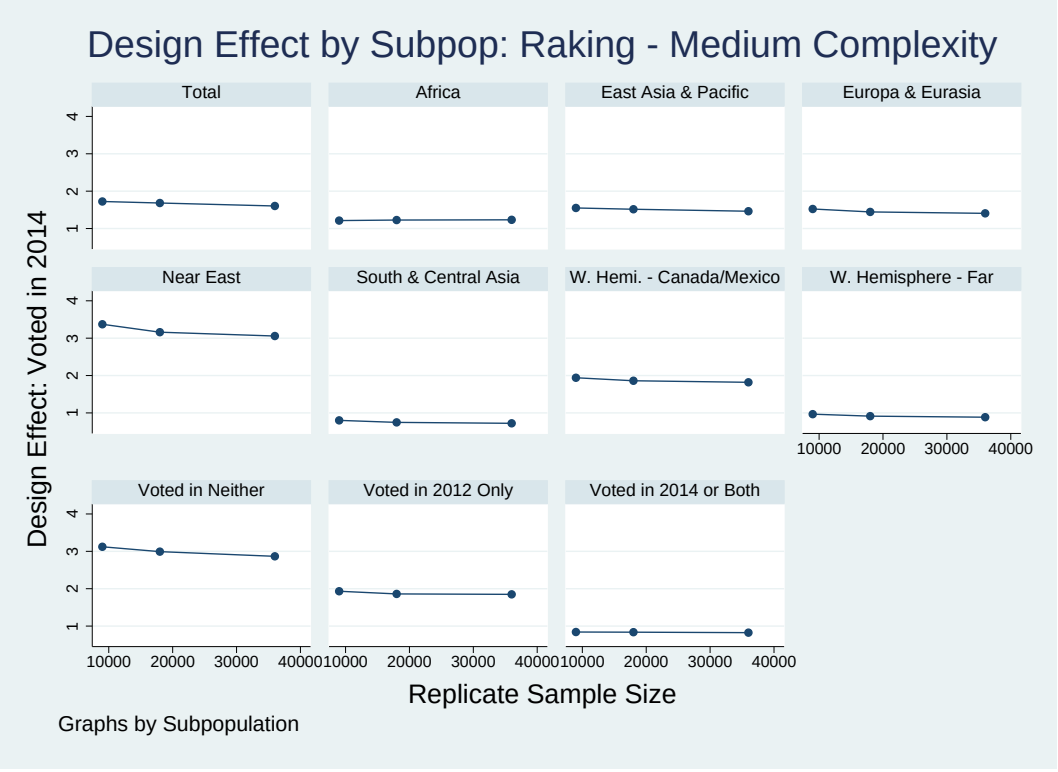
Valliant, R., Dever, J. A., & Kreuter, F. (2013), *Practical tools for designing and weighting survey samples*, New York: Springer.

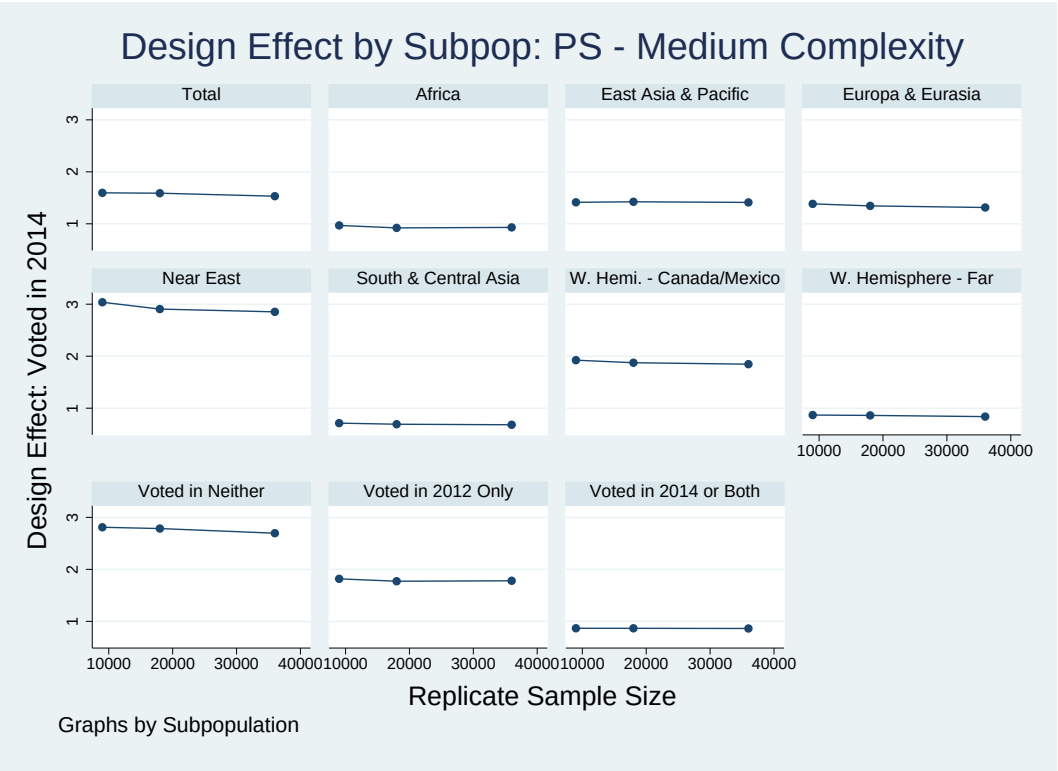
Williams, R. L., & Chromy, J. R. (1980), "SAS sample selection macros," in *Proceedings of the Fifth Annual SAS Users Group International Conference*.

7. Appendix: Design Effects by Weighting Method






Design Effect by Subpop: Raking - Medium Complexity
Total
Africa
East Asia & Pacific
Europa & Eurasia
Near East
South & Central Asia
W. Hemi. - Canada/Mexico
W. Hemisphere - Far
Voted in Neither
Voted in 2012 Only
Voted in 2014 or Both
Design Effect: Voted in 2014
Replicate Sample Size
Graphs by Subpopulation


Design Effect by Subpop: PS - Medium Complexity
Total
Africa
East Asia & Pacific
Europa & Eurasia
Near East
South & Central Asia
W. Hemi. - Canada/Mexico
W. Hemisphere - Far
Voted in Neither
Voted in 2012 Only
Voted in 2014 or Both
Design Effect: Voted in 2014
Replicate Sample Size
Graphs by Subpopulation