RESEARCH REPORT SERIES
(Statistics #2022-03)

Methodology and Theory for Design-Based Calibration of Low-Response Household Surveys with Application to the Census Bureau 2019-20 Tracking Survey

Eric V. Slud[1,2],
Darcy Morris[1]

[1] Center for Statistical Research and Methodology, U.S. Census Bureau;
[2] Department of Mathematics, University of Maryland College Park

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: June 22, 2022

June 22, 2022

Methodology and Theory for Design-Based Calibration of Low-Response Household Surveys

with Application to the Census Bureau 2019-20 Tracking Survey

Eric Slud and Darcy Morris, Census Bureau CSRM

ABSTRACT: Motivated by the Census Bureau's 2019-2020 Tracking Survey, conducted in two modes as a telephone probability survey and also as a nonprobability web survey, this report studies methodological issues concerning household survey estimation following weight-adjustment. The issues of interest include: the creation of base-weights for telephone RDD (random-digit-dialing) surveys; the adjustment of base-weights in probability and nonprobability surveys by generalized raking methods to calibrate respondents to national population targets for so-called 'post-stratification' variables; the handling of these base-weighting and calibration methods when some of the respondents' post-stratification variable values are missing; the mathematical theory justifying the large-sample behavior and variance estimation for survey estimates; and assessment of the success of weight-adjustment in making survey respondents represent the US population. Well-established theory (Deville and Särndal, 1992 JASA) says that when inverse-inclusion-probability survey weights are calibrated to true totals in a probability survey with complete response, the design-based estimates are design-consistent and have reduced variances, and the weights move very little. Yet in raking real low-response surveys like the Tracking Surveys, the movement of most weights is large. New design-based theory provided in this report justifies generalized raking in settings where the correct weights (defined here for the first time in a design-based framework) satisfy a parametric model, and large-sample theory is established for adjusted-weight survey estimators and their variance estimates whether such a model holds or not. This theory makes precise a sense in which weight calibration and variance estimation in surveys with either low response or non-probability designs is unavoidably model-based. To assess the validity of survey weight adjustments made by potentially misspecified models, survey methodologists often compare with known national targets the survey-weighted estimated proportions with respect to benchmark variables not used in the weight adjustment. In the Tracking Survey, this is done using the raked weights for several benchmark variables and also for augmented demographic benchmarks obtained by cross-classifying demographic variables whose marginal totals have been used in calibration.

Key words and phrases: survey calibration, nonresponse adjustment, raking, nonprobability survey, benchmark variables, item-missing data

1 Introduction

The general setting for the survey inference problems we consider is a frame population list $\mathcal{U}$ of N units, n units of which are respondents in a designed probability sample (of a larger number of units) drawn from $\mathcal{U}$. Unit nonrespondents from the larger designed sample are ignored, or else no information about them is available, and we allow the possibility that the respondent-set $\mathcal{R}$ is much smaller than the set $\mathcal{S}$ originally sampled. For purposes of survey inference, we maintain the fiction that $\mathcal{R} = \mathcal{S}$, a fiction that forces us to modify the design base-weights, the original single-inclusion probabilities $\pi_i^o = P(i \in \mathcal{S})$, to reflect the response mechanism as part of the sampling design. It is assumed that the design weights $w_i^o = 1/\pi_i^o$ are available for all $i \in \mathcal{R}$, along with the survey outcomes Y_i and covariates X_i for the respondent units. The covariates X_i are vectors of demographic and other observations on survey respondents, which may be categorical or numeric, and some of which may be missing for some survey respondents.

The setting described here applies to many political and social-science probability surveys, including random-digit-dialing (RDD) telephone surveys. Characteristic features of such surveys include (i) the low response rate (small size of n as a fraction of the total number of residential telephone numbers initially 'sampled' and called), and (ii) the definition of the frame $\mathcal{U}$ as the union of multiple frames, which in the RDD setting are the separate frames of landline and cell telephones. The proper definition of the base-weights for units in $\mathcal{U}$ requires knowledge of the national population of households with telephones of each type, landline and cell, as well as the proportion of the total with phones of both types. It is assumed that essentially 100% of the US population has some kind of telephone service – although that may not be correct for the homeless or for some rural and American Indian areas. All respondents to the 2019-2020 Tracking Survey were asked questions about standard demographic categories (Age, Sex, Educational level, Race group, Owner/renter status, geographic location [from which are derived Census Region and population density of household location], Household size (# adults), Marital Status, whether a Language other than English is spoken in the home), together with a few benchmark questions related to Volunteering behavior, Blood donation, and Health. For all these survey variables, there are high-quality national surveys from which national proportions of the outcome categories and some low-order cross-tabulations of them can be derived.

A standard approach to the analysis of surveys like those described above, including surveys with low response-rate, is to adjust the design- or base- weights subject to the constraints that weighted survey totals of population subsets (domains) defined by specified 'poststratification variables' must agree with the totals from reliable estimates from large national surveys or the census (Valliant, Dever and Kreuter 2018)[1]. That external source of national totals is generally the demographic Population Estimates produced yearly by the

[1]In this report, the terms post-stratification and calibration will generally be used synonymously.

Census Bureau as an update to the decennial census, augmented by variables tabulated in the yearly American Community Survey or Current Population Survey which themselves have been adjusted to agree with the Population Estimates. Nowadays, the same methods are applied (with constant base-weights, *cf.* Valliant et al. 2018) to surveys that are obtained from questionnaires administered online to self-selected internet panels in which broad invitations are issued but no probability design is used to sample potential respondents.

The specific survey that motivated this report is the 2020 Census Tracking survey that was designed to help inform advertizing decisions during the collection of the 2020 Census. This survey collected daily data (aggregated monthly) on noninstitutionalized US adults (aged 18 and above) from September 2019 through June 2020 on awareness, attitudes, self-reported intent to participate in the U.S. 2020 Census, perceptions about how data would be used, selected topics on messaging, and other questions about major events happening during 2019-2020. The survey was conducted by a contractor (Team Y&R 2020), and the data are publicly available through the University of Michigan Inter-university Consortium for Political and Social Research. A similar survey had been conducted in 2010. This current survey used two different samples and data collection modes, analyzed and described here as separate surveys. One was a nationally representative probability telephone survey and the other a non-probability opt-in web survey. Both had very low response rates ($< 8\%$), and the raw demographics of respondents in both surveys differed markedly from the general US population. In both surveys,'survey weights' were derived (initially by the contractor, using standard raking adjustments, a popular method of calibration or 'poststratification') with the goal of making the respondents better represent the US population.

This report is organized as follows. Section 2 is an overview of current research topics in survey methodology bearing on the choice of weight-adjustment method for a (low-response-rate) probability survey. As part of this overview, we describe alternative types of weight adjustment and briefly tell what methods exist to estimate variances of survey totals estimated with these methods. In the last heading of the section are previewed some standard methods of assessment of adjusted weights in making a survey's respondent population representative of a targeted national population. Section 3 explains how base-weights are derived in RDD surveys with some missing information relevant to determining the types of telephones accessible to respondent households. Section 4 describes alternative methods of handling occasionally missing poststratification variables in implementing otherwise standard 'generalized raking' weight-adjustments. The formal mathematical assumptions and theory supporting survey inference following generalized-raking of low-response surveys (or other surveys requiring large adjustments, such as those with frame coverage errors) is given in Appendices A-B. The theory is exclusively design-based, treating the finite population surveyed as a large array of unknown but nonrandom individual-level covariates $\mathbf{X}_i$ and outcomes Y_i. Probability sample design fully describes the mechanism of choice of sampled units, but in surveys with low response-rate there is a further mechanism, unknown and unmodeled, by which a subset of sampled units become respondents. The formal theory

provides natural, but apparently new, assumptions defining when the set of *final weights* observable along with respondent attributes can serve as large-sample asymptotically correct weights. The theory also shows how parametric assumptions related to the weight-movement metric G used in defining generalized calibration enable conclusions about survey estimates and variances comparably general to those of Deville and Särndal (1992) but in settings with large weight-adjustments. Sections 5 and 5.1 respectively interpret the application of the formal and technical results to survey totals and variance estimates in real surveys like the Tracking Survey after weight-adjustment by calibration. Section 6 illustrates the theory of the earlier sections on data from the RDD and Web Tracking surveys and implements several data-analytic assessments of the adequacy of the weight-adjustments, both those made by the contractor and by the methods of this paper, in the Tracking Survey. Section 7 provides a summary discussion and conclusion of the report, also sketching implications of the paper's theory and assessment methodology for the choice of benchmark variables.

The novel methodological elements of this report are: (i) a method of variable-by-variable treatment of missing item data in base-weighting (Sec. 3) and generalized-raking weight adjustment (Sections 4, 4.3 and 4.4); (ii) new design-based large-sample theory for generalized-raking weight adjustment when either a parametric model-assumption holds for the ratio of true over base weights or when no such model holds (Appendix and Sec. 5); (iii) valid variance formulas for survey-weighted estimates in (ii) when the sampling before weight-adjustment is arguably Poisson (Appendix and Sec. 5.1), and (iv) use in Sec. 6.2 of a metric for discrepancies of adjusted-weight estimates from targets using cross-classified categorical variables to approximate the total-variation distance, and (v) statement in Sec. 7 of analytical criteria for choice of effective benchmark variables.

2 Current research issues in survey weight-adjustment

For the data setting of Section 1, a first task is to modify the initial or base weights w_i^o (generally taken to be identical for all respondents i in a nonprobability survey) to reflect known national category proportions for a set of demographic and geographic variables. The process of modifying these inital weights to a set w_i of final weights is called *weight-adjustment*, with the goal of enabling estimates of population totals t_Y of survey attributes Y_i to be calculated from survey-respondents in the survey-weighted form $\hat{t}_Y = \sum_{i \in \mathcal{R}} w_i Y_i$. The modifications from w_i^o to w_i are generally made with the aid of non-constant auxiliary numeric variables (covariates) $\underline{X}_i = (X_{i,k}, k = 1, \ldots, p)$ to satisfy exact or approximate calibration constraints

$$N^{-1} \sum_{i \in \mathcal{R}} w_i X_{i,k} \;=\; \bar{X}_k \quad \text{for} \quad k = 0, \ldots, p \tag{1}$$

where the target population means $\bar{X}_k = t_{X_k}/N$ are known from external sources, and we define $X_{i,0} \equiv \bar{X}_0 \equiv 1$. We consider various weight-adjustment schemes of this sort,

summarizing from the survey research literature and extending where necessary to account for the possibility that some respondent covariate items $X_{i,k}$ may be missing.

Two different kinds of model underlie weight adjustment. The first concerns frame coverage, positing that the frame may be faulty but respondents to the survey are 'accessible' in the sense of being correctly identified on the frame list, reachable by the mode(s) of survey delivery and able to accept the survey burden and the way in which questions are asked. The second viewpoint is that the frame is correct but that the characteristics of individuals predispose some to respond to the survey and others not. A shorthand for these different formulations is that they respectively address frame coverage and propensity to respond. The first suggests models in which survey weights (inverse inclusion probabilities) differ from intended weights in the general population by a biasing mechanism involving only those characteristics (typically, geography or address-type) that would be known in advance from the frame, while the second suggests models for the probability of response in terms of individual characteristics that would not be known before the subject responds.

The research issues that are most important for the present overview are:

(I). Models versus Metrics for w_i/w_i^o

A modeling idea generally ascribed (by survey methodologists) to Oh and Scheuren (1983) is the 'post-randomization model' that treats the decision to respond for a sampled individual as a binary decision independent of data all other individuals conditionally given a covariate vector Z_i of individual variables, observable or not. This idea is taken up by biostatisticians and social scientists in the form of a parametric model for conditional probabilities $P_\theta(i \in \mathcal{R} \,|\underline{Z}_i;\, i \in S)$, where $\underline{Z}_i$ is a vector of covariates including any used for calibration as $X_{i,k}$. A different, apparently nonparametric, optimization-based approach is *generalized-raking calibration* as formulated by Deville and Särndal (1992) to minimize over final weights w_i a metric $\sum_{i\in\mathcal{R}} G(w_i/w_i^o)$ subject to (1). Economists and biostatisticians (*cf.* Tsiatis 2006) are interested in efficient estimation of parameters, calibration theorists in qualitative nonparametric results not depending heavily on the choice of G. These approaches are similar when the weights w_i^o are very far from those needed to satisfy the constraints (1).

(II). Missing Survey Items $X_{k,i}$ – Model-based Imputation versus Complete-case analysis

Survey methodologists broadly recognize the need for a systematic *imputation* method of filling in missing items, either singly for each covariate or sequentially for jointly missing items. Indeed, widely used survey software (Lumley 2010) assumes that there are no missing items, i.e., that they have already been filled in before calibration or that calibration omits respondent records containing any missing data. Model-based imputations are used by various authors interested in Multiple Imputation (Carlin 2015): a current set of techniques known as *Chained Imputation* is embedded in the `MICE` software of van Buuren (2015). Survey methodologists preparing for weight-adjustment often fill in missing items using item-wise hot-deck imputation methods (Andridge and Little 2010) with a nonpara-

metric flavor. An idea tried in the `ANESrake` R-package (Pasek et al. 2014) in the context of raking is that each *Iterative Proportional Fitting* pass requires only the filling-in or omission of missing items in a current raking variable, and we generalize that idea below to the class of generalized-raking calibration methods of Deville and Särndal (1992). It turns out that this method is equivalent to mean-imputation within each raking pass.

(III). Alternative Choices for Calibration Variables and Methods

Almost all of the various ratio-adjustment, linear-calibration, raking and other post-stratification adjustment methods of weight-adjustment in the survey methodology literature fit within the model-based or generalized-raking methods summarized in paragraph (I) above (Valliant et al. 2018). (One example of a method, widely cited in the Causal Inference literature, that is not acknowledged to fall in the latter rubric, is 'entropy balancing' calibration (Hainmuller 2012), but we will see below that this is another instance of generalized raking and is subject to the advantages and disadvantages of such calibration methods.) In all, there is a choice of 'post-stratification' variables to use. In the calibration methods of Deville and Särndal (1992) and Deville, Särndal and Sautory (1993), there is a choice of metric G. Some of those metrics enforce weight-ratio trimming (uniform upper and lower on w_i/w_i^o), but for those that do not (including ordinary raking), some practitioners do further 'weight-trimming'. In addition, there is the possibility of allowing more post-stratification variables to be used by relaxing some of the constraints (1) to be 'soft' in the sense that they are not forced to hold exactly but a penalty term containing a weighted sum of squared discrepancies is added to the metric $\sum_{i\in\mathcal{R}} G(w_i/w_i^o)$ being minimized over $\{w_i\}_{i\in\mathcal{R}}$ (Slud and Thibaudeau 2010). This last idea is familiar to survey methodologists from the Benchmarking literature, but is not usually presented as a weight-adjustment approach.

(IV). Variance Estimation of $\hat{t}_Y$ after Weight-adjustment

Variance estimation following imputation and weight-adjustment is still largely an open problem. Rao and Shao (1992) provided a method of variance estimation for randomized hot-deck imputation (without calibration of weights, but involving some adjustment of weights related to the imputation), and Deville and Särndal (1992) justified the use of GREG variance formulas following generalized-raking calibration under the assumption that the base-weights w_i^* are correct inverse-inclusion probabilities. Perhaps practitioners of Multiple Imputation hope that their variance estimates apply equally well when weights are adjusted following model-based adjustment methods such as MICE, but that hope has so far not been theoretically justified. In the present report, variance formulas valid for large samples are provided in a generalized-raking context under an assumption that the base-weights differ by a parametric model from the correct base-weights. Those results, demonstrated in Appendix B and summarized in Section 5.1, are the material of this report with the greatest methodological novelty, but they do not address the level of variability due to missing covariate items.

(V). Assessment of Weight-adjustments

An important element of survey measurement methodology is the comparison of survey-weighted estmates of national outcome proportions from benchmark responses versus proportions measured by well-established high-response-rate national surveys (Valliant et al. 2018). Such external measures of quality are needed particularly in surveys with low respnse-rate, as part of a general assessment of survey biases. This kind of comparison has been a staple of the survey measurement literature comparing the quality of results of low-response-rate probability surveys versus nonprobability (web) surveys, as in MacInnes et al. (2018), Pasek and Krosnick (2020) and Yeager et al. (2011). We return to these quality assessments in Section 6 in relation to the two arms of the Tracking Survey, and as part of a more general discussion in Section 7 of what constitutes an effective benchmark variable.

3 Base-Weights in the RDD Phone Tracking Survey

It is customary in RDD surveys to ignore unit-level nonresponse, or rather to treat it as an essential aspect of the sampling design. The 'base weights' w_i^o are the reciprocals of the effective single-inclusion probabilities, and this Section describes their construction in the context of the 2019-2020 Phone Tracking Survey. The standard approach, chosen by Team Y&R, is that of Buskirk and Best (2012), applicable to a survey receiving data from (at most) one respondent adult (aged 18+) in each household, with respondents viewed as having been a Simple Random Sample from the frame (Cell-phone or Landline) within which they were sampled. Formulas (3)-(4) of Buskirk and Best (2012) are elaborated here because some data are missing regarding household size and dual landline-cellphone status in the Phone Tracking Survey, and our treatment of the missing data is non-standard. 'Household size' is used here only for landlines and refers to all adults accessible by that landline. It is acknowledged that residential situations with multiple landlines are not adequately addressed.

The basic assumptions of Buskirk and Best (2012), which we adopt here, are

(BB1) all adults in a household are equally likely to be sampled in a landline call,

(BB2) each cell-phone can reach only a single adult, and

(BB3) sampling from the landline and cell-phone frames is independent.

Assumption (BB1) is reasonable in the Tracking Survey because data are collected (subject to some callback nonresponse) on the adult in the household with the next birthday. Assumption (BB2) is common in RDD work, although anecdotal evidence suggests it underestimates the number of adults actually reachable, and therefore will lead to baseweight over-estimates. (BB3) correctly reflects the actual conduct of RDD surveys.

Because dual-telephone status of respondent persons and numbers of adults in respondent landline households are sometimes missing (as is true respectively of 1.24% and 2.43% of respondents in the Phone Tracking Survey), we require two further assumptions about the ignorability of the mechanism causing missing data.

(BB4) The proportion of landline households who would fail to provide number of adults if sampled is the same among respondents and nonrespondents, and

(BB5) Within each of the subpopulations of adults reachable by landline and those reachable by cellphone, the fraction of persons who would fail to provide information about dual cell/landline telephone status is the same among respondents and nonrespondents.

(BB4) seems a harmless assumption, except that callback for nonresponse might be more frequent in larger households. Similarly, (BB5) might fail in larger households in which some household adults do not know which other adults have cell-phones. Since (BB2) might also be less valid in larger households, we recommend – after using (BB1)-(BB5) in constructing base-weights – to use household size as a poststratifying variable in constructing final adjusted weights for RDD telephone surveys, and this is apparently not standard practice.

We assume that the overall size of the adult population N_{pop} with telephones is known, along with the 'target' national proportions (p_{CO}, p_{LO}, p_D) of persons reachable by telephone in the nation with Cellphone-only, Landline-only, or Both (Dual). The latter proportions may be taken from a source such as the National Health Interview Survey (Blumberg and Luke 2018). The 2019 estimated adult US population was 209.1 million, and Table 1 of Blumberg and Luke (2018) give the proportion of the national adult population in the period Jan-June 2018 without telephones as 3.2% and with unknown Dual status as 0.1%, and 55.2% with Cell-phone only, 4.1% Landline only, and 37.4% with both.

In terms of these target telephone-status proportions, we define respective population fractions p_L in the Landline and p_C in the Cellphone universes as

$$p_L = p_{LO} + p_D \quad , \qquad p_C = p_{CO} + p_D$$

In this Section, we treat the respondent-set $\mathcal{R}$ and sample $\mathcal{S}$ as identical, indexed by i or j, and let n denote the size of this set. If there were no missing household-size or dual-status information, then based on respective numbers n_L, n_C, and n of Landline, Cell-phone and Total respondents, Buskirk and Best's (2012) formula (3) defines base-weights in three steps:

(i) For each respondent i sampled by landline telephone, the inclusion probability is

$$\pi_i^L \; = \; \{n_L/(p_L \cdot N_{pop})\} \, / \, \texttt{ADULTS}_i \tag{2}$$

where $\texttt{ADULTS}_i$ denotes the number of adults living in the household with person i.

(ii) For each respondent i sampled by cell-phone, the inclusion probability is

$$\pi^C \;=\; n_C/(p_C \cdot N_{pop}) \tag{3}$$

(iii) Unnormalized weight w_i^o for individual i is the reciprocal of the overall inclusion probability taking account of the status of i belonging to the groups CO (Cell-phone only), LO (Landline only), or D (dual), and $\bar{w}_i^o$ are normalized to average 1 among respondents:

$$w_i^o \;=\; 1 \Big/ \left(I_{[i\in LO\cup D]} \cdot \pi_i^L \;+\; I_{[i\in CO\cup D]} \cdot \pi^C \;-\; I_{[i\in D]} \cdot \pi_i^L \cdot \pi^C \right), \qquad \bar{w}_i^o = \frac{n\, w_i^o}{\sum_j \, w_j^o} \tag{4}$$

Buskirk and Best (2012, in their formula 4) provide a simplified version of formula (4) with the third denominator term in w_i^o removed, because that third term is generally very small. (In the Phone Tracking Survey, $n_L = 20528$, $n_C = 41082$, $n = 42161$, and the L and C inclusion probabilities are small enough that their product is negligible.)

3.1 Case of Missing Data

It remains only to explain the modifications of formula (4) when some respondent individuals i are missing the information D (i.e., missing $I_{[i\in D]}$), and some individuals sampled in the landline frame are missing $\texttt{ADULTS}_i$. For clarity, let LL and CP respectively denote the sets of persons in the land-line and cell-phone frames.

Step 1°. First, the frame population proportions of dual phone status $p_{D|L} = p_D/p_L$ and $p_{D|C} = p_D/p_C$ are estimated from the non-missing data by

$$\hat{p}_{D|L} \;=\; \hat{P}(i \in D \,|\, i \in LO \cup D) \;=\; \frac{\sum_j \, I_{[j\in D]}}{\sum_j \, I_{[j\in LL]}}$$

$$\hat{p}_{D|C} \;=\; \hat{P}(i \in D \,|\, i \in CO \cup D) \;=\; \frac{\sum_j \, I_{[j\in D]}}{\sum_j \, I_{[j\in CP]}}$$

where the sums over j are restricted to respondents in the survey with $I_{[j\in D]}$ not missing.

Step 2°. Next, the numbers of respondents n_L, n_C in the survey who respectively belong to the LL and CP frames are estimated using (BB5) as

$$\begin{aligned} n_L \;=\;& (\#\text{sampled by Landline}) \;+\; (\#\text{sampled by Cell in D}) \\ & +\, (\#\text{sampled by Cell with D missing}) \cdot \hat{p}_{D|C} \end{aligned} \tag{5}$$

$$\begin{aligned} n_C \;=\;& (\#\text{sampled by Cellphone}) \;+\; (\#\text{sampled by Landline in D}) \\ & +\, (\#\text{sampled by Landline with D missing}) \cdot \hat{p}_{D|L} \end{aligned} \tag{6}$$

Step 3°. We propose to impute the missing $\texttt{ADULTS}_i$ data among the survey respondents using only a constant weighted average from the non-missing $\texttt{ADULTS}_j$ values sampled by Landline or known to fall in the Dual-phone-status group D, as follows.

$$(1/\texttt{ADULTS})_{avg} = \Big(\text{sum of non-missing } (1/\texttt{ADULTS}_j) \text{ for j sampled by Landline or known in D}$$
$$+ \text{ (sum of non-missing } (1/\texttt{ADULTS}_j) \text{ for j sampled by Cell and missing D)} \cdot \hat{p}_{D|C}\Big) \Big/$$
$$\Big(\#\text{non-missing ADULTS sampled by Landline or known in D}$$
$$+ (\#\text{non-missing adults sampled by Cell and missing D)} \cdot \hat{p}_{D|C}\Big)$$

All missing values $1/\texttt{ADULTS}_j$ in the collected dataset are replaced by $(1/\texttt{ADULTS})_{avg}$.

Step 4°. Because of the complexity of the formulas, we provide the modified formula for weights w_i^o (before renormalizing as in (4) to obtain $\bar{w}_i^o$ averaging 1) ignoring the small third denominator term in (4) corresponding to the possibility of being sampled in both frames in the same dual-frame survey. Let the quantities π_i^L and π^C again be given respectively by formulas (2) and (3) after modifying the definition of n_L, n_C by (5) and (6).

For respondents i with non-missing D, we apply the formulas (2) and (3) and define

$$1 \,/\, w_i^o \;=\; I_{[i \in LO \cup D]} \cdot \pi_i^L \;+\; I_{[i \in CO \cup D]} \cdot \pi^C$$

For respondents i sampled by Landline with missing D, define

$$1 \,/\, w_i^o \;=\; \pi_i^L \;+\; \hat{p}_{D|L}\, \pi^C$$

For respondents i sampled by Cellphone with missing D, define

$$1 \,/\, w_i^o \;=\; \hat{p}_{D|C}\, \pi_i^L \;+\; \pi^C$$

Step 5°. For all respondents, $\bar{w}_i^o \;=\; n\, w_i^o / \sum_j w_j^o$. As in Buskirk and Best (2012), base-weights $\bar{w}_i^o$ summing to n are later scaled via calibration to sum to N.

Remark 1 *Step 3° uses only non-missing Landline* $\texttt{ADULTS}_i$ *values; only those values for sampled individuals in the Landline frame are used in constructing inclusion probabilities.*

Remark 2 *The form of Step 4° in the Base-weights recipe given here is designed to conform to the average of multiply imputed base-weights if the dual-telephone-status variable D_i were imputed when missing from survey respondents known to have Landline phones by Bernoulli($\hat{p}_{D|L}$) trials and from those known to have Cellphones by Bernoulli($\hat{p}_{D|L}$) trials.*

Remark 3 *Note that in Step 5° and also in (4) if the third denominator term is ignored, the telephone population-size N_{pop} plays no role in $\bar{w}_i^o$ because it is a constant factor in all w_i^o and cancels out of $\bar{w}_i^o$.*

4 Treatment of Missing Items in Poststratification

A review of documentation for raking and calibration software indicates that available weight-adjustment methods are generally designed to work without missing items among the (usually demographic and categorical) variables used for calibration. The software and documentation reviewed so far include: the `R survey` package of T. Lumley; the `anesrake` `R` package of Josh Pasek, which refers to a 2009 ANES technical report by M. DeBell and J. Krosnick along with a 2010 Stanford report of J. Pasek and a full 2014 ANES report by Pasek, DeBell and Krosnick; the `WesVar` package with 2002 documentation given in a StatCanada symposium report of G. H. Choudhry and R. Valliant; an IBM document by Jon Peck (2011) on raking in SPSS; and textbook literature (Agresti 2013, and Bishop et al. 1975) on raking and Iterative Proportional Fitting in contingency tables. The documentation for survey software suggests either that missing items be imputed in preprocessing steps, or that raking be done on records for complete cases (survey respondents with no missing items in any variables used for calibration). In the present Section, we sketch algorithms for weight adjustment in which there are some missing items and different records may be missing different items, but where a simultaneous imputation of all missing data is not needed.

These software packages share a common framework. Let $(Y_i : i \in \mathcal{S})$ be attribute observations on the respondents in a survey of size n. Again ignore the distinction between sample and respondents, with the understanding that inclusion-probabilities π_i and weights $w_i = 1/\pi_i$ account both for probability of sampling and possibly incomplete unit-response. Each unit i is assumed to provide observed categorical $\mathbf{X}_i$ vector covariates $X_{i,k}$, with first entry identically 1 and with 'missing' or `NA` an allowed category for the other entries. Let D_k denote the non-`NA` levels of $X_{i,k}$, and assume that the population-control counts $c_{X,k,d} = \sum_{i \in \mathcal{U}} I_{[X_{i,k}=d]}$ over all distinct values $d \in D_k$ are known or fixed in advance from estimates (possibly from the current survey), along with the population total N. Then define $c_{X,k,+} = \sum_{d \in D_k} c_{X,k,d}$. If the population counts are fixed from external sources with no missing values, then $c_{X,k,+} = N$, but if the control counts are fixed from a survey with missing $X_{i,k}$ items, then the counts $c_{X,k,+}$ may differ from N and vary with k.

Throughout this discussion, in a current survey $(\mathbf{X}_i,\ i \in \mathcal{S})$ may have some missing entries. Let $\mathbf{R}_i$ denotes a binary vector of the same dimension p as X_i, with components $R_{i,k}$ defined as the indicators that $X_{i,k}$ is observed rather than missing. When control totals are fixed from the survey in terms of base weights w_i^o scaled to sum to the true population size N, the control counts are

$$c_{X,k,d} \ = \ \sum_{i \in \mathcal{S}} R_{i,k}\, w_i^o\, I_{[X_{i,k}=d]} \quad , \qquad c_{X,k,+} \ = \ \sum_{i \in \mathcal{S}} R_{i,k}\, w_i^o$$

The goal is to provide adjustments to the weights w_i^o depending only on $\mathbf{X}_i$ and $\mathbf{R}_i$, in order to apply the new weights to many interesting outcomes and benchmark variables Y_i.

A unifying notation, for categorical or continuous covariates used in generalized raking calibration, is to take $\mathbf{X}_i$ as a numeric vector of components $X_{i,k}$ with known population averages $\bar{X}_k = t_{X_k}/N = N^{-1}\sum_{i\in\mathcal{U}} X_{i,k}$, such that the vectors $(X_{i,k},\ i \in \mathcal{S})$ are linearly independent. This can be done by replacing categorical covariates by dummy indicator-variables $I_{[X_{ik}=d]}$ for all levels $d = 2, \ldots, |D_k|$ other than $d = 1$ which serves as a reference category, and then eliminating all dummy columns that are perfectly expressed by linear combinations of dummy columns from earlier covariates. Such a restriction to linearly independent numeric columns $(X_{i,k},\ i \in \mathcal{S})$ is maintained in Sections 4.3, 4.4, and beyond.

4.1 Raking when Items May be Missing

Personal communication with Josh Pasek (author of `anesrake`) confirmed that his version of raking is done using in each raking pass all non-missing records for the survey variable of that pass, and agrees with the `anesrake` software documentation. The relevant part of the `rakeonvar.default.R` function (which is called by `rakelist.R` called by `anesrake.R`) in the `anesrake` package is

```
if (lwo == (lwt + 1)) {
    mis <- sum(weightvec[weighton == (lwt + 1)])
    weightto <- c(weightto * ((sum(weightvec) - mis)/sum(weightto)),
        mis)
}
for (i in 1:lwo) {
    weightvec[weighton == i] <- weightvec[weighton == i] *
        (weightto[i]/sum(weightvec[weighton == i])
}
```

Assume a survey starts from an initial set of population-control totals $c_{X,k,d}$ as described in the opening paragraphs of Section 4 above. The `anesrake` code above implements a raking pass with formulas below, for the categorical variable $(X_{i,k},\ i \in \mathcal{S})$, where w_i^A and w_i^B respectively denote the updated weight-vector at the beginning and end of the pass.

- Redefine the control vector, adjusted in non-missing categories, and a missing-category control equal to the sum of the weights of the observations with a missing value.

$$c^A_{X,k,d} \;=\; \frac{c_{X,k,d}}{c_{X,k,+}} \sum_{i\in\mathcal{S}} R_{i,k}\, w_i^A, \quad d \in D_k; \qquad c^A_{X,k,\texttt{NA}} \;=\; \sum_{i\in\mathcal{S}} (1 - R_{i,k})\, w_i^A \tag{7}$$

- The new weights are calculated as the new adjusted target for the category times the i'th weight proportion of the total of weights for that category.

$$w_i^B \;=\; w_i^A \,\frac{c_{X,k,d_i}\,\sum_{i\in\mathcal{S}}\,R_{i,k}w_i^A}{c_{X,k,+}\,\sum_{j\in\mathcal{S}}\,w_j^A\,I_{[X_{jk}=d_i]}} \;=\; \frac{w_i^A\;c_{X,k,d_i}^A}{\sum_{j\in\mathcal{S}}\,w_j^A\,I_{[X_{jk}=d_i]}}\;\;,\quad i\in\mathcal{S},\;\;R_{i,k}=1 \tag{8}$$

For the special case of the missing (`NA`) category where $X_{ik}=$ `NA`,

$$w_i^B \;=\; w_i^A\,\frac{c_{X,k,\texttt{NA}}^A}{\sum_{j\in\mathcal{S}}\,w_j^A\,I_{[X_{j,k}=\texttt{NA}]}} \;=\; w_i^A \tag{9}$$

These equations are for a single raking pass for the k'th raking variable. Each raking variable goes separately through this procedure. This way of handling missing data within each raking iteration is exactly the same as that described next in equation (10), and also agrees approximately *but not algebraically* with the generalized raking formulation described in Section 4.4 below.

4.2 Unified Notation for `anesrake` Raking Formula

This subsection provides a formula for direct raking as in `anesrake` when there are missing items in unit-responders, in a consistent notation. This is not the algorithm used later in solving the system of equations (17) that extends the scope of weight adjustment using missing-item data to linear and generalized-raking calibration along the lines of Deville, Särndal and Sautory (1993).

The method of `anesrake` is to rescale weight in each raking pass among the units so that weight-proportions of the respondent $X_{i,k}$ values follow the known population proportions. Suppose that a pass is about to be done using totals for the k'th covariate entries ($X_{i,k},\; i\in\mathcal{S}$). Let $\mathcal{R}_k$ denote the set of indices $i\in\mathcal{S}$ for which $R_{i,k}=1$, i.e., for which the k'th entry is observed; D_k the set of distinct categorical data-values for $X_{i,k}$ across i; and $c_{X,k,d}$ the control total counts for variables $X_{i,k}=d$. Let w_i^A denote the weights just before this raking pass, and w_i^B the weights just after. We define $w_i^B=w_i^A$ for $i\notin\mathcal{R}_k$ and rescale the weights for $i\in\mathcal{R}_k$ by a common factor, leading to the raking formula for each fixed k, for all $d\in D_k$ and $i\in\mathcal{R}_k$ for which $X_{i,k}=d$:

$$w_i^B \;=\; \frac{w_i^A\;c_{X,k,d}}{\sum_{d\in D_k}\,c_{X,k,+}}\;\;\sum_{j\in\mathcal{R}_k}\,w_j^A\;\Big/\sum_{j\in\mathcal{R}_k}\,w_j^A\,I_{[X_{j,k}=d]} \tag{10}$$

In this step, $\sum_{j\in\mathcal{R}_k}w_j^B=\sum_{j\in\mathcal{R}_k}w_j^A$, although later steps raking on other marginals can modify this weighted total. Formula (10) describes a single raking-pass, for the k'th entry of the categorical survey variables $\mathbf{X}_i$. Passes cycle successively through all entries k for post-stratification variables, ideally until convergence.

4.3 Linear Calibration with Missing Items

In this and succeeding sections, the notation changes somewhat. From now on, the covariate column-vectors $(X_{i,k},; i \in \mathcal{S})$ are now *numeric* and linearly independent columns whose population totals are known, along with the overall population size N. To work within this framework, cateogrical variables are replaced by those of their dummy indicator-columns $I_{[X_{ik}=d]}$ that are not prefectly predicted by earlier columns.

We reformulate in this subsection the standard derivation of linear calibration, to accommodate missing items. The usual derivation minimizes $\sum_{i\in\mathcal{S}} (w_i - w_i^o)^2/w_i^o$ subject to $\sum_{i\in\mathcal{S}} w_i \underline{X}_i = t_X \equiv \sum_{i\in U} \underline{X}_i$, where i indexes survey respondents with unit-nonresponse-adjusted weights w_i, and $\underline{X}_i = (1, X_{i,2}, \ldots, X_{i,p+1})$ are vectors of linearly independent covariates. This formulation assumes no $\underline{X}_i$ entries are missing, so that the first entry of t_X is the population size $t_{X,1} = N$. When that is true, the same optimization problem is rewritten in the form

$$\min_{\underline{w}} \sum_{i\in\mathcal{S}} \frac{(w_i - w_i^o)^2}{2\, w_i^o} \quad \text{subject to} \quad \sum_{i\in\mathcal{S}} w_i \left(X_{i,k} \,-\, t_{X,k}/N\right) \,=\, 0 \quad \forall\, k = 1, \ldots, p+1 \tag{11}$$

Next let $\bar{X}_k = N^{-1} \sum_{i\in U} X_{i,k} = N^{-1} t_{X,k}$ for $k \geq 1$. In (11), the sum involving only weights would be unchanged by missing items, while for each k the last sum could be restricted to $i \in \mathcal{R}_k$ in case there were missing items. The resulting modified linear calibration problem becomes:

$$\min_{\underline{w}} \sum_{i=1}^{n} \frac{(w_i - w_i^o)^2}{2\, w_i^o} \qquad \text{subject to} \qquad \sum_{i\in\mathcal{S}} w_i = N, \quad \sum_{i\in\mathcal{R}_k} w_i \,(X_{i,k} - \bar{X}_k) \,=\, 0 \quad \forall\, k \geq 2 \tag{12}$$

Minimizing with Lagrange multipliers $-(\beta_1 + \sum_{k=2}^{p+1} \beta_k \bar{X}_k,\, \beta_2,\, \ldots,\, \beta_{p+1})$ for the constraints leads to equations

$$w_i/w_i^o \,=\, 1 + \beta_1 \,+\, \sum_{k=2}^{p+1} \beta_k \bar{X}_k \,+\, \sum_{k=2}^{p+1} I_{[i\in\mathcal{R}_k]}\, \beta_k \,(X_{i,k} - \bar{X}_k) \tag{13}$$

That is, in terms of the quantities w_i^o, $X_{i,k}$, and $\bar{X}_k$, one determines the column vector $\beta = (\beta_1, \ldots, \beta_{p+1})$ of Lagrange multipliers by using (13) to define w_i and then by solving the linear system of constraints in (12) for β. The solution is expressed neatly in terms of the matrix X^* with first column $\mathbf{1}$ (denoting a vector of n 1's) and defined for $k \geq 2$ by $X^*_{i,k} = \bar{X}_k I_{[i\notin\mathcal{R}_k]} + X_{i,k}\, I_{[i\in\mathcal{R}_k]}$, which can be described as the original design matrix X with entries in the k'th column *imputed* to the externally fixed population average $\bar{X}_k$ wherever the i'th unit's item k value is missing. Then (13) is written equivalently as

$$w_i \;=\; w_i^o \left\{1 + \beta_1 + \sum_{k=2}^{p+1} \beta_k \bar{X}_k + \sum_{k=2}^{p+1} \beta_k (X^*_{i,k} - \bar{X}_k)\right\} \;=\; w_i^o \,+\, w_i^o \left(X^* \beta\right)_i \tag{13$'$}$$

and the constraint in (11) or (12) is written equivalently as $X^{*tr}\,\underline{w} = t_X$. Letting $W^o = \mathrm{diag}(\underline{w}^o)$ denote the $n \times n$ diagonal matrix with initial weights w_i^o along the diagonal, we combine the form of w_i in (13′) and the constraint to write the equation determining the Lagrange multipliers and the linear-calibration adjusted weights as:

$$t_X - X^{*tr}\underline{w}^o \;=\; X^{*tr}W^oX^*\beta \quad , \qquad \frac{w_i}{w_i^o} \;=\; 1 \,+\, \Big(X^*\,(X^{*tr}W^oX^*)^{-1}(t_X - X^{*tr}\underline{w}^o)\Big)_i \quad (14)$$

This derivation shows that the linearly calibrated adjusted weights w_i arising when respondents may have missing item-data are precisely the same as the usual formula for g-weights derived from GREG models (Särndal et al. 1992, p. 232) when the design matrix X is replaced by the design matrix X^* with population-averages imputed for missing items.

4.4 Generalized Raking with Missing Items

As formulated in Deville and Särndal (1992) and Deville, Särndal and Sautory (1993), raking and other calibration extensions satisfies (11) with a different metric between w_i/w_i^o and 1 in the first summation, and we find a similar extension to missing items as in (12). According to Deville, Särndal and Sautory (1993), the optimization problem (11) is replaced by

$$\min_{\underline{w}} \sum_{i\in\mathcal{S}} w_i^o\, G(w_i/w_i^o) \qquad \text{subject to} \qquad \sum_{i\in\mathcal{S}} w_i\, X_{i,k} \;=\; t_{X,k} \qquad \forall\; k = 1,\ldots,p+1 \quad (15)$$

when no items are missing, where $G(x)$ is smooth and satisfies $G(1) = G'(1) = 0, G''(1) = 1$. The most common choices for G apart from $(x-1)^2/2$ embodied in (11) are the functions

$$G_{rak}(x) = x\log(x) - x + 1, \quad G_{logis}(x) = \frac{(1-L)(U-1)}{U-L}\left\{(x-L)\log(\frac{x-L}{1-L}) + (U-x)\log(\frac{U-x}{U-1})\right\}$$

respectively associated with raking and with a "logistic" form of calibration guaranted to yield weight-ratios w_i/w_i^o in a fixed interval (L, U) containing 1. Unlike linear calibration, the generalized-raking methods using G_{rak} or G_{logis} are guaranteed to result in positive weights.

The problem (15) can be generalized to allow missing item-data in a way exactly analogous to the way (12) generalized (11). The resulting optimization problem is

$$\min_{\underline{w}} \sum_{i\in\mathcal{S}} w_i^o\, G(w_i/w_i^o) \quad \text{subject to} \quad \sum_{i\in\mathcal{S}} w_i = N, \quad \sum_{i\in\mathcal{R}_k} w_i\,(X_{i,k} - \bar{X}_k) \;=\; 0 \quad \forall\; k \geq 2 \quad (16)$$

and the solution is easily seen to have the same form (with Lagrange multipliers parametrized by (13)) as the generalized-raking solution with the design matrix X replaced by the population-average imputed design matrix X^*. The equation that generalizes (13′) to determine the weights and Lagrange multiplier vector β in this setting is

$$G'(w_i/w_i^o) \;=\; \big(X^*\beta\big)_i \;\; \forall\; i \quad \Longrightarrow \quad t_X \;=\; X^{*tr}\,\underline{w} \;=\; X^{*tr}\left(w_i^o \cdot (G')^{-1}\big((X^*\beta)_i\big)\right)_{i\in\mathcal{S}} \quad (17)$$

This system of equations reduces precisely to (14) when $G(x) = (x-1)^2/2$.

When $G \equiv G_{rak}$, $G'(x) = \log(x)$ and the second part of equation (17) says for all k,

$$t_{X,k} \;=\; \sum_{i \in \mathcal{S}} X^*_{i,k}\, w^o_i\, \frac{w_i}{w^o_i} \quad \forall\, i \,\in\, \mathcal{R}_k \qquad \Longrightarrow \qquad \sum_{i \in \mathcal{R}_k} w_i\,(X_{i,k} - \bar{X}_k) \;=\; 0$$

which is precisely the balance equation imposed by scaling the w_i weights for $i \in \mathcal{R}_k$ in a raking pass (10) on the categorical-variable calibration incorporating the k'th column of X. For each ordering of covariates, the general theory of convergence of raking implies that the solution obtained by successive single-variable raking passes is still unique in the presence of missing data. However, despite sharing the same constraint equations, the Iterative Proportional Fitting (IPF) style raking done in (10) does *not* in general lead to the same final adjustments as the optimized calibration-style raking in (17).

Remark 4 *(Entropy-balancing vs. Raking) Generalized-raking adjustments obtained by solving equations (17) are usually applied with design matrices* $\mathbf{X}^*$ *consisting of dummy columns for a set of single or pairwise-interacting cateogrical variables. An apparently different method of weight-adjustment has recently been advanced in Census Bureau research by Rothbaum and Bee (2021), after having been used in propensity-weighting for Causal Inference since the publication of Hainmuller (2012). This method, called Entropy Balancing, involves the solution of calibration equations like (17) below, with a few characteristic differences. First, in Causal Inference one is often interested in calibrating based on known properties of the national distribution of a continuous variable, such as income. In that setting, Hainmuller (2012) and later investigators work with design matrices with columns containing low-order powers of these continuous outcome variables, with normalized national totals equated to the corresponding known low-order moments of the national distribution. A second apparent difference is that Hainmuller defined his balancing problem by minimizing the 'entropy' metric function* $G_{ent}(x) = x \log x$ *subject to constraints. However, in the Deville-Särndal minimization problem (16), the number* n *of respondents is fixed and known, as is the population total* $N = \sum_{i \in \mathcal{S}} w_i$*. This implies that, subject to* $\sum_{i \in \mathcal{S}} w_i = N$ *plus other constraints, minimizing (16) for* $G(x) = G_{rak}(x) = x \log x - x + 1$ *is exactly the same as minimizing (16) for* $G(x) = G_{ent}(x)$*. Therefore, the entropy-balancing idea of weight-adjustment is exactly the same as generalized raking with metric-function* $G = G_{rak}$*.* □

4.5 `ANESrake` versus Generalized Raking

The solution to the generalized-raking calibration equations (17) can be found by the function `calibrate` in Lumley's (1999) `R`-package `survey`. Using that package for linear calibration (with argument `calfun = "linear"`) leads directly to the weighted-least-squares GREG solution (14), with sample-dependent g-weights (Särndal et al. (1992, p. 232). In the raking

case, $G = G_{rak}$ in (17) and `calfun = "raking"` in the `calibrate` function. However, when there is missing data in the raking calibration, the unique weight-solution is not obtained by a cyclical sequence of single-pass updates as in (10) using single-variable balance equations. The distinction between the two methods of raking with missing item-data can be seen through the observation that the solution to (17) does not depend on the ordering of the raking variables indexed by k, while the `anesrake` solution does. The differences between the two solutions, as well as the dependence of the `anesrake` solution on ordering, is confirmed numerically among the data results in Section 6.

5 Model-based Nature of Raking under Large Movement of Weights

The well-known paper of Deville and Särndal (1992) has for decades been cited as large-sample theoretical support for survey-weighted estimation and variance estimation using weights adjusted for unit nonresponse by means of 'generalized raking', as described in Sections 4.4 and Appendix A. However, its theory assumes that this calibration is done to correct calibration-variable totals using base weights defined as correct inverse single-inclusion probabilities, without unit nonresponse or missing respondent-data. A hallmark of the Deville-Särndal theory is that for samples of large size n, the calibration-step moves each weight an amount of order $1/\sqrt{n}$. All these assumptions may be challenged, but the weakest one is to require that the population underlying the calibration totals and the population used to define the designed base-weights are the same. The failure of this assumption, either due to strong self-selection of respondents or because the calibration totals reflect a population very different frm the respondents, results in the commonly observed phenomenon that calibration moves the individual base weights by amounts that are not small.

The published base or final weights for respondents in a survey can seldom be interpreted as estimates of the inverse probabilities of selection-and-response for units in the target population, because it is not clear what information response depends on. In a design-based setting, where the finite population attributes are viewed as nonrandom, the sampled units are selected by a random mechanism fully known to the investigator, but response occurs by self-selection. The indicator of response might be random, but even for respondents often depends on characteristics not accessible to the investigator. A first step in understanding the movement of weights toward a correct representation of population is to define what that means in formal terms for large samples. The mathematical framework for this is to view the population and sample as stages in a triangular array of larger and larger populations, with relative frequencies of characteristics settling down to limits along these stages. This framework has been adopted by many theoreticians of survey sampling (Krewski and Rao 1981, Deville and Särndal 1992, Rubin-Bleuer and Kratina 2005, Fuller 2009) and is now

the established framework for design-based asymptotic theory. The Appendix of this report extends the design-based theory of Deville and Särndal (1992) in essential ways. First, (A.3) defines the large-sample limiting properties needed for a system of population weights $\{w_i^*\}_{i\in\mathcal{U}}$ to be asymptotically correct with respect to attributes $\mathbf{X}_i$ and outcomes Y_i. This definition is novel in being fully design-based, but it is not the only possible definition: in particular, this one restricts sample-weighted sums of outcomes Y_i in the presence of covariates $\mathbf{X}_i$ in a way compatible with the *missing at random* condition of Little and Rubin (2019). In this setting, one seeks large sample consistency and asymptotic normality and variance estimation of survey-weighted estimates of Y-totals using weights adjusted by generalized raking. The second extension of calibration theory given in Appendix B (Theorems 1–2) is to establish these large-sample results (for variances, in a Poisson-sampling setting, under a natural additional population-level assumption (A.6)) when correct weights satisfy a parametric regression model (A.4) in terms of the post-stratification variables $\mathbf{X}_i$. This extension of Deville and Särndal's 1992 results is accomplished in their own design-based terms, although similar results are known in essentially model-based theoretical treatments of parametric or semiparametric response-propensity models.

The operational conclusion of the new theory in the Appendix is that, when the parametric model (A.3) holds and there are no missing covariate entries, the estimated totals $\hat{t}_Y$ after (generalized) raking are consistent and asymptotically normal, with variances that can be consistently estimated by exactly the same estimator that would be used after linear calibration starting from the calibrated weights $\hat{w}_i$. That is, after calculating $\hat{w}_i$ using software such as the `calibrate` function in the `survey` package in `R` (Lumley 2010), the variances can also be estimated using these estimated weights in the role of base-weights; re-running the *linear* calibration on the same data with the same covariates and design; and substituting the GREG residuals into a standard design-based variance estimation formula such as (35) in Theorem 2 for the Poisson sampling setting appropriate for RDD and Web surveys.

A further technical innovation in the Appendix is the extension of the theoretical results to the case where the parametric model (A.4) relating correct weights w_i^* to base-weights w_i^o is misspecified. This is unfortunately the most common case. Theorems 3 and 4 provide similar information to that obtained by robustified variances in misspecified model M-estimation for *iid* data. If asymptotically correct weights can be estimated by some means, such as a more fully parametrerized model enabling consistent but not necessarily asymptotically normal estimation (for example, consistency without any bound on the rate of convergence), then Theorem 4 shows how to estimate the variance of survey-weighted estimates following weight-adjustment by raking. This could be useful in statistical checks on the goodness of fit of a specific calibration scheme in estimating benchmark totals.

5.1 Variance Estimation after Raking

A preliminary, 'naive' estimator of variance[2] of a survey-weighted total $\hat{t}_{w,Y} = \sum_{i \in \mathcal{S}} w_i Y_i$ can be obtained by ignoring the post-stratification of the weights w_i and treating the Y_i's as though they are random and *iid* over $\mathcal{U}$ with variance σ_Y^2. The formula is obtained by modifying the PPS with-replacement variance formula in Cochran (1977, p. 254) in Hájek ratio form and separating the estimator of superpopulation y-attribute variance from the weights, is

$$\hat{V}_{naive}(\hat{t}_{w,Y}) \;=\; (\hat{\sigma}_Y^2\, N^2) \sum_{i \in \mathcal{S}} w_i^2 \,/\big(\sum_{i \in S} w_i\big)^2 \tag{18}$$

The same variance formula, with w_i^2 replaced by $w_i(w_i - 1)$, is asymptotically correct for the scaled-weight or Hájek estimator $\hat{t}_{w,Y} \cdot N / \sum_{i \in \mathcal{S}} w_i$ in large surveys if the Y_i attribute values are constants and units $i \in \mathcal{U}$ are Poisson- (i.e., independently) sampled with inclusion probabilities $\pi_i = 1/w_i$, as $N = |\mathcal{U}|$ and the expected sample size $n = \sum_{\imath \in \mathcal{U}} \pi_i$ tend to ∞. However, the estimator (18) errs by ignoring the weight changes due to calibration or raking that constrain weighted totals $\sum_{i \in \mathcal{S}} w_i \mathbf{X}_i$ close to 0 for the post-stratifying variables $\mathbf{X}_i$. Thus, (18) is expected to overstate the variance for attributes Y_i with significant regression on the poststratifying variables.

Remark 5 *In remotely conducted surveys like the RDD and Web arms of the Tracking Survey, with no clustering in their design, the sampling design including response may not be far from Poisson (independent across sampled units), although the inclusion probabilities depend on many unknown factors. With this in mind, we develop variance formulas as though the base-weighted sample follows an unequally-weighted Poisson design.* □

Because of raking weight-adjustment, reported margins of error (MOE) in survey demographic totals for marginal post-stratifying variables may be extremely small. This holds in the American Community Survey (ACS) , for example. After its multiple raking stages, very small MOEs are reported for County level demographics in sex and coarse age- and race-categories (some partially cross-classified!) represented in the Census Bureau's Population Estimates which serve as targets for raking-calibration. It is important, therefore, to account correctly for the effect of raking weight adjustment on the variances of survey totals.

Consider the estimation of variance after raking for large surveys with minimal or no missing item data. Formula (18) for variance is too simplistic. At the next higher level of sophistication, if weights are calibrated by generalized raking to true totals based on true inverse-inclusion-probability design weights $w_i^o = 1/\pi_i^o$, the theory of Deville and Särndal (1992), which is re-proved in Theorems 1 and 2, establishes that the variances of survey totals $\hat{t}_{w,Y}$

[2]In the context of the Tracking Survey, Paul Biemer suggested this estimator of variance as a rough comparator for the GREG-weighted and model-based variances discussed later in this Section.

are close in large samples to the variances of GREG residuals of Y_i regressed on the post-stratifying variables. The formula is (18) with Y_i replaced by the GREG residuals from regression on the full set of post-stratifying variables $\mathbf{X}_i$:

$$\hat{V}_{GREG}(\hat{t}_{w,y}) \;=\; \frac{N^2}{n}\Big[\sum_{i\in\mathcal{S}} w_i^o\,(Y_i-\hat{\beta}'\mathbf{X}_i)^2\,\Big/\sum_{i\in\mathcal{S}} w_i^o\Big]\;\cdot\;\sum_{i\in\mathcal{S}}\; w_i^o\cdot(w_i^o-1)\,\Big/\Big(\sum_{i\in\mathcal{S}} w_i^o\Big)^2 \qquad (19)$$

where $\hat{\beta}$ are the GREG regression coefficients for Y_i in terms of $\mathbf{X}_i$ based on w_i^o weighted regression. Finally, if the parametric model (A.4) for correct weights in terms of base weights holds, then the theory of this paper (Theorem 2 and formula (35) in Appendix B) justifies replacing (19) by the same formula with w_i^o replaced by $\hat{w}_i$ and with $\hat{\beta}$ denoting coefficients obtained from $\hat{w}_i$-weighted least squares regression.

6 Data Results from the Tracking Survey

This section provides an extended description, from preprocessing to descriptive exhibits to final results, of the weight adjustments developed for both the RDD and Web components of the Tracking Survey introduced in Section 1. The general reference for the data collected is the Team Y&R Project (2020). Other papers and reports discussing data quality, comparisons between the RDD and Web data and results, and interpretations of those comparisons, can be found in Ellis et al. (2022a,b).

6.1 Pre-processing, Outcome Variables and Data Recoding

The choice of more than 200 measured survey variables was made in the Census Bureau's agreement with the contractor, Team Y&R (2020), which itself sub-contracted with `ReconMR`, a commercial data-collector. Data were collected from a single adult in each respondent household, chosen as the person aged at least 18 with the nearest birthday. Survey variables included personal and household demographics, geographic location, telephone type (landline or cell) and availability, and benchmark variables chosen for comparability with data collected at national level in high-quality government surveys. Data were collected on a monthly basis. In the RDD survey, approximately 1400 respondents were interviewed per month from September through December 2019, and a total of 36675 respondents supplied data from January through June 2020. The Web survey was designed to be 50% larger, 2100 respondents per month in 2019 and 54,000 in all of 2020. In both surveys, the actual data collection was self-contained and self-terminating in each month of 2019, but spilled over across months during 2020. This aspect of the data collection was highlighted by Team Y&R, which did data-imputations, base-weighting (for the RDD survey) and raking weight-adjustments separately for each month of 2019 and in 2020 re-computed base-weights daily

and weight-adjustments weekly, separately for RDD and Web (Team Y&R 2019, p. 10). Because of the separate data collections monthly in 2019 and in all of 2020, the Census Bureau Tracking Survey analysis team weighted and analyzed the data separately for each 2019 month and for the 2020 months lumped together.

The contractor chose a standard set of variables, `household size`, number of `Adults` in household, and `telephone type` (landline or cell or both), for use in base-weighting the RDD survey, and the Census Bureau analysts used these same variables. (Base-weights for all respondents were equal by definition in the Web Survey.) The contractor also chose 8 variables for post-stratification of both surveys: 6-group `Age` cross-classified by `Sex`, 3-level `Education` by `Sex`, 4-level `Census Region`, 3-level `Education` by 5-group `Age`, 2-level `Education` by (indicator of) `White Alone non-Hispanic`, `Owner/Renter`, and (quartile of) `Population Density`. The contractor used these variables for poststratifying the Web survey, and in the RDD survey these same variables together with `TelStatus` (a categorical variable telling whether the respondent used landline or cell telephone or had both telephone types available). The Census Bureau analytical team used the same variables as the contractor (for both surveys) together with the additional variable `Adults` telling whether there were 1,2,3 or more than 3 adults in the respondent household. These post-stratifying variables were targeted to 2018 national proportions from the ACS 5-year 2014-2018 data (which themselves were calibrated to the 2018 Population Estimates published by the Census Bureau). Missing values for the base-weighting and post-stratification variables were imputed in the Y& R analyses, for purposes of weight-adjustment only, by a randomized hot-deck imputation method using (a subset of) the same set of post-stratifying variables.

The outcome variables of interest in the Tracking Survey measured its respondents' attitudes toward the Census Bureau and decennial census and (in later 2020 months) self-response to the census. Results for these outcome variables will be reported elsewhere (Ellis et al. 2022a,b). The analytical team was also tasked with assessing and comparing the quality of data from both the RDD and Web surveys. The methodology for this assessment, following that of Yeager et al. (2011) and MacInnes et al. (2018), consisted of comparing the survey-weighted estimates from the two surveys with each other and national target proportions, on the primary demographic variables used in post-stratification, on secondary demographic variables (such as `Marital Status`, `non-English` language spoken in the home, and on a few benchmark variables (indicators of `Volunteering`, `Giving Blood`, and self-reported `Health` levels, and an indicator of Work for Pay). The national target proportions for the demographic and Work-for-Pay variables were derived from ACS, with other targets respectively derived from the Current Population Survey (for volunteering), NHANES (for donating blood and activity) and the National Health Interview Survey (for health). There was one further benchmark question included on the Tracking Survey, a question drawn from NHANES about physical activity. However, because of different ordering and skip-patterns of the `Activity` question in the Tracking survey versus the NHANES survey, that question turned out not to be usable as a benchmark and is therefore not discussed further.

The data collected from the Tracking Survey respondents had many missing items (coded as 'unknown'), but our descriptive summaries in this report are limited to the post-stratification and benchmark variables used to compare the effectiveness of weight adjustment. Across the whole Telephone survey, with 42334 respondents, the counts of missing items were:

AGE	HISP	REG	EDUC	RACE	TELSTAT	ADULTS	CPS_VOL	BLOOD
1693	1080	1903	547	1978	524	972	386	265

`SEX` and `RENT` responses were never missing. In addition, the self-reported Health variable from a survey question similar to that asked in NHANES, showed 57 missing values, but the question was administered in the Telephone survey only in the 2019 months (which had a total of 5649 respondents). For each of these variables, the missing items appeared at approximately the same rate (as a proportion of all respondents) across the 2019 months and over all of January to June 2020.

In the Web Tracking survey, the data collectors must have imposed as a requirement for acceptable responses that the poststratifying demographic variables not be missing or unknown. The `Age, Hisp, Reg, Educ` and `Race` responses were never missing or 'unknown' in the Web survey, and the counts of household `Adults` (which we use in poststratifying but the Team Y& R analysts did not) were missing for just 302 respondents out of a total of 62494. The rate of missing benchmark responses was not markedly smaller in the Web survey than in the RDD survey, with 462 missing for `CPS_VOL` and 265 for the NHANES-styled question about donating `Blood`.

We tabulated the extent to which multiple demographic variables were simultaneously missing in the Telephone survey. For the seven variables `Age, Hisp, Reg, Race, Educ, TelStat`, and `Adults`, Table 1 shows the proportions of respondents in each month of data (monts 9-12 were in 2019, 1-6 were in 2020) with numbers of missing variables 0 to 7. The Table shows that the proportions in the categories 0:7 were remarkably stable across months, regardless of the different month numbers sampled (approximately 6000 per month in 2020 months 1-6, 1400 per month in 2019 months 9-12). The proportion of respondents with at least one missing demographic variable (equal to 1 minus the entry in the first row of Table 1) ranged from 0.124 to 0.151.

Some re-coding of basic demographic variables was done in the Tracking surveys, separately for the respondents in the RDD and Web samples, to reduce detailed ordinal variables (such as `Age, Household Size, Educ` (educational level) to categorical variables with a small number of levels. Thus, ages in the range 18 to 97+ were converted to `AgeGp5` intervals 18-24, 25-34, 35-44, 45-64, 65-97; 12 Educational levels were reduced to 3: High school or less, some college, and BA+; and household sizes ranging from 1 to 8 were truncated at 4. The number of adults in the household, which was used in base-weighting the RDD survey, was also capped at 4. While missing `HH_Size` and missing `Adults` could each occur without the other, by definition `Adults` was taken to be 1 when `HH_Size` was 1. As a result

Table 1: Monthly proportions of RDD survey respondents with numbers 0-7 of missing demographic variables (out of 7). Column for each month sums to 1.

	Months									
missing	1	2	3	4	5	6	9	10	11	12
0	0.864	0.858	0.876	0.863	0.875	0.869	0.876	0.876	0.863	0.849
1	0.094	0.104	0.094	0.096	0.093	0.089	0.082	0.092	0.091	0.107
2	0.024	0.022	0.017	0.021	0.019	0.024	0.031	0.018	0.026	0.025
3	0.008	0.008	0.006	0.008	0.006	0.008	0.005	0.006	0.014	0.006
4	0.004	0.003	0.002	0.004	0.002	0.002	0.001	0.005	0.004	0.003
5	0.003	0.003	0.002	0.002	0.001	0.002	0.003	0.001	0.002	0.005
6	0.002	0.002	0.002	0.005	0.003	0.004	0.001	0.001	0.001	0.004
7	0.001	0.000	0.000	0.001	0.000	0.001	0.001	0.000	0.000	0.001

of the re-coding summarized in this paragraph, all survey variables were categorical, and the generalized-raking operations were done (separately for RDD and Web data) on design matrices $\mathbf{X} = \left(X_{i,k}\right)$ consisting of an initial column of 1's followed by the dummy columns (one fewer than the number of levels) for the post-stratification variables.

The generalized-raking weight adjustment in (17) required further recoding. Item nonresponse was initially coded in the raw data as 'unknown' category levels, usually $98, 99$. The unknowns were first changed to 'missing' (`NA`); then the categorical values were written out as multiple dummy-variable columns (the columns that will later become $(X_{i,k},\, i \in \mathcal{R})$ of the design matrices with rows $\mathbf{X}_i$), all dummy-entries of which were stored as `NA` whenever the corresponding categorical variable (from which dummy-column k was derived) was missing. Finally, all the missing elements $X_{i,k}$ were replaced by respondent averages

$$\bar{X}_k = \sum_{j \in \mathcal{R}} I_{[X_{j,k} \neq \texttt{NA}]}\, w_i^o\, X_{j,k} \,/\, \sum_{j \in \mathcal{R}} I_{[X_{j,k} \neq \texttt{NA}]}\, w_i^o$$

where w_i^o are Base-weights in the RDD and Uniform (w_i^o all equal) in the Web survey.

6.2 Base-weighting & Poststratification in RDD & Web Surveys

The steps for base-weighting and poststratification weight adjustment in the Tracking surveys were described in Sections 3, 4, and 4.4. These are largely standard except for the treatment of missing poststratification variables in respondent data. The contractor Y&R used standard formulas for base- and poststratification-weighting (raking) after first filling in missing respondent data by a randomized hotdeck procedure using the `R` package `hotdeck`. As explained further below, we (the Census Bureau analysis team) rejected that approach for three reasons. First, at least in the smaller monthly samples, randomized hot-deck im-

putation injects a marked degree of randomness into the base-weights and final weights. Second, the use of complete-data formulas following imputation makes every formula for base-weighting and raking depend on a joint imputation model filling in all data, while our method of Sections 4–5 depends only on filling in the variable needed for one balance equation index k at a time in (17). A third reason is that our single-variable imputations could be done by mass-imputation based on known national categorical proportions, while the joint imputation models (of `hotdeck` and other available packages like `MICE`) are generated only from observed data and are noisy and unvalidated. Model-based imputation methods might be better if respondent samples are large and can take account of interactions expressing local geographic patterns of missing data, but even then a routine approach to model-development is likely to lead to noisy and unreliable model specification.

For the RDD survey, we implemented the dual-frame approach of Buskirk and Best (2012) in the case of missing `Adults` or `TelStatus` data in Steps $1^\circ - 5^\circ$ of Section 3. Team Y&R had filled in these missing data by randomized hot-deck imputation before applying the complete-data base-weighting formulas of Buskirk and Best (2012). In the Web Survey, which involved self-selection in response to a general unweighted list of invitations to supply data, there was no probability design and thus no base-weighting, so that w_i^o are all chosen equal to the reciprocal of the number of respondents in each *survey time-block* (that is, in each month of 2019, and in the set of all months of 2020 lumped together).

After estimating base-weights in the RDD survey for each survey time-block, we compute in the RDD survey the post-stratified adjusted weights with respect to the target ACS variables, ten variables `AGE6xSex`, `EDU3xSEX`, `REG`, `ED3xAGE5`, `EDU2xWNH`, `AGE2xWNH`, `RENT`, `POPDENSITY`, `TELSTAT`, `ADULTS` for the RDD survey, and the same set without `TELSTAT` was used to post-stratify the Web Survey. We poststratified a few different ways within the generalized raking framework of Sections 4.3 and 4.4, using $G = G_{lin}, G_{rak}$, and another 'logistic' choice that enforces pre-chosen bounds on weights. All these produced similar adjusted weights. Because a few adjusted weights with linear calibration (G_{lin}) were negative, and those done by raking (G_{rak}) were nicely bounded without any further intervention (such as weight-trimming) or modification of G), we present only the raking results in exhibits, and these are the weights we used in developing survey estimates and their standard errors.

Note that all base-weights and adjusted weights in the descriptions that follow are normalized to have average 1 within each survey time-block within both the RDD and Web Surveys. Table 2 displays summary statistics for the distribution of RDD Base-weights across the five Tracking Survey time-blocks. These weights draw meaningful differences between the probabilities of selecting different households, but with a factor of no more than about 15 from smallest to largest in each time-block, the difference is not concerningly large. The spread is roughly as large as for the base-weights constructed by Team Y&R, using randomized hot-deck imputation to correct for missing `Adult` and `TelStat`.

The distribution of these base-weights across different survey time-blocks is quite stable.

Table 2: Base Weights in the RDD Tracking Survey within each survey time-block, normalized to average 1, summarized through overall range and interquartile range.

	Sep2019	Oct2019	Nov2019	Dec2019	Yr2020
Min	0.558	0.558	0.561	0.561	0.564
1st Qu	0.763	0.764	0.765	0.764	0.766
Median	1.204	1.104	1.200	1.197	1.191
3rd Qu	1.204	1.211	1.200	1.197	1.191
Max	8.316	8.283	5.275	5.279	8.583

The apparent anomaly of some maximum base-weights being particularly small is not so concerning when we note that the respective numbers of survey base-weights greater than 5 in the five survey time-blocks are respectively $2, 1, 1, 1, 8$ and that the number of respondents in the `Yr2020` time-block is roughly 26 times the number in each of the 2019 months.

One way to assess the spread of base-weights and the movement from base-weights to adjusted weights is through the subject-level ratios of those weights. These are shown by survey time-block in Table 3, both for the RDD and the Web surveys. The distributions look rather stable across survey time-block and even across RDD and Web. In this Table, the slightly concerning feature may be the very small size of the smallest weight-ratios in the first three months of the RDD survey. There may be a real difference between the smallest weights obtained by raking in those months as compared with `Dec2019` and `Yr2020`, but it is not large, because the numbers of respondents with adjusted weights < 0.06 are respectively $5, 5, 6, 0, 0$ in the RDD survey and $0, 3, 0, 0, 0$ in the Web survey. Another similar table (not shown here) of the distributions across survey time-blocks of the adjusted weights in the RDD survey (without dividing by base-weights) actually has less spead than the upper portion of Table 3, with respective maximum values $6.319, 7.239, 6.042, 5.490, 7.690$.

The ratio of maximum to minimum adjusted weights within each survey time-block are definitely larger for our adjusted weights raked as in Section 4.4 than for the Team Y&R weights. The differences between these sets of adjusted weights is illustrated for the Nov. 2019 time-block of the RDD Survey in the scatterplot of Figure 1. In this month as in the other time-blocks, for both surveys, the adjusted weights computed by Y&R are on one hand more scattered than ours, due to the hot-deck imputations that preceded the Y&R raking steps. On the other hand, the Y&R final weights are then compressed by weight-trimming both above and below. The Y&R weights for the five time-blocks in the RDD survey are trimmed below with lower limit in a range $[0.18, 0.23]$ for 2019 months and $[0.12, 0.15]$ for 2020 months, and are trimmed above with upper limit in a range $[3.6, 4.7]$ in 2019 months and $[4.2, 4.9]$ for 2020 months.

In any case, it is completely clear from Table 3 that very substantial weight adjustments are required to calibrate the Telephone and Web Tracking Surveys to the target proportions

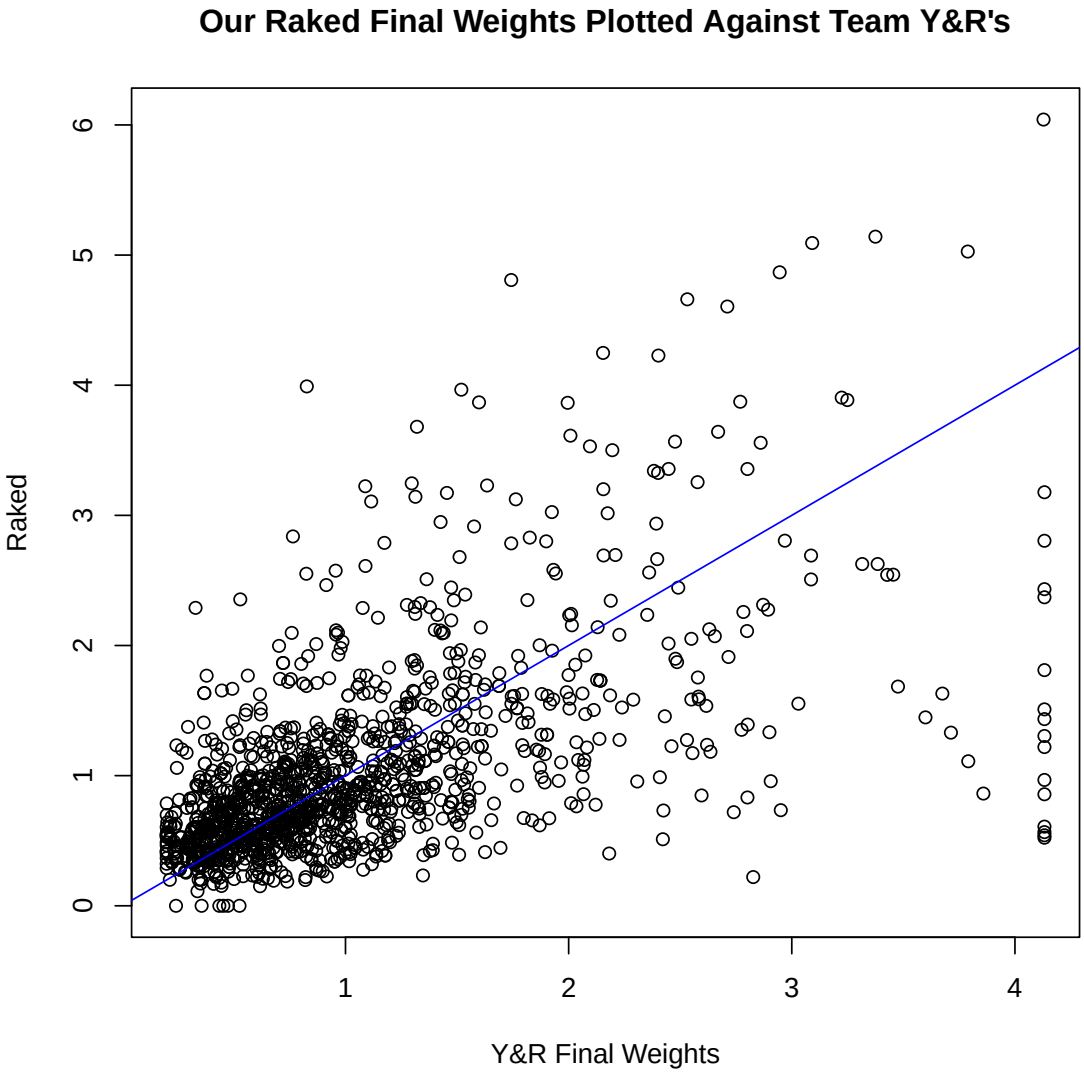

Figure 1: Scatterplot of adjusted weights for subjects in Nov. 2019 RDD Tracking Survey, as computed by the methods of this paper (mean imputation, no weight-trimming) versus those of Team Y&R. Note that the horizontally placed weights have been trimmed both on the right and left.

Table 3: Ratios of Adjusted over Base Weights in the RDD and Web Tracking Surveys within each survey time-block, summarized through overall range and interquartile range.

RDD	Sep2019	Oct2019	Nov2019	Dec2019	Yr2020
Min	0.000	0.000	0.000	0.089	0.075
1st Qu	0.566	0.517	0.570	0.511	0.581
3rd Qu	1.291	1.260	1.287	1.320	1.333
Max	5.247	8.264	7.109	9.754	9.188
Web	Sep2019	Oct2019	Nov2019	Dec2019	Yr2020
Min	0.100	0.042	0.148	0.158	0.089
1st Qu	0.590	0.626	0.583	0.602	0.593
3rd Qu	1.237	1.154	1.230	1.236	1.244
Max	5.293	8.568	6.442	8.071	6.357

used here for 9 or 10 ACS survey variables. Another way to confirm this is through the descriptive statistics of average changes in respondent weights contained in Table 4. This is only to be expected due to the self-selection of respondents in very low-response surveys, whether those surveys have a probability sampling design like the Tracking RDD Survey or reflect acceptance of an invitation to participate in a Web Survey where the composition of the invited population is unknown. Therefore the classic theory of Deville and Särndal (1992) for large-sample distributional behavior of post-calibration survey-weighted estimators does not apply, and the newer theoretical developments of Section 5 are needed.

Table 4: Average absolute weight changes Δ by respondent, and fraction of respondents with small weight changes (approximately 1/2 the mean), from Base to Final adjusted weights, in RDD and Web Tracking Survey within each survey time-block. Both Base and Final weights within each survey are normalized to have average value 1.

	Survey	Sep2019	Oct2019	Nov2019	Dec2019	Yr2020
mean(abs(Δ))	RDD	0.499	0.573	0.502	0.548	0.482
	Web	0.427	0.436	0.441	0.427	0.430
frac($\lvert\Delta\rvert < 0.2$)	RDD	0.268	0.240	0.273	0.229	0.271
	Web	0.270	0.320	0.259	0.288	0.295

Remark 6 *For linear calibration as in Section 4.3, based on nonsingular matrix inversion, redundant columns are removed from the design matrix. The need for this arises because several of the paststratification variables cross-classify pairs of variables with common underlying categorical variables (specifically* `AGE, EDUC3, SEX` *and indicator of* `White-alone`

non-Hispanic*). Thus, even in the presence of missing data certain columns are redundant. Removing redundant columns also makes sense for generalized raking: after this step, all the quasi-Newton numerical maximization steps can be based on matrix inversion rather than generalized inverses or methods based on more careful singular value decomposition.* □

Remark 7 *Section 4.5 mentioned the verification for the Tracking telephone survey data that the* anesrake *IPF-style raking with update steps (10) differs algebraically from calibration-style raking (17). We did this in two ways: by direct comparison of IPF-style and calibration-style raked weights, and also by comparing the IPF-style raked weights calculated with the 3 categorical marginal variables (*AGE×SEX, Race, NHWA*) entered in their 6 possible orderings. (Here* NHWA *is the indicator variable for Non-Hispanic persons self-identifying their race as White alone.) The results were that, across the 1412 respondent records, the mean absolute difference between the IPF-style raked weights (with variables entered in any order) and the calibration-style raked weights were approximately* 0.03*, while the mean absolute differences between the IPF-style raked weights computed via (10) with variables entered in different orders ranged from* 0.001 *to* 0.003*. These calculations show that the ordering of variables in the successive raking passes of (10) cannot be completely ignored but may not be important for most purposes unless there is a high rate of item missingness. There were very few entries (18 total) in which the* anesrake *weights based on differently ordered marginal totals were in fact different, and all of these occurred in units with at least one missing item.* □

6.3 Metrics of Successful Weight-Adjustment

A preliminary assessment of the quality of the adjusted survey data from the two Tracking Surveys is based on the level of agreement achieved after poststratification between survey-weighted estimates of population proportions of demographic and benchmark outcomes with their targets obtained from high-quality national surveys. The estimated proportions are compared with the targets, with the discrepancies standardized by estimated standard errors of the survey estimates. This was the methodology used by Yeager et al. (2011) and MacInnes et al. (2018) in their comparative studies of the quality of RDD probability surveys and of Web surveys. We follow the same basic methodology here. A fuller account of the same methodology applied to more questions from the Tracking Surveys can be found in Ellis et al. (2022a,b). In this Section, we restrict attention to the following subset of 7 demographic and benchmark outcome proportions: Age 18-34, Age 45-64, WA.Hsp, BA, HH1:2Rent, HH3+Own, CPS_Vol.Y. (Here WA denotes White Alone, BA Black Alone, Hsp Hispanic, and HH1:2Rent denotes Renter households of size up to 2, while HH3+Own denotes Owner households of size 3 or more.) The target values for these 7 outcome proportions, drawn from the 2019 Population Estimates or ACS, are as follows:

Age18:34	Age45:64	WA.Hsp	BA	HH1:2RENT	HH3+OWN	VOL.Y
0.298	0.327	0.146	0.130	0.236	0.251	0.278

Of these outcomes, the age-categories were poststratification variables, the `WA.Hsp` and `HHsize x Rent` categories are pairwise interactions of categories exactly or roughly used in poststratification (since number of Adults and not HH-size was the poststratification variable used), and and BA and volunteering activity as measured in CPS were not used in poststratification. All these outcome variables have national target proportions available from 2019 single-year ACS figures. (Other simple pairwise interaction variables, such as `HHsize x SEX`, are apparently not available from published ACS tables although they could be estimated from ACS public-use microdata.)

Table 5 displays the survey-weighted estimates of the 7 outcome variables at the unweighted, base-weighted, and poststratified stages of weighting, for the RDD and Web surveys, by survey time-block. Then Table 6 shows, for each outcome variable and survey time-block in the RDD and Web surveys, the final weighted survey estimate after raking together with its estimated standard errors (SE), the discrepancies between the final weighted estimate and its target, and finally the ratio of this discrepancy over the SE. This last standardized ratio is like a Wald statistic to determine whether the discrepancy is large: the p-value associated with this standardized ratio referred to a standard normal distribution would be the p-value for a hypothesis test that the discrepancy is large. In Table 6, the threshold absolute value for this standardized ratio is chosen as 3 because so many different estimated outcomes are being scrutinized.

The upshot of the comparisons in Table 5 is first of all that all weighted estimates are extremely stable across survey time-blocks, that the outcome estimates made unweighted or with base weights are hardly different (although we saw in Table 2 that there is considerable variability in the Base-weights), that the outcome variables involving poststratification variables are better estimated by the Raked weights than the uniform (all equal) or the Base weights, and that our Raked weights (computed without weight-trimming and with single-variable mass imputation rather than hot-deck imputation) are generally closer to the targets than the Y&R adjusted weights. The Raked-weight estimates for the Age outcomes in the RDD survey are particularly close to the targets, with particularly small SEs shown in Table 6, because Age was a raking variable and it had very little missing data. For Age and a few other variables in Table 6 closely related to poststratification variables, SEs were very small because there was no missing data among the Web Survey's raking variables except for a few missing `Adults` responses.

It is striking how far some of the final raked estimates can be from their targets. For example the `WA.Hsp` population proportion estimated in the RDD survey is about half the target proportion, and the `Vol.Y` benchmark proportion is very poorly estimated in both the RDD and Web surveys. No doubt, these large discrepancies are ultimately due to the very different compositions of the RDD and Web unweighted respondents and the marked difference of those self-selected populations from the national population. The unweighted-population characteristics for the outcome-variables `Age18:34`, `WA.Hsp`, `HH1:2Rent` and `Vol.Y` are very

Table 5: Estimates of each of 7 outcome variable proportions for each survey time-block, at various stages of weighting for each of the two Tracking Surveys.

Time	Telephone				Target	Web			Variable
	Unwt	Base	Raked	YR.fnl		Unwt	Raked	YR.fnl	
Sep2019	0.196	0.229	0.278	0.282	0.298	0.303	0.290	0.301	**Age18:34**
Oct2019	0.170	0.193	0.275	0.291		0.303	0.290	0.303	
Nov2019	0.192	0.219	0.275	0.284		0.313	0.290	0.303	
Dec2019	0.192	0.216	0.277	0.286		0.292	0.290	0.300	
Yr2020	0.196	0.218	0.278	0.285		0.305	0.290	0.304	
Sep2019	0.354	0.349	0.318	0.330	0.327	0.341	0.331	0.342	**Age45:64**
Oct2019	0.340	0.342	0.314	0.330		0.341	0.331	0.341	
Nov2019	0.334	0.337	0.314	0.327		0.333	0.331	0.342	
Dec2019	0.338	0.340	0.316	0.326		0.337	0.331	0.342	
Yr2020	0.337	0.338	0.317	0.329		0.325	0.331	0.339	
Sep2019	0.052	0.055	0.072	0.060	0.146	0.100	0.083	0.083	**WA.Hsp**
Oct2019	0.044	0.049	0.079	0.062		0.086	0.075	0.075	
Nov2019	0.051	0.057	0.070	0.054		0.076	0.072	0.075	
Dec2019	0.057	0.063	0.086	0.079		0.083	0.075	0.075	
Yr2020	0.055	0.059	0.073	0.063		0.096	0.083	0.084	
Sep2019	0.077	0.075	0.112	0.085	0.130	0.131	0.123	0.120	**BA**
Oct2019	0.063	0.063	0.102	0.085		0.129	0.137	0.130	
Nov2019	0.081	0.082	0.116	0.081		0.131	0.146	0.134	
Dec2019	0.074	0.070	0.120	0.081		0.134	0.138	0.133	
Yr2020	0.093	0.092	0.120	0.091		0.109	0.117	0.105	
Sep2019	0.147	0.155	0.194	0.156	0.236	0.220	0.209	0.195	**HH1:2Rent**
Oct2019	0.143	0.151	0.202	0.162		0.202	0.217	0.207	
Nov2019	0.149	0.159	0.195	0.157		0.202	0.211	0.186	
Dec2019	0.148	0.153	0.212	0.165		0.199	0.206	0.183	
Yr2020	0.161	0.169	0.204	0.161		0.188	0.213	0.191	
Sep2019	0.301	0.317	0.276	0.346	0.251	0.267	0.259	0.287	**HH3+Own**
Oct2019	0.316	0.328	0.279	0.367		0.301	0.266	0.297	
Nov2019	0.304	0.314	0.264	0.338		0.277	0.278	0.316	
Dec2019	0.311	0.323	0.259	0.339		0.289	0.290	0.318	
Yr2020	0.307	0.318	0.262	0.348		0.308	0.273	0.313	
Sep2019	0.439	0.430	0.383	0.390	0.278	0.339	0.350	0.344	**Vol.Y**
Oct2019	0.447	0.445	0.402	0.390		0.314	0.315	0.316	
Nov2019	0.435	0.433	0.415	0.419		0.336	0.348	0.345	
Dec2019	0.446	0.445	0.402	0.404		0.339	0.352	0.350	
Yr2020	0.418	0.410	0.380	0.380		0.352	0.359	0.354	

different from the Web to the RDD respondent populations and from both to the ACS target. Nothing guarantees that generalized-raking weight-adjustment will make the respondent population in a low-response-rate survey resemble the national population.

These remarks about the measurement of difference between adjusted-weight survey estimates and their targets are reinforced in Table 6. Perhaps the most interesting additional conclusion from that table is how similar from RDD to the Web survey is the pattern of significant differences between estimates and targets. This finding is surprising in light of the previous comparisons by Yeager et al. (2011) and MacInnes et al. (2018), generally in favor of the superior quality of RDD surveys. The failure to distinguish the quality of estimates between the two Tracking Surveys may reflect problems with RDD data quality highlighted in the reports of Ellis et al. (2022a,b), and may in particular be due to the data collectors ensuring that poststratification variables had almost no missing data in the Web survey while 10-15% of RDD subjects had missing data among these variables.

A more detailed measurement of weight-adjustment quality requires a different metric. Table 6 already showed that variables successfully estimated close to their national targets by poststratification may in their pairwise interactions be quite poorly adjusted. (For example, in the Web Survey `HHsize` and `Renter` proportions are nearly perfectly adjusted to their national targets by the raked adjustment, but `HH1:2Rent` is consistently off by a few percentage points.) Taking this idea further, we could cross-classify several demographic/geographic poststratifying variables and define as an alternative metric of adjustment quality between two different sets of weights the mean of absolute differences of their cell-by-cell weighted estimates.The quality of adjustment of a single set of weights to national targets could similarly be measured by the cell-wise sum of absolute discrepancies. An obstacle to implementing the latter metric of adherence to targets is that it is not easy to find multiply cross-classified characteristics reliably estimated in national statistics. The Census Bureau's Population Estimates are one such public source, a yearly national tabulation updating the decennial census. The file `SC-EST2019-ALLDATA5.csv` (Population Division 2020) was our source for 2019 national cross-classified proportions, aggregated to 5-group `AgeGp5`, 4-group `Race` (White Alone, Black Alone, Asian Alone, and Other), 4-Division `Region` (with categories grouping states into Northeast, Midwest, South, and West), `SEX` (M,F), and `Hisp` (Yes, No). The resulting 5-way table partitions the adult US 2019 population of 255,200,373 into 320 cells, and after dividing all cell counts by the total we treat it as a 5-way table of population proportions summing to 1.

The degree of cross-classification that makes sense in assessing agreement between survey and target proportions derived from multiple categorical variables depends on the sample size and extent of missing or unknown items. Therefore, we compare our adjusted weights and targets at a sequence of orders of cross-classification. For each order k of cross-classification, from 1 up to a maximum number K of variables for which cross-classified national data exist (here, $K = 5$), we compute for each set of k variables the sum of absolute cross-classified

Table 6: For 7 outcomes and 5 survey time-blocks, column of entries: estimated proportion with raked weights, SE, discrepancy from target, and absolute discrepancy over SE (bolded when > 3)

Survey	Time	[18,34]	[45,64]	WA.Hsp	BA	HH1:2Rent	HH3+Own	Vol.Y
RDD	09/19	0.278	0.318	0.072	0.112	0.194	0.276	0.383
		0.004	0.000	0.009	0.010	0.009	0.010	0.013
		-0.020	-0.009	-0.073	-0.018	-0.042	0.025	0.105
		5.081	*	**8.245**	1.804	**4.711**	2.540	**7.845**
	10/19	0.275	0.314	0.079	0.102	0.202	0.279	0.402
		0.020	0.017	0.011	0.013	0.016	0.017	0.018
		-0.024	-0.013	-0.067	-0.028	-0.034	0.028	0.124
		1.204	0.761	**5.804**	2.128	2.097	1.641	**6.888**
	11/19	0.275	0.314	0.070	0.116	0.195	0.264	0.415
		0.018	0.017	0.009	0.015	0.015	0.015	0.018
		-0.023	-0.013	-0.075	-0.014	-0.041	0.013	0.137
		1.300	0.748	**8.034**	0.946	2.700	0.924	**7.762**
	12/19	0.277	0.316	0.086	0.120	0.212	0.259	0.402
		0.018	0.017	0.012	0.015	0.016	0.015	0.017
		-0.022	-0.011	-0.060	-0.010	-0.024	0.008	0.124
		1.185	0.620	**5.069**	0.650	1.486	0.504	**7.133**
	2020	0.278	0.317	0.073	0.120	0.204	0.262	0.380
		0.004	0.003	0.002	0.003	0.003	0.003	0.003
		-0.021	-0.010	-0.072	-0.010	-0.032	0.011	0.102
		5.893	2.865	**34.68**	**3.640**	**10.62**	**3.693**	**30.33**
Web	09/19	0.290	0.331	0.083	0.123	0.209	0.259	0.350
		0.003	0.000	0.006	0.006	0.006	0.008	0.011
		-0.008	0.004	-0.063	-0.006	-0.027	0.008	0.072
		3.350	*	**10.87**	0.999	**4.626**	1.061	**6.564**
	10/19	0.290	0.331	0.075	0.137	0.217	0.266	0.315
		0.013	0.013	0.008	0.010	0.012	0.013	0.013
		-0.008	0.004	-0.070	0.007	-0.019	0.015	0.037
		0.654	0.351	**9.324**	0.750	1.613	1.205	2.811
	11/19	0.290	0.331	0.072	0.146	0.211	0.278	0.348
		0.012	0.013	0.007	0.010	0.011	0.012	0.013
		-0.008	0.004	-0.073	0.016	-0.025	0.027	0.070
		0.678	0.359	**10.75**	1.689	2.197	2.266	**5.209**
	12/19	0.290	0.331	0.075	0.138	0.206	0.290	0.352
		0.012	0.013	0.006	0.010	0.011	0.012	0.013
		0.008	0.004	-0.070	0.008	-0.030	0.039	0.074
		0.677	0.358	**10.85**	0.870	2.723	**3.263**	**5.669**
	2020	0.290	0.331	0.083	0.117	0.213	0.273	0.359
		0.002	0.003	0.001	0.002	0.002	0.002	0.003
		0.008	0.004	-0.063	-0.013	-0.023	0.022	0.081
		3.426	1.726	**43.60**	**7.086**	**9.919**	**9.480**	**30.32**

Table 7: Tabulated sum of absolute cellwise differences between RDD survey-weighted estimates and Pop-Estimate targets, by weighting-stage, survey time-block, and order of cross-classification. The five variables used in these cross-classifications were: `AgeGp5`, (4-category) `Race`, `Region`, `Hisp`, and `Sex`. The four weighting-stage results are shown in a column of 4 for each time-block and order of cross-classification.

	Weight	Order	1	2	3	4	5
Sep2019	Unwt		0.129	0.208	0.288	0.383	0.495
	Baswt		0.115	0.188	0.259	0.356	0.469
	Raked		0.057	0.114	0.181	0.288	0.437
	Y&Rwt		0.062	0.113	0.189	0.297	0.428
Oct2019	Unwt		0.160	0.260	0.357	0.461	0.587
	Baswt		0.147	0.239	0.323	0.425	0.556
	Raked		0.064	0.130	0.202	0.311	0.462
	Y&Rwt		0.061	0.119	0.205	0.329	0.487
Nov2019	Unwt		0.130	0.224	0.308	0.409	0.526
	Baswt		0.116	0.198	0.273	0.370	0.492
	Raked		0.059	0.125	0.201	0.302	0.446
	Y&Rwt		0.070	0.133	0.214	0.328	0.473
Dec2019	Unwt		0.151	0.250	0.334	0.431	0.542
	Baswt		0.139	0.233	0.311	0.403	0.520
	Raked		0.055	0.119	0.198	0.306	0.449
	Y&Rwt		0.051	0.112	0.186	0.283	0.416
Yr2020	Unwt		0.112	0.189	0.261	0.336	0.409
	Baswt		0.104	0.170	0.232	0.299	0.370
	Raked		0.049	0.098	0.147	0.196	0.254
	Y&Rwt		0.049	0.096	0.147	0.208	0.277

cellwise differences between survey-estimated and/or target proportions. (In calculating this sum, survey data with missing variables are omitted, so that the cellwise survey proportions are defined by summing the weights for all survey subjects with the non-missing levels of survey data for the cross-classified variables and dividing by the total of weights for survey subjects with non-missing data for those variables.) We report for each k from 1 to $K = 5$ the average of these summed-absolute-deviation metrics over all sets of k cross-classifiying variables. Table 7 shows these discrepancy metrics between survey weights and Pop-Estimates targets, by survey time-block, for Unweighted, Base-Weighted, Raked, and Y&R-adjusted RDD survey data. The corresponding Table for the Web survey looked similar, and so is not shown.

Table 8: Tabulated sum of absolute cellwise differences between survey-weighted estimates and Pop-Estimate targets averaged over all 3-way cross-classifications of 5 variables, displayed by survey time-block and weighting-stage, for both the RDD and Web surveys.

	RDD				Web		
	Unwt	Baswt	Raked	Y&Rwt	Unwt	Raked	Y&Rwt
Sep2019	0.288	0.259	0.181	0.189	0.251	0.217	0.219
Oct2019	0.357	0.323	0.202	0.205	0.252	0.212	0.220
Nov2019	0.308	0.273	0.201	0.214	0.197	0.185	0.177
Dec2019	0.334	0.311	0.198	0.186	0.205	0.195	0.191
Yr2020	0.261	0.232	0.147	0.147	0.202	0.159	0.158

The summed absolute deviation metric here is a categorical version of `total variation distance` between two probability measures. Unsurprisingly, this metric is larger for very detailed (4- and 5-way) cross-classifications, and it is remarkably high at orders 4 and 5, even after raking. It is generally but not perfectly consistent across months for the same level of survey weighting. In Table 7, at each order of cross-classification, the survey weights ranked from highest to lowest metric value are usually in the order `Unwt`, `Baswt`, `Y&Rwt`, and `Raked`. By this metric, as in Table 6, our Raking often but not always outperforms the Y&R raking with randomized hotdeck imputation and weight-trimming.

Table 8 displays the differences in these summed-absolute-difference metrics month-to-month and between the RDD and Web surveys, restricting attention only to the 3-way cross-classifications. (The RDD portion of Table 8 was already displayed in the Order-3 column of Table 7.) By this measure, our raking and the Y&R raking are about equally good. But in this table, the distance between the 3-way cross-classified unweighted RDD-survey is seen to be definitely farther from the national targets than was the Web-survey population. This is yet another way in which the populations accessed by the data-collectors favored the Web survey by comparison with the RDD.

6.4 Comparison among SE estimators at the final weighting stage

The standard errors in Table 6 are obtained as described in Section 6, the theory for which is developed rigorously only for the case where the base-weights are regarded as fixed and given, where the poststratification variables are never missing for survey respondents, and (A.4) holds for the true survey weights. Throughout our discussion, supported by Remark 5 in Section 5.1, we assume that unequally-weighted Poisson sampling adequately describes the RDD and Web data-collection. We display in Table 9 the differences between SE estimates done at the three different levels of sophistication discussed in Section 6, the highest of which has informed the values computed for Table 6. The simplest version, formula (18) treats the

final (raked) weights w_i as though they were given and known, and does not take into account the data-dependent aspects of base-weighting or raking. The next, more reasonable, variance formula is (19), computed using a weighted-least-squares set of GREG model estimates with the (uniform or) base-weights w_i^o. The most general variance formula again has the form (19), but now with GREG coefficients computed using the adjusted weights $\hat{w}_i$ from (17), and with $\hat{w}_i$ replacing w_i^o. As suggested in Section 5.1, the first of these three SEs is generally larger than the second because the second takes into account the raking weight adjustment. This is especially noticeable for the variables closely related to raking variables (`Age`, and the `HHsize` by `Rent` interaction), less so for other variables. The other general statement from Section 5.1 confirmed in Table 9 is that the third SE estimate that takes into account the large change from base to final weights (i.e., that relies on the correctness of **(A.4)** rather than the correctness of the base-weights w_i^o) is larger than the estimates (19). The third SE estimate is hardly larger than the second for some of the outcome variables (`WA.Hsp`, `BA`, `HH1:2Rent`, `HH3+Own`) but the excess is often greater for the other variables, although for unexplained reasons the difference between the second and third SE estimates varies noticeably across month and from the RDD to the Web survey.

7 Summary & Conclusions

Many of the answers to questions asked in sample surveys are strongly conditioned on demographics. Therefore, surveys whose respondent populations have major demographic categories appearing in markedly different proportions than the national population are adjusted to make these differences disappear. These adjustments, generically called *post-stratification*, are usually made via cell-based ratios or by some form of (generalized) raking (Valliant et al. 2018). However, for reasons of simplicity and numerical feasibility, the adjustments are usually based on control totals of important categorical demographic variables singly or in cross-classified pairs. Since many survey outcome variables measuring social characteristics or attitudes are strongly predicted by single demographic categories or pairs of them, these outcome variables as measured by surveys align better with reliable national measurements after poststratification than before. However, many other survey outcome variables do not align so predictably with marginal demographics, and these outcome variables – that may have received responses from particular respondent populations that are very unrepresentative of the national target population – may not align better with the national population after poststratification. These remarks apply equally well to surveys conducted with probability samples as to nonprobability web samples and are especially relevant to surveys with extremely low response rates (e.g., rates of 10% or less), like the Tracking Surveys.

There is nevertheless a large body of social-science and survey-measurement research that shows well-designed poststratified probability surveys achieving fairly good agreement on non-demographic benchmark survey questions (such as whether an individual holds a

Table 9: For each of 7 outcomes and 5 survey time-blocks, column of 3 differently calculated SEs for survey-weighted estimates for Poisson sampling with final raked weights. First is the naive SE (18) treating final weights as though known; the 2nd is the large-sample form (19) treating base-weights w_i^o are essentially correct, but accounting for calibration; and 3rd is the variance (19) under **(A.4)**, with w_i^o replaced by $\hat{w}_i^o$.

Survey	Time	[18,34]	[45,64]	WA.Hsp	BA	HH1:2Rent	HH3+Own	Vol.Y
RDD	09/19	0.0149	0.0155	0.0087	0.0107	0.0129	0.0147	0.0158
		0.0038	0.0000	0.0087	0.0097	0.0087	0.0098	0.0134
		0.0040	0.0000	0.0089	0.0098	0.0089	0.0099	0.0134
	10/19	0.0164	0.0170	0.0099	0.0111	0.0142	0.0160	0.0173
		0.0136	0.0154	0.0067	0.0078	0.0112	0.0149	0.0157
		0.0197	0.0169	0.0115	0.0131	0.0160	0.0168	0.0180
	11/19	0.0154	0.0159	0.0088	0.0111	0.0131	0.0148	0.0163
		0.0140	0.0152	0.0080	0.0091	0.0118	0.0144	0.0158
		0.0180	0.0168	0.0094	0.0148	0.0153	0.0145	0.0176
	12/19	0.0155	0.0161	0.098	0.0115	0.0137	0.0151	0.0165
		0.0137	0.0153	0.0084	0.0085	0.0114	0.0149	0.0157
		0.0183	0.0172	0.0118	0.0148	0.0159	0.0153	0.0174
	2020	0.0029	0.0030	0.0017	0.0021	0.0025	0.0028	0.0030
		0.0027	0.0030	0.0015	0.0019	0.0023	0.0029	0.0030
		0.0035	0.0033	0.0021	0.0026	0.0030	0.0031	0.0034
Web	09/19	0.0114	0.0119	0.0069	0.0083	0.0103	0.0110	0.0121
		0.0030	0.0000	0.0062	0.0066	0.0060	0.0076	0.0108
		0.0025	0.0000	0.0058	0.0064	0.0058	0.0076	0.0110
	10/19	0.0125	0.0130	0.0073	0.0095	0.0114	0.0122	0.0129
		0.0104	0.0111	0.0068	0.0079	0.0098	0.0111	0.0112
		0.0129	0.0128	0.0075	0.0099	0.0119	0.0127	0.0133
	11/19	0.0116	0.0121	0.0066	0.0091	0.0105	0.0115	0.0123
		0.0110	0.0112	0.0065	0.0080	0.0097	0.0107	0.0115
		0.0124	0.0125	0.0068	0.0096	0.0112	0.0121	0.0134
	12/19	0.0116	0.0120	0.0067	0.0088	0.0103	0.0116	0.0123
		0.0107	0.0111	0.0065	0.0082	0.0098	0.0107	0.0114
		0.0125	0.0126	0.0065	0.0096	0.0110	0.0121	0.0130
	2020	0.0023	0.0023	0.0014	0.0016	0.0020	0.0022	0.0024
		0.0022	0.0022	0.0014	0.0015	0.0019	0.0022	0.0023
		0.0025	0.0026	0.0014	0.0018	0.0023	0.0024	0.0027

passport, or self-reports being in very good health, or many others) with the national proportions of categorical responses seen on those same questions in reliable national high-response-rate surveys (Yeager et al. 2011, MacInnes et al. 2018). That research has also tended to conclude that the agreement on such benchmarks after adjustment with national targets is worse with nonproprobability surveys than with well-designed probability surveys. Despite that research, there is no general methodological basis for assurance that unrepresentative respondent populations can be adjusted, by post-stratification or other means, to give useful information about survey responses that depend in complicated ways on the interaction between demographic and geographic categories and unmeasured variables.

One possible reason why low-response surveys, including political polls and other widely cited surveys of public attitudes, may fail to achieve representative results is that the self-selection of respondents depends importantly on high-order (at least 3-way) interactions among demographic categories. These interactions may remain important even in the presence of information about group affiliations (party, church, single-issue advocacy groups) and past voting. Yet many academic survey practitioners cite recent literature (e.g., Kreuter et al. 2010) to argue that low-response surveys may nevertheless have low bias, because of that paper's persuasive argument that it is difficult to find single survey variables that are simultaneously predictive of response and survey outcome variables. We ignore at our peril the predictive value of multiway interactions of demographic/geographic variables (often used in stratification), even though single cells of multiway contingency tables defined from these variables contain small population proportions. Benchmark variable categories should therefore be selected to cut across standard demographic categories, and possibly to combine pockets of the population exhibiting complex demographic interactions.

In Section 6 on data results from the Tracking Survey, we have considered several ways of assessing of survey weight-adjustment when the magnitude of adjustments is large, using survey outcome variables (benchmark variables) or variable combinations (cross-classified demographic/geographic variables) not used in poststratification. Theoretical results developed in the Appendix and interpreted in Section 5 establish under some assumptions the large-sample approximate normal distribution of survey-weighted estimates using poststratified weights, results which were not previously available when the base-weights are far from the adjusted weights. The theory was used to develop valid large-sample variance estimators in the setting of the Tracking Surveys, where (*cf.* Remark 5) survey subjects were arguably sampled independently. The theory applies to the Telephone survey because it was conducted under a probability design. With less plausible reliance on the independent-sampling and model assumptions, the theory applies also to the Web survey. The survey-outcome estimates and their estimated standard errors were computed and compared in Section 6 on the Tracking Survey data. Figures and tables were used to compare the survey estimates between the RDD and Web surveys and across survey months. Standardized discrepancies between estimated proportions of benchmark and demographic categories from national targets were assessed (using the `t`-tests implicit in Table 6) and compared between RDD and

Web surveys. In addition, we introduced for both surveys a novel method of assessment of summed absolute deviations (between adjusted-weight estimates and targets) over cross-classified multiway tables of higher orders (3, 4 and 5). In the Tracking Survey application, the estimates and standard errors were fairly stable across months and did not show conclusive superiority of the RDD or Web surveys. However, the Web survey turned out to have considerably less missing survey-item data than the RDD survey. The companion reports of Ellis et al. (2022 a,b) study more deeply whether the unexpectedly similar performance of the RDD and Web estimates in the Tracking Surveys may have been due to ascertainable lapses in RDD data-quality.

A related issue treated in (Section 4 of) the report is the handling of partially missing data in survey weight-adjustment. The contractor had developed base and final weights using randomized hot-deck imputation methods (Team Y&R 2019), while our estimates injected no new randomness and amounted to mass imputation of the missing-data proportions during calibration by the national target proportions. In benchmark-variable comparisons, our estimates based on adjusted weights showed slightly better agreement across RDD and Web and between these and the national targets than did the estimates based on Team Y&R adjusted weights. However, the difference in performance was small and did not show up at all when assessed by the new metric of summed absolute discrepancies over the cells of multiway cross-classified contingency tables.

This report has emphasized the inherently model-based nature of survey weight-adjustment in the setting where survey response-rates are low and the magnitude of adjustments is large. The weights being adjusted are equal in number to the respondents in the survey, more than 1400 in all self-contained time-blocks of the RDD or Web Tracking surveys. Each calibration method, raking or linear calibration or any of the other generalized raking methods contemplated by Deville and Särndal, implicitly assumes (as in **(A.4)**) that the ratio of the correct weights over the base-weights satisfies a parametric model of very modest dimension (less than 40) equal to the number of calibration constraints imposed by national-target constraints for all categorical levels of the poststratification variables. The construction of weights to satisfy the constraints is mathematically a very under-determined problem solved in practice by what amounts to a parametric model assumption on weights. This is a chink in the design-based armor of survey methodologists who advertize their results as not being model-based.

References

Andridge, R. and Little, R. (2010), A review of hot deck imputation for survey non-response, *Internat. Statist. Rev.* **78**, 40-64.

Agresti, A. (2013), **Categorical Data Analysis**, 3rd ed. Wiley.

Bishop, Y., Fienberg, S. and Holland, P. (1975), **Discrete Multivariate Analysis: Theory and Practice**, MIT Press.

Blumberg, S. and Luke, J. (2018), Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January-June 2018, National Center for Health Statistics, released Dec. 2018.

Buskirk, T. and Best, J. (2012) Venn Diagrams, Probability 101 and Sampling Weights Computed for Dual Frame Telephone RDD Designs, *Proc. Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association.

Carlin, J. (2015), Multiple Imputation: Ch. 12 in: **Handbook of Missing Data Methodology**, CRC / Chapman & Hall.

Chen, J. and Shao, J. (2001), Jackknife variance estimation for nearest-neighbor imputation, *Jour. Amer. Statist. Assoc.*, **96**, 260269.

Cochran, W. (1977), **Sampling Techniques**, 3rd ed. Wiley.

Devaud, D. and Tillé, Y. (2019), Deville and Särndal's calibration: revisiting a 25-years-old optimization problem, *TEST* **28**, 1033-1065.

Deville, J. and Särndal, C.-E. (1992), Calibration Estimators in Survey Sampling, *Jour. Amer. Statist. Assoc.* **87**, 376-382.

Deville, J., Särndal, C.-E. and Sautory, O. (1993), Generalized Raking Procedures in Survey Sampling, *Jour. Amer. Statist. Assoc.* **88**, 1013-1020.

Ellis, R., Krosnick, J. et al. (2022a,b), Summary Reports on Data Quality and Comparisons between Probability and Nonprobability surveys in the 2019-2020 Census Bureau Tracking Survey, in preparation.

Fuller, W. (2009), **Sampling Statistics**, Wiley.

Hainmuller, J. (2012), Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis* **20**, 25-46.

Kreuter, F. , Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T., Cases-Corder, C., Lernay, M., Peytchev, A., Grovers, R. and Raghunathan, T. (2010), Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys, *Jour. Roy. Statist. Soc. Ser. A* **173**, 389-407.

Lumley, T. (2010), **Complex Surveys**, Wiley.

MacInnes, B., Krosnick, J., Ho, A. and Cho, M.-J. (2018), The accuracy of measurements with probability and nonprobability survey samples, *Public Opinion Quarterly* **82**, 707-744.

Oh, H. and Scheuren, F. (1983) Weighting adjustment for unit nonresponse. In: *Incomplete Data in Sample Surveys*, vol. 2, Eds. Madow, W., Olkin, I. and Rubin, D. New York: Academic Press, 143-184.

Pasek, J. (2010), ANES Weighting Algorithm, A Description, Stanford Univ. preprint.

Pasek, J. DeBell, M. and Krosnick, J. (2014), Standardizing and Democratizing Survey Weights: The ANES Weighting System and `anesrake`, Stanford Univ. preprint.

Pasek, J. and Krosnick, J. (2020), Relations Between Variables and Trends Over Time in RDD Telephone and Nonprobability Sample Internet Surveys, *Jour. Survey Statist. & Methodology* **8**, 37-61.

Oh, H. and Scheuren, F. (1983) Weighting adjustment for unit nonresponse. In: **Incomplete Data in Sample Surveys**, vol. 2, Eds. Madow, W., Olkin, I. and Rubin, D. New York: Academic Press, 143-184.

Pollard, D. (1980), **Convergence of Stochastic Processes**, Springer.

Population Division (2020), Annual State Resident Population Estimates for 5 Race Groups (5 Race Alone or in Combination Groups) by Age, Sex, and Hispanic Origin: April 1, 2010 to July 1, 2019, US Census Bureau (release date: June 2020), `www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html`

Rao, J.N.K. and Shao, J. (1992), Jackknife variance estimation with survey data following hot-deck imputation, *Biometrika* **79**, 811-822.

Rothbaum, J. and Bee, A. (2021), Coronavirus infects surveys too: survey nonresponse bias and the coronavirus pandemic, Census Bureau Preprint.

Rubin-Bleuer, S. and Kratina, i. (2005), On the two-phase framework for joint model and design-based inference, *Ann. Statist.* **33**, 2789 - 2810.

Särndal, C.-E., Swensson, J. and Wretman, J. (1992), **Model-Assisted Survey Sampling**, Springer.

Slud, E. and Thibaudeau, Y. (2010), Simultaneous Calibration and Nonresponse Adjustment, Census Bureau CSRM Research Report RR2010-3.

Team Y&R (2019), Tracking Survey Post-Processing and Analysis Plan, internal Census Bureau project documentation.

Team Y&R, 2020 Census Tracking Survey. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-11-09.

Valliant, R., Dever, J. and Kreuter, F. (2018), **Practical Tools for Designing and Weighting Surveys**, Springer.

van Buuren, Stef (2015), Fully Conditional Specification: Ch. 13 in **Handbook of Missing Data Methodology**, CRC / Chapman & Hall.

Yeager, D., Krosnick, J., Chang, L., Javitz, H., Levendusky, M., Simpers, A. and Wei, R. (2011), Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability snd non-probability samples, *Public Opinion Quarterly* **75**, 707-747.

Appendix on Design-Based Asymptotic Theory

A Setting and Notation

The general setting for the survey inference problems we consider is a frame population list $\mathcal{U}$ of N units, n units of which are respondents in a designed probability sample drawn from $\mathcal{U}$. Unit nonrespondents from the larger designed sample are ignored, or else no information about them is available, and we allow the possibility that the respondent-set $\mathcal{R}$ is much smaller than the set $\mathcal{S}$ originally sampled. For purposes of survey inference, we maintain the fiction that $\mathcal{R} = \mathcal{S}$, a fiction that forces us to modify the design base-weights, the original single-inclusion probabilities $\pi_i^o = P(i \in \mathcal{S})$, to reflect the response mechanism as part of the sampling design. It is assumed that the design (base-) weights $w_i^o = 1/\pi_i^o$ are available for all $i \in \mathcal{R}$, along with the survey outcomes Y_i and covariates X_i for the respondent units. The covariates X_i are vectors of demographic and other observations on survey respondents, which may be categorical or numeric, and some of which may be missing for some survey respondents.

In this setting, the first task is to modify the initial or base weights w_i^o (generally taken to be identical across i for nonprobability surveys) to reflect known overall proportions for a set of demographic and geographic variables. The process of modifying these inital weights to a set w_i of final weights is called *weight-adjustment*, with the goal of enabling estimates of population averages $\bar{Y}$ of survey attributes Y_i to be calculated from survey-respondents in the survey-weighted form $\hat{\bar{Y}} = N^{-1} \sum_{i \in \mathcal{R}} w_i Y_i$. The modifications $w_i^o \mapsto w_i$ are generally done with the aid of auxiliary variables (covariates) $\underline{X}_i = (X_{i,k},\ k = 1, \ldots, p) \in \mathbb{R}^{p+1}$ to satisfy exact or approximate calibration constraints

$$N^{-1} \sum_{i \in \mathcal{R}} w_i X_{i,k} \ = \ \bar{X}_k \quad \text{for} \quad k = 0, \ldots, p \tag{20}$$

where the target population means $\bar{X}_k$ are known from external sources, and weights are scaled in such a way that $X_{i,0} \equiv \bar{X}_0 \equiv 1$. We consider various weight-adjustment schemes of this sort. Although some respondent covariate items $X_{i,k}$ may realistically be missing, in this technical Appendix we assume that none of the covariate items $X_{i,k}$ are missing.

As formulated in Deville and Särndal (1992) and Deville, Särndal and Sautory (1993), linear calibration, raking and other generalized raking extensions follow the same pattern. Each is expressed as an optimization problem constrained by (20) in terms of a summed unit-level metric between w_i/w_i^o and 1:

$$\min_{\underline{w}} \sum_{i \in \mathcal{S}} w_i^o \, G(w_i/w_i^o) \qquad \text{subject to} \qquad \frac{1}{N} \sum_{i \in \mathcal{S}} w_i \, X_{i,k} \ = \ \bar{X}_k \qquad \forall \ k = 0, \ldots, p \tag{21}$$

where $G(x)$ is a known smooth function satisfying $G(1) = G'(1) = 0, G''(1) = 1$. The most common choices for G, apart from $G_{lin}(x) \equiv (x-1)^2/2$ leading to linear calibration, are

$$G_{rak}(x) = x\log(x) - x + 1, \quad G_{logis}(x) = \frac{(1-L)(U-1)}{U-L}\left\{(x-L)\log(\frac{x-L}{1-L}) + (U-x)\log(\frac{U-x}{U-1})\right\}$$

respectively associated with raking and with a "logistic" form of calibration guaranteed to yield weight-ratios w_i/w_i^o in a fixed interval (L, U) of positive numbers containing 1. Unlike linear calibration, generalized-raking methods using G_{rak} or G_{logis} are guaranteed to result in positive weights. Although G is not always assumed strictly convex, we do assume it.

The constrained optimization problem (21) is equivalently written in terms of Lagrange multipliers $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ as the unconstrained minimization of

$$\min_{\underline{w}} \left\{ \left(\frac{1}{N}\sum_{i\in\mathcal{S}} w_i^o\, G(w_i/w_i^o) \;-\; \sum_{k=0}^{p} \beta_k \left(\frac{1}{N}\sum_{i\in\mathcal{S}} w_i\, X_{i,k} - \bar{X}_k\right)\right\} \qquad (21)'$$

Minimization is equivalent (with unique solution because of convexity of G) to solution of

$$G'(w_i/w_i^o) \;=\; \big(\mathbf{X}\beta\big)_i \quad \forall\; i\,, \qquad \big(\bar{X}_k\big)_{k=0}^{p} \;=\; \frac{1}{N}\,\mathbf{X}^{tr}\left(w_i^o\, G'^{\,-1}\big((\mathbf{X}\beta)_i\big)\right)_{i\in\mathcal{S}} \qquad (22)$$

where the $n \times (p+1)$ design-matrix $\mathbf{X}$ is defined by $\mathbf{X}_{i,k} = X_{i,k}$ for $i \in \mathcal{S},\ k = 0, \ldots, p$, and $n = |\mathcal{S}|$. Where needed below, the rows of the design-matrix $\mathbf{X}$ are denoted by $\mathbf{X}_i \in \mathbb{R}^{p+1}$ as (column) $(p+1)$-vectors. That is, with w_i^o, $X_{i,k}$, and $\bar{X}_k$ known in advance, one determines the column vector $\beta = (\beta_0, \ldots, \beta_p)$ of Lagrange multipliers from the first equation in (17) to define w_i as a function of β and then solving for β in the second equation in (22).

A.1 Assumptions

There are far more sampled responding units (and therefore weights) than covariates, generally, so the property of weights performing adequately is generally 'underdetermined' in the sense that many systems of N weights could in principle be used with equal success in reproducing known population totals. In practice, the weights in generalized raking are found through an essentially parametric system of equations, that is, the weight-ratios w_i/w_i^o are parameterized in terms of Lagrange multipliers β.

We collect in this section the mathematical large-sample assumptions needed to define a design-based sense in which the set of weights $\{w_i^*\}_{i\in\mathcal{S}}$ can serve as 'true' weights. These assumptions and properties will then be used in later sections to establish sufficient conditions for large-sample limiting behavior of the estimates $\sum_{i\in S}\, \hat{w}_i\, Y_i$, where $\hat{w}_i$ are the generalized-raking estimates obtained by solving (22) for $\beta = \hat{\beta}$ and $\{w_i\}_{i\in\mathcal{S}} = \{\hat{w}_i\}_{i\in\mathcal{S}}$.

We continue with the notations of Appendix A and repeat here the preliminary assumptions or restrictions made there, clarifying the asymptotic framework for large n, N.

(A.1) $G : (0,\infty) \to (\infty)$ is twice continuously differentiable, with $G(1) = G'(1) = 0$, $G''(1) = 1$, and everywhere $G''(\cdot) > 0$.

Because G is twice continuously differentiable and strictly convex, G' is increasing and invertible, and $(G')^{-1}$ is strictly increasing and differentiable. Moreover, in what follows, the domain and range of the function G will be restricted in such a way that $(G')^{-1}$ is uniformly positive. Two functions related to G' play an important role in the results of this Appendix:

$$\kappa(z) \equiv \frac{d}{dz}(G')^{-1}(z) \quad , \qquad \gamma(z) \equiv \kappa(z)/(G')^{-1}(z) \tag{23}$$

(A.2) For $n \geq n_0 > 0$, and $N = N(n) > n$, there are a finite population $\mathcal{U} = \mathcal{U}_N = \{1, \ldots, N(n)\}$ of indices, a fixed vector $(w_i^o,\ i = 1, \ldots, N)$ of base weights, and arrays of (known or observed) scalar constants $X_{i,k}, Y_i$ for $i = 1, \ldots, N,\ k = 0, \ldots, p$, where $X_{i,0} \equiv 1$. The absolute values of $X_{i,k}, Y_i$ are all $\leq C$, and the individual weights w_i^o are all $\leq C'N/n$, where C, C' are constants not depending on n, N.

The weights w_i^o, w_i studied in this Appendix are un-scaled first-order inclusion weights approximating the reciprocals of conditional first-order selection probabilities $P(i \in \mathcal{S} \,|\, \mathbf{X}_i)$. The survey of the finite population is used to supply estimates $\sum_{i\in\mathcal{S}} w_i Y_i$ of unknown total outcomes $\bar{Y} = \sum_{i\in\mathcal{U}} Y_i$ in a real population for which the corresponding $N^{-1}\sum_{i\in\mathcal{S}} w_i X_{i,k}$ is intended to represent the actual known average outcome $\bar{X}_k = N^{-1}\sum_{i\in\mathcal{U}} X_{i,k}$.

The next assumptions define what it means for a system $\{w_i^*\}_{i\in\mathcal{U}}$ to be an asymptotically correct set of weights, relative to auxiliary predictor/poststratification variables $\mathbf{X}_i$ and outcome variables Y_i. Because assumption **(A.3)** requires sums weighted by w_i^* to satisfy a Central Limit Theorem, it restricts also the probability sampling design resulting in the survey samples $\mathcal{S}$.

(A.3) There exists for each $N = N(n)$ a system of positive weights $\{w_i^*\}_{i=1}^N$ satisfying the following properties, as $n, N \to \infty$:

$$\frac{\sqrt{n}}{N}\sum_{i\in\mathcal{S}} w_i^* \left(\begin{pmatrix} Y_i \\ \mathbf{X}_i \end{pmatrix} - \begin{pmatrix} \bar{Y} \\ \bar{\mathbf{X}} \end{pmatrix} \right) \xrightarrow{\mathcal{D}} \mathcal{N}_{p+2}(\underline{\mathbf{0}},\, B) \tag{A.3.1}$$

where the $(p+2)\times(p+2)$ matrix B is specific to the attributes $Y, \mathbf{X}$, and for all bounded measurable $h : \mathbb{R}^{p+1} \to \mathbb{R}$,

$$N^{-1}\sum_{i\in\mathcal{S}} w_i^*\, h(\mathbf{X}_i) \xrightarrow{P} \int h(\mathbf{z})\,\nu(d\mathbf{z}) \quad , \quad N^{-1}\sum_{i\in\mathcal{S}} w_i^*\, h(\mathbf{X}_i)\, Y_i \xrightarrow{P} \int h(\mathbf{z})\,\eta(d\mathbf{z}) \tag{A.3.2}$$

where ν is a nondegenerate probability measure and η a signed measure supported on $[-C, C]^{p+1} \subset \mathbb{R}^{p+1}$, and

$$A \equiv \int \mathbf{z}^{\otimes 2}\,\nu(d\mathbf{z}) \qquad \text{is positive-definite} \qquad (A.3.3)$$

where $\mathbf{x}^{\otimes 2} \equiv \mathbf{x}\,\mathbf{x}' = \mathbf{x}\,\mathbf{x}^{tr}$.

By combination of (A.3.1) and (A.3.2) respectively using $h(\mathbf{z}) = \mathbf{z}$ or $h(\mathbf{z}) = 1$,

$$\bar{\mathbf{X}} - \int \mathbf{z}\,\nu(d\mathbf{z}) \to \mathbf{0} \quad , \qquad \bar{Y} - \int \eta(d\mathbf{x}) \to \mathbf{0} \qquad \text{as} \qquad n, N \to \infty$$

Because weight-adjustment will be performed via generalized-raking optimization (21), our next assumption expresses a parametric relationship between some asymptotically correct system of weights w_i^* and a linear combination of covariates, $\sum_{k=0}^{p} X_{i,k}\,\beta_k^o$. Although one may view (A.3)-(A.4) as a parametric form for $w_i^* = 1/P(i \in \mathcal{S}\,|\mathbf{X})$, the notion of 'asymptotically correct weights' given in (A.3) is actually more general.

(A.4) There exists a coefficient vector $\beta^o = \beta \in \mathbb{R}^{p+1}$ such that a system of weights w_i^* (defined for each N) satisfying (A.3) has the form $w_i^* = w_i^o \cdot (G')^{-1}(\mathbf{X}_i'\,\beta^o)$, where G is as in (A.1).

The weights w_i^* are intended to be uniformly positive. For many possible choices of G, including G_{rak} and G_{logis}, the values $(G')^{-1}$ are automatically positive. To cover other G's such as G_{lin}, we make an assumption.

(A.5) There exist $c_1, r_1 > 0$ such that, for all $\|\beta - \beta^o\| \le r_1$ and $\mathbf{x} = (x_0, \ldots, x_p) \in \mathbb{R}^{p+1}$ satisfying $|x_k| \le C$, also $(G')^{-1}(\mathbf{x}'\,\beta) \ge c_1$.

Throughout the proof arguments in the next Section, $\beta \in \mathbb{R}^{p+1}$ will be restricted to lie in the ball $B(r_1, \beta^o)$ of radius r_1 around the point β^o defined in (A.4). Together with assumption (A.5) and the bound $|X_{i,k}| \le C$ given in (A.2), this leads to the uniform bounds

$$\|\beta - \beta^o\| \le r_1 \quad , \qquad |\beta'\mathbf{X}_i| \le C_0 < \infty \quad , \qquad 0 < c_1 \le (G')^{-1}(\beta'\mathbf{X}_i) \le C_1 < \infty \qquad (24)$$

B Technical Results and Proofs

The first required technical results say that for large n, with probability close to 1, the random vector-valued function

$$K(\beta) \equiv N^{-1} \sum_{i \in S} w_i^o\,\mathbf{X}_i\,(G')^{-1}(\mathbf{X}_i'\,\beta) - \bar{\mathbf{X}} \qquad (25)$$

has a unique root in $B(r_1, \beta^o)$. The interpretation is that the Lagrange multipliers β and generalized-raking weights w_i defined by (22) as solutions to (21) are uniquely defined.

Lemma 1 *Assume (A.1)-(A.5). Then there exists $c > 0$ such that with probability converging to 1 as $n \to \infty$, the symmetric $p \times p$ Jacobian matrix $\nabla_\beta^{tr} K(\beta)$ with (j,k) entry $\partial K_j(\beta)/\partial\beta_k$ (for $0 \le j,k \le p$) has smallest eigenvalue $\ge c$. Moreover, with probability approaching 1 as $n \to \infty$, the function $K(\beta)$ has a unique root $\hat\beta \in B(r_1, \beta^o)$ and $\hat\beta - \beta^o = O_P(n^{-1/2})$ in the sense that*

$$\limsup_{n\to\infty} P(\|\beta - \beta^o\| \ge b/\sqrt{n}) \to 0 \qquad as \quad b \to \infty$$

Proof. First, (A.4) implies that

$$K(\beta^o) \;=\; N^{-1} \sum_{i\in\mathcal{S}} w_i^* \, \mathbf{X}_i \;-\; \bar{\mathbf{X}} \tag{26}$$

By the properties of G collected in (A.1) and the restriction of $|\mathbf{X}_i'\,\beta|$ as in (24),

$$\kappa(z) \;\equiv\; \frac{d}{dz}\,(G')^{-1}(z) \quad , \qquad 0 \;<\; c_2 \le\; \kappa(z) \;\le\; C_2 < \infty \quad \text{for} \quad |z| \le C_0 \tag{27}$$

Then for each $\beta \in B(r_1, \beta_0)$, by (A.4) and the definitions (23),

$$\nabla_\beta^{tr}\, K(\beta) \;=\; N^{-1} \sum_{i\in\mathcal{S}} w_i^o \, \mathbf{X}_i^{\otimes 2}\, \kappa(\mathbf{X}_i'\,\beta) \;=\; N^{-1} \sum_{i\in\mathcal{S}} w_i^* \, \left[\kappa(\mathbf{X}_i'\beta)/(G')^{-1}(\mathbf{X}_i'\beta^o)\right] \mathbf{X}_i^{\otimes 2}$$

The function of $\mathbf{X}_i$ in the last summation is uniformly bounded according to (24) and (27), so by (A.3.2) applied to (matrix components of) $h_*(\mathbf{x},\beta) = \mathbf{x}^{\otimes 2}\,\kappa(\mathbf{x}'\beta)/(G')^{-1}(\mathbf{x}'\beta^o)$,

$$\nabla_\beta^{tr}\, K(\beta) \;\xrightarrow{P}\; \int \mathbf{x}^{\otimes 2} \left[\kappa(\mathbf{x}'\beta)/(G')^{-1}(\mathbf{x}'\beta^o)\right] \nu(d\mathbf{x}) \qquad \text{as} \qquad n,\, N \to \infty \tag{28}$$

Convergence in (28) occurs for each β, but the integrands $h_*(x,\beta)$ are uniformly continuous in β on the bounded region $B(r_1,\beta_0)$ and therefore can be 'bracketed' uniformly within ϵ by a finite number of uniformly bounded functions to which (A.3.2) also applies. It follows by the standard bracketing arguments explained in Pollard (1984, Sec. II.2) that convergence in (28) is actually uniform in β over the ball $B(r_1,\beta^o)$: as $n,\, N \to \infty$,

$$\sup_{\beta\in B(r_1,\beta^o)} \left|\, \nabla_\beta^{tr}\, K(\beta) \;-\; \int \mathbf{x}^{\otimes 2} \left[\kappa(\mathbf{x}'\beta)/(G')^{-1}(\mathbf{x}'\beta^o)\right] \nu(d\mathbf{x}) \,\right| \;\xrightarrow{P}\; 0 \tag{29}$$

By the Intermediate Value form of the Mean Value Theorem, with $\tilde\beta$ denoting a value on the line segment between β_0 and β, and with $h_*(\mathbf{x},\beta)$ as defined just above (28),

$$K(\beta) \;=\; K(\beta^o) \;+\; \left[\left\{\nabla^{tr} K(\tilde\beta) \;-\; \int h_*(\mathbf{x},\tilde\beta)\,\nu(d\mathbf{x})\right\} \;+\; \int h_*(\mathbf{x},\tilde\beta)\,\nu(d\mathbf{x})\right)\Big] \,(\beta - \beta_0)$$

In the last displayed equation, $K(\beta_0) = O_P(1/\sqrt{n})$ by (26) and (A.3.1), and by (27) the bracketed matrix multiplying $\beta - \beta^o$ has minimum eigenvalue uniformly bounded $\geq (1-\delta)\,\lambda_*\,c_2/C_1$ for $\beta \in B(r_1, \beta^o)$, for arbitrary $\delta > 0$, where λ_* is the smallest eigenvalue of A in (A.3.3). Therefore, the displayed equation implies for arbitrary $\epsilon > 0$ there is a large constant C^* such that for large enough n with probability $\geq 1-\epsilon$, $\|K(\beta^o)\| \leq C^*/\sqrt{n}$ and

$$\inf\big\{\|K(\beta)\| : \ (2\,C_1\,C^*)/(c_2\,\lambda_*\,\sqrt{n}) \leq \|\beta - \beta^o\| \leq r_1\big\} \ \geq \ 2C^*/\sqrt{n}$$

Therefore, with high probability $\|K(\beta)\|$ has a minimum on $B(r_1, \beta^o)$ that actually lies within a ball of radius $2C_2\,C^*/(c_3\,\lambda_*\,\sqrt{n})$ around β^o. By the Implicit Function Theorem, this minimum is unique and is actually a root of $K(\beta)$. The Lemma is proved. □

The consequence of Lemma 1 is that the solutions β, w_i of (22) are uniquely defined as $\hat\beta$ and $\hat w_i = (G')^{-1}(\mathbf{X}_i'\,\hat\beta)$. The asymptotic distribution of $\hat\beta$ enables statements about the asymptotic distribution and variance estimation for calibrated survey estimates $\sum_{i\in\mathcal{S}} \hat w_i\,Y_i$.

Theorem 1 *Assume (A.1)-(A.5), recall the notation $\gamma(z)$ defined in (23), and define*

$$M \ = \ \int \mathbf{x}^{\otimes 2}\,\gamma(\mathbf{x}'\beta^o)\,\nu(d\mathbf{x}) \quad , \qquad \mathbf{m} \ \equiv \ \int \mathbf{x}\,\gamma(\mathbf{x}'\beta^o)\,\eta(d\mathbf{x}) \ - \ \bar Y \int \mathbf{x}\,\gamma(\mathbf{x}'\beta^o)\,\nu(d\mathbf{x})$$

and denote by B_{22} the lower-right $(p+1)\times(p+1)$ sub-matrix of B. (Thus, B_{22} is the asymptotic variance matrix of $\sqrt{n}\,K(\beta^o)$ in (A.3.1).) Then as $n, N \to \infty$,

$$\sqrt{n}\,\big(\hat\beta \ - \ \beta^o\big) \ + \ M^{-1}\,\sqrt{n}\,K(\beta^o) \ \xrightarrow{P} \ 0 \tag{30}$$

$$\sqrt{n}\,\big(\hat\beta \ - \ \beta^o\big) \ \xrightarrow{\mathcal{D}} \ \mathcal{N}\Big(0,\ M^{-1}\,B_{22}\,M^{-1}\Big) \tag{31}$$

$$\frac{\sqrt{n}}{N}\,\Big(\sum_{i\in\mathcal{S}} \hat w_i\,Y_i \ - \ \bar Y\Big) \ - \ \frac{\sqrt{n}}{N}\,\Big[\sum_{i\in\mathcal{S}} w_i^*\,\big(Y_i \ - \ \bar Y \ - \ \mathbf{m}'M^{-1}\,(\mathbf{X}_i - \bar{\mathbf{X}})\big)\Big] \ \xrightarrow{P} \ 0 \tag{32}$$

$$\frac{\sqrt{n}}{N}\,\Big(\sum_{i\in\mathcal{S}} \hat w_i\,Y_i \ - \ \bar Y\Big) \ \xrightarrow{\mathcal{D}} \ \mathcal{N}\Big(0\,,\ \begin{pmatrix} 1 \\ -M^{-1}\,\mathbf{m}\end{pmatrix}^{tr} B \begin{pmatrix} 1 \\ -M^{-1}\,\mathbf{m}\end{pmatrix}\Big) \tag{33}$$

Proof. As in the displayed equation following (29), with $\hat\beta$ replacing β so that $K(\hat\beta) = \mathbf{0}$,

$$\mathbf{0} \ = \ K(\beta^o) \ + \ \Big[\int \mathbf{x}^{\otimes 2}\big(\kappa(\mathbf{x}'\,\tilde\beta)/(G')^{-1}(\mathbf{x}'\,\beta^o)\big)\,\nu(d\mathbf{x}) \ + \ o_P(1)\Big]\,(\hat\beta - \beta^o)$$

where $\tilde\beta$ is now on the line segment between β^o and $\hat\beta$. Multiplying through by $\sqrt{n}$, and rearranging terms, we have

$$\sqrt{n}\,\big(\hat\beta - \beta^o\big) \ = \ \big[-\int \mathbf{x}^{\otimes 2}\big(\kappa(\tilde\beta'\mathbf{x})/(G')^{-1}(\tilde\beta'\mathbf{x})\big)\,\nu(d\mathbf{x}) \ + \ o_P(1)\big]^{-1}\,\sqrt{n}\,\big(K(\beta^o) - \bar{\mathbf{X}}\big)$$

Because κ and $(G')^{-1}$ are continuous and uniformly bounded on the support of ν, which is compact and falls in the region $\{\mathbf{x}: \ |x_k| \leq C \ \text{ for } \ k = 0, \ldots, p\}$, the square-bracketed term in the last equation (before) inverting converges in probability to $-M$. Thus we have proved (30). Since $K(\beta^o) = N^{-1} \sum_{i \in \mathcal{S}} w_i^* \, \mathbf{X}_i$ by (A.4), the limiting distribution of $\sqrt{n}\,(K(\beta^o) - \bar{\mathbf{X}})$ is $\mathcal{N}(\mathbf{0}, \, B_{22})$ by (A.3.1), and this fact together with (30) immediately implies (31).

Now first-order Taylor expansion of the summands of $\hat{\bar{Y}} - \bar{Y} \equiv N^{-1} \sum_{i \in \mathcal{S}} \hat{w}_i \, Y_i - \bar{Y}$ using $\hat{w}_i = w_i^o \cdot (G')^{-1}(\hat{\beta}' \mathbf{X}_i)$ and the constraint-relation $\sum_{i \in \mathcal{S}} \hat{w}_i = N$ yields

$$N^{-1} \sum_{i \in \mathcal{S}} \hat{w}_i Y_i - \bar{Y} \ = \ N^{-1} \sum_{i \in \mathcal{S}} w_i^* \left((G')^{-1}(\mathbf{X}_i' \hat{\beta}) \big/ (G')^{-1}(\mathbf{X}_i' \beta^o) \right) (Y_i \, - \, \bar{Y}) \ =$$

$$\frac{1}{N} \sum_{i \in \mathcal{S}} w_i^* \, (Y_i - \bar{Y}) \, + \, \big(\frac{1}{N} \sum_{i \in \mathcal{S}} w_i^* \left[\kappa(\mathbf{X}_i' \, \beta^o) / (G')^{-1}(\mathbf{X}_i' \, \beta^o) \right] \mathbf{X}_i \, (Y_i - \bar{Y}) \big)^{tr} \, (\hat{\beta} - \beta^o) + o_P\big(\frac{1}{\sqrt{n}}\big)$$

Multiply through by $\sqrt{n}$ and apply (A.3.2) to $h(\mathbf{z}) = \mathbf{z} \, \kappa(\mathbf{z}' \tilde{\beta}) \big/ (G')^{-1}(\mathbf{z}' \beta^o)$, to find

$$\frac{\sqrt{n}}{N} \big(\sum_{i \in \mathcal{S}} \hat{w}_i Y_i - \bar{Y} \big) \ = \ \frac{\sqrt{n}}{N} \sum_{i \in \mathcal{S}} \left(w_i^* \, Y_i \, - \, \bar{Y} \right) \, + \, \mathbf{m}^{tr} \sqrt{n} \, (\hat{\beta} - \beta^o) \, + \, o_P(1)$$

Next substitute for $\sqrt{n} \, (\hat{\beta} - \beta^o)$ using (30) and (26) to learn

$$\frac{\sqrt{n}}{N} \, \big(\sum_{i \in \mathcal{S}} \hat{w}_i Y_i - \bar{Y} \big) \ = \ \frac{\sqrt{n}}{N} \sum_{i \in \mathcal{S}} w_i^* \Big(Y_i - \bar{Y} \, - \, \mathbf{m}^{tr} \, M^{-1} \, (X_i - \bar{X}) \Big) \, + \, o_P(1)$$

$$= \begin{pmatrix} 1 \\ -M^{-1}\mathbf{m} \end{pmatrix}^{tr} \frac{\sqrt{n}}{N} \sum_{i \in \mathcal{S}} w_i^* \begin{pmatrix} Y_i - \bar{Y} \\ \mathbf{X}_i - \bar{\mathbf{X}} \end{pmatrix} + o_P(1) \ \xrightarrow{\mathcal{D}} \ \mathcal{N}\Big(0, \, \begin{pmatrix} 1 \\ -M^{-1}\mathbf{m} \end{pmatrix}^{tr} B \begin{pmatrix} 1 \\ -M^{-1}\mathbf{m} \end{pmatrix} \Big)$$

where the last convergence in distribution is an immediate consequence of (A.3.1). $\square$

The asymptotic variance expressions given in (31) and (33) in Theorem 1 immediately enable estimators for asymptotic survey estimates based on calibrated weights. However, the survey variance estimates look different based on the magnitude of survey weight-adjustments (i.e., on whether β^o is $\mathbf{0}$ or not) and on the choice of generalized-raking method through the function G. There are four cases to consider within the framework of (A.1)-(A.5): that addressed by Deville and Särndal (1992) where the base weights w_i^o are essentially correct ($\beta^o = \mathbf{0}$), the case of general β^o with linear calibration ($G = G_{lin}$), the case of general β^o with raking ($G = G_{rak}$), and the case of general β^o with all other (convex) G.

We present formal theoretical results on consistent asymptotic variance estimation only in the context of Poisson sampling. We must do that because the asymptotic variance B in (A.3.1) is not parametrically restricted and is generally accessible only through a Horvitz-Thompson style variance estimator. Under suitable conditions, including restrictions on joint

inclusion probabilities, it is accessible through the design-based unbiased variance estimators using joint inclusion probabilities or through replication-based variance formulas (Balanced Repeated Replication as in Wolter 2007 including Successive Difference Replication as in Fay and Train 1996), with theoretical justification – as far as it goes – along the lines of Krewski and Rao (1981).

Theorem 2 *Assume (A.1)-(A.5), and assume in addition that sampling is Poisson (and nonresponse subsumed in base-weights is based on independent unit-level decisions) in the sense that the random variables $R_i = I_{[i\in\mathcal{S}]}$ are independent random variables for all $i \in \mathcal{U}$, with correct weights and base-weights restricted by:*

$$E(R_i) = \pi_i^* = 1/w_i^* \quad , \qquad 0 < c_3 \leq w_i^o \leq C_3 < \infty \tag{34}$$

Assume also that

(A.6) *As $n, N \to \infty$, the limit $lim_{n\to\infty}\,(n/N^2)\,\sum_{i\in\mathcal{U}}\,(w_i^*-1)\begin{pmatrix} Y_i-\bar{Y} \\ \mathbf{X}_i-\bar{\mathbf{X}}\end{pmatrix}^{\otimes 2}$ exists.*

Then the limiting variance matrix B in (A.3.1) is equal to the limit in (A.6), and a consistent estimate of it based on survey data is

$$\hat{B} \;=\; \frac{n}{N^2}\,\sum_{i\in\mathcal{S}}\,\hat{w}_i\,(\hat{w}_i-1)\begin{pmatrix} Y_i-\hat{\bar{Y}} \\ \mathbf{X}_i-\hat{\bar{\mathbf{X}}}\end{pmatrix}^{\otimes 2}$$

where $\hat{\bar{Y}} = N^{-1}\sum_{i\in\mathcal{S}}\,\hat{w}_i\,Y_i \quad , \qquad \hat{\bar{\mathbf{X}}} = N^{-1}\sum_{i\in\mathcal{S}}\,\hat{w}_i\,\mathbf{X}_i.$

The estimator

$$\hat{V} \;=\; \frac{n}{N^2}\,\sum_{i\in\mathcal{S}}\,\hat{w}_i\,(\hat{w}_i^o-1)\,\big(Y_i-\hat{\bar{Y}}-\hat{b}'\,(\mathbf{X}_i-\hat{\bar{\mathbf{X}}})\big)^{\mathbf{2}} \tag{35}$$

is consistent for the asymptotic variance of $(\sqrt{n}/N)\,\big(\sum_{i\in\mathcal{S}}\,\hat{w}_i\,Y_i\,-\,\bar{Y}\big)$, but the estimator $\hat{b}$ in (35) for $M^{-1}\mathbf{m}$ from Theorem 1 takes somewhat different forms in the following cases.

(i). $\beta^o = \mathbf{0}$, then $(G')^{-1}(\mathbf{X}_i'\beta^o) = 1 = \kappa(\mathbf{X}_i'\beta^o)$, and $\max_i\,|\hat{w}_i - w_i^o| = o_P(1)$, so $\hat{w}_i$ can be replaced by w_i^o in $\hat{V}$ and $\hat{\bar{Y}}$, $\hat{\bar{\mathbf{X}}}$, and

$$\hat{b} \;=\; \big(\sum_{i\in\mathcal{S}}\,w_i^o\,\mathbf{X}_i^{\otimes 2}\big)^{-1}\,\sum_{i\in\mathcal{S}}\,w_i^o\,\mathbf{X}_i\,(Y_i-\hat{\bar{Y}})$$

(ii). β^o is general and $G \equiv G_{lin}$, then $(G')^{-1}(\mathbf{X}_i'\beta^o) \;=\; (1+\mathbf{X}_i'\beta^o)$, $\kappa(\mathbf{X}_i'\beta^o) = 1$, and

$$\hat{b} \;=\; \big(\sum_{i\in\mathcal{S}}\,w_i^o\,\mathbf{X}_i^{\otimes 2}\big)^{-1}\,\sum_{i\in\mathcal{S}}\,w_i^o\,\mathbf{X}_i\,(Y_i-\hat{\bar{Y}})$$

(iii). β^o is general and $G \equiv G_{rak}$, then $(G')^{-1}(\mathbf{X}_i'\beta^o) = \kappa(\mathbf{X}_i'\beta^o)$, so $\gamma(\mathbf{X}_i'\beta_i^o) \equiv 1$, and

$$\hat{b} \;=\; \big(\sum_{i\in\mathcal{S}} \hat{w}_i\, \mathbf{X}_i^{\otimes 2}\big)^{-1} \;\sum_{i\in\mathcal{S}} \hat{w}_i\, \mathbf{X}_i\, (Y_i - \hat{\bar{Y}})$$

(iv). β^o and G are general, and (recalling $\gamma(z)$ from (23)),

$$\hat{b} \;=\; \big(\sum_{i\in\mathcal{S}} \hat{w}_i\, \gamma(\mathbf{X}_i'\hat{\beta})\, \mathbf{X}_i^{\otimes 2}\big)^{-1} \;\sum_{i\in\mathcal{S}} \hat{w}_i\, \gamma(\mathbf{X}_i'\hat{\beta})\, \mathbf{X}_i\, (Y_i - \hat{\bar{Y}})$$

Proof. Consider the sum

$$B^* \;=\; \frac{n}{N^2} \sum_{i=1}^{N} R_i\, w_i^*\, (w_i^* - 1) \begin{pmatrix} Y_i - \bar{Y} \\ \mathbf{X}_i - \bar{\mathbf{X}} \end{pmatrix}^{\otimes 2}$$

Then $E(B^*)$ is equal to the design variance of $(\sqrt{n}/N)\,(Y_i - \bar{Y}, (\hat{\bar{\mathbf{X}}} - \bar{\mathbf{X}})')'$, and also to the matrix whose limit is assumed to exist in (A.6), and (since $\pi_i^* w_i^* = 1$), for any $a \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^{p+1}$,

$$\mathrm{Var}\Big(\begin{pmatrix} a \\ \mathbf{v} \end{pmatrix}' B^* \begin{pmatrix} a \\ \mathbf{v} \end{pmatrix}\Big) \;=\; \frac{n^2}{N^4} \sum_{i=1}^{N} (w_i^* - 1)^3\, \big(\big(a \cdot (Y_i - \bar{Y} + \mathbf{v}'(\mathbf{X}_i - \bar{\mathbf{X}})\big)^4 \;=\; O_P(1/n)$$

where the last order-of-magnitude estimate follows from the uniform boundedness of $Y_i, \mathbf{X}_i$ and the bound $w_i^* \le C'\, C_1\, N/n$ from (A.2), (A.4), and (24). Thus as $n, N \to \infty$, by Chebychev's inequality

$$B^* - \frac{n}{N^2} \sum_{i=1}^{N} (w_i^* - 1) \begin{pmatrix} Y_i - \bar{Y} \\ \mathbf{X}_i - \bar{\mathbf{X}} \end{pmatrix}^{\otimes 2} \;\xrightarrow{P}\; \mathbf{0}$$

Moreover, by similar reasoning,

$$\hat{\bar{Y}} \;\xrightarrow{P}\; \bar{Y} \;\;,\quad \hat{\bar{\mathbf{X}}} \;\xrightarrow{\mathbf{P}}\; \bar{\mathbf{X}}$$

and we knew from Theorem 1 that

$$\sup_{i\in\mathcal{U}} \Big|\, \exp(\mathbf{X}_i'\hat{\beta}) \,-\, \exp(\mathbf{X_i}'\beta^o) \Big| \;\xrightarrow{P}\; 0$$

This implies that the n summands of $\hat{B}$ and B^* over $i \in \mathcal{S}$ differ termwise uniformly by $(N/n)^2$ multiplied by a term converging to 0, so that for each $a, \mathbf{v}$,

$$\Big|\begin{pmatrix} a \\ \mathbf{v} \end{pmatrix}'(\hat{B} - B^*)\begin{pmatrix} a \\ \mathbf{v} \end{pmatrix}\Big| \;\le\; o_P(n\, \frac{n}{N^2}\,(N/n)^2) \;=\; o_P(1)$$

It follows that B in (A.3.1) is equal to the limiting matrix in (A.6), and $\hat{B}$ is a consistent estimator for B.

The rest of the proof is done only for case (iv): cases (i)-(iii) are special instances of case (iv). By (32) in Theorem 1), we are estimating the asymptotic variance of

$$\begin{pmatrix} 1 \\ -M^{-1}\,\mathbf{m} \end{pmatrix}' \frac{\sqrt{n}}{N} \sum_{i\in\mathcal{S}} w_i^* \begin{pmatrix} Y_i - \bar{Y} \\ \mathbf{X}_i - \bar{\mathbf{X}} \end{pmatrix}$$

By (33) in Theorem 1, the asymptotic variance is

$$\begin{pmatrix} 1 \\ -M^{-1}\mathbf{m} \end{pmatrix}' B \begin{pmatrix} 1 \\ -M^{-1}\mathbf{m} \end{pmatrix}$$

so the Theorem is proved if we establish $\hat{b}$ in (iv) as as consistent estimator of $M^{-1}\,\mathbf{m}$. However, by the limiting relations (A.3.2) and the boundedness of the function $\gamma(\mathbf{x}'\beta^o)$ on the supports of $\nu(\cdot)$ and $\gamma(\cdot)$,

$$\mathbf{m} \;=\; \text{P-lim}_{n\to\infty}\, N^{-1} \sum_{i\in\mathcal{S}} w_i^*\,(Y_i - \bar{Y})\,\gamma(\mathbf{X}_i'\beta)\,\mathbf{X}_i \;\;, \quad M \;=\; \text{P-lim}_{n\to\infty}\, N^{-1} \sum_{i\in\mathcal{S}} w_i^*\,\gamma(\mathbf{X}_i'\beta)\,\mathbf{X}_i^{\otimes 2}$$

Again recognizing that $\max_i\,|\hat{w}_i - w_i^o| = o_P(N/n)$ and $\max_i\,\gamma(\mathbf{X}_i'\hat{\beta}) - \gamma(\mathbf{X}_i'\beta^o)| = o_P(1)$, we conclude that $\hat{b} = \hat{M}^{-1}\,\hat{\mathbf{m}}$ where the consistent estimators $\hat{M}$ for M and $\hat{\mathbf{m}}$ for $\mathbf{m}$ are given by

$$\hat{M} \;=\; N^{-1} \sum_{i\in\mathcal{S}} \hat{w}_i\,\gamma(\mathbf{X}_i'\hat{\beta})\,\mathbf{X}_i^{\otimes 2} \quad , \qquad \hat{\mathbf{m}} \;=\; N^{-1} \sum_{i\in\mathcal{S}} \hat{w}_i\,(Y_i - \bar{Y})\,\gamma(\mathbf{X}_i'\hat{\beta})\,\mathbf{X}_i$$

and the Theorem is proved. □

Theorems 1 and 2 under case (i) are the main results of Deville and Särndal (1992), re-proved here with slightly more attention than those authors gave to arguments based on the uniform law of large numbers derived from (A.3.2).

It is particularly noteworthy in Theorem 2 that the variance estimator $\hat{V}$ can be obtained as a GREG variance by using $\hat{w}_i$ in place of the base weights. Although results on variances are presented here only in the context of Poisson sampling, they will generalize to other contexts – such as hierarchical stratified and clustered sampling designs with many independently or SRS sampled final clusters – whenever design-based estimates replacing $\hat{B}$ for variance matrices B are provably consistent. In such settings, applying the variance formula with weights $\hat{w}_i$ to GREG residuals from $\hat{w}_i$ weighted regresssion of Y_i on $\mathbf{Z}_i$ will also work.

Up to this point, our theoretical results concern only surveys in which an asymptotically correct system $\{w_i^*\}_{i=1}^N$ of weights (as defined in (A.3)) has ratios w_i^*/w_i^o satisfying a finite-dimensional parametric condition (A.4). Such a parametric-model condition essentially requires that design weights be calibrated to be correct using finitely many calibration

constraints with a termwise metric for weight changes based on a function G. However, such a condition is a lot to ask. Can anything be said about calibrated weights, or of survey estimators based on them, when the condition holds only inexactly or not at all ? That is the subject of the next Theorems.

For simplicity, we restrict attention in what follows to $G(z) \equiv G_{rak}(x) \equiv x\log(x) - x + 1$, so that $(G')^{-1}(z) = \kappa(z) = e^z$, $\gamma(z) \equiv 1$. Assumptions (A.1) and (A.5) are no longer needed, but Assumption (A.3) again defines an assumed 'asymptotically correct' set of weights w_i^*, but it is assumed that the design weights satisfy a limiting property.

(A.4') There exist a nondegenerate finite probability measure $\tilde{\nu}$ and signed measure $\tilde{\eta}$ (both necessarily supported on $\{1\} \times [-C, C]^p \subset \mathbb{R}^{p+1}$) such that as n, $N \to \infty$,

$$N^{-1} \sum_{i \in \mathcal{S}} w_i^o \, h(\mathbf{X}_i) \xrightarrow{P} \int h(\mathbf{z}) \, \tilde{\nu}(d\mathbf{z}) \quad , \quad N^{-1} \sum_{i \in \mathcal{S}} w_i^o \, h(\mathbf{X}_i) \, Y_i \xrightarrow{P} \int h(\mathbf{z}) \, \tilde{\eta}(d\mathbf{z}) \tag{A.4.1}$$

and

$$\tilde{A} \equiv \int \mathbf{z}^{\otimes 2} \, \tilde{\nu}(d\mathbf{z}) \qquad \text{is positive-definite} \tag{A.4.2}$$

and

$$\text{there exists} \quad \beta_* \in \mathbb{R}^{p+1} : \qquad \int \mathbf{z} \, e^{\mathbf{z}' \beta_*} \, \tilde{\nu}(d\mathbf{z}) = \bar{X} \tag{A.4.3}$$

$$\sqrt{n} \left[N^{-1} \sum_{i \in \mathcal{S}} w_i^o \, e^{\mathbf{X}_i' \beta_*} \begin{pmatrix} Y_i \\ \mathbf{X}_i \end{pmatrix} - \int \mathbf{z} \, e^{\mathbf{z}' \beta_*} \begin{pmatrix} \tilde{\eta}(d\mathbf{z}) \\ \tilde{\nu}(d\mathbf{z}) \end{pmatrix} \right] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \, \tilde{B}) \tag{A.4.4}$$

These conditions are only slightly more restrictive than (A.3), although (A.4.1)–(A.4.2) follow immediately from (A.3) if $\log(w_i^*/w_i^o)$ is a bounded function $g^\dagger(\mathbf{X}_i, U_i)$ of the variables $\mathbf{X}_i$ along with some independent *iid* sequence U_i. We continue to work with the function $K(\cdot)$ defined by (25).

Lemma 2 *Assume (A.2)-(A.3) and (A.4′), and let $G \equiv G_{rak}$. Then (A.5) holds with β_* in place of β^o, and there exists $c > 0$ such that with probability converging to 1 as $n \to \infty$, the symmetric $p \times p$ Jacobian matrix $\nabla_\beta^{tr} K(\beta)$ with (j,k) entry $\partial K_j(\beta)/\partial \beta_k$ (for $0 \le j, k \le p$) has smallest eigenvalue $\ge c$. Moreover, with probability approaching 1 as $n \to \infty$, the function $K(\beta)$ defined in (25) has a unique root $\hat{\beta}_* \in B(r_1, \beta_*)$ and $\hat{\beta}_* - \beta_* = O_P(n^{-1/2})$.*

Proof. The steps are very much like those of Lemma 1, but with Taylor expansions taken around the base-point β_* in place of β^o. (A.5) and (24) are immediate by choice of G with $(G')^{-1}(z) = e^z$, with β_* replacing β^o. Next, $K(\beta_*) = O_P(1/\sqrt{n})$ by (A.4.4) within (A.4′),

and the first-order Taylor expansion of $K(\beta)$ around β_* leads exactly as in Lemma 1 (now with $\kappa(z)/(G')^{-1}(z) \equiv 1$ and $\tilde{\nu}$ replacing ν) to

$$\sup_{\beta \in B(r_1, \beta_*)} \left| \nabla_\beta^{tr} K(\beta) \, - \, \int \mathbf{x}^{\otimes 2} \, \tilde{\nu}(d\mathbf{x}) \right| \, \xrightarrow{P} \, 0 \tag{36}$$

By the Mean Value Theorem, with $\beta \in B(r_1, \beta_*)$ and $\tilde{\beta}$ on the line from β_* to β,

$$K(\beta) \, = \, K(\beta_*) \, + \, \Big[\big\{ \nabla^{tr} K(\tilde{\beta}) \, - \, \int \mathbf{z}^{\otimes 2} \, \tilde{\nu}(d\mathbf{z}) \big\} \, + \, \int \mathbf{z}^{\otimes 2} \, \tilde{\nu}(d\mathbf{z}) \big) \Big] \, (\beta - \beta_*) \tag{37}$$

In the last displayed equation, the bracketed matrix multiplying $\beta - \beta_*$ has minimum eigenvalue uniformly bounded $\geq (1 - \delta) \, \lambda_* \, c_2 / C_1$ for $\beta \in B(r_1, \beta_*)$, for arbitrary $\delta > 0$, where λ_* is the smallest eigenvalue of $\tilde{A}$ in (A.4.2). Therefore, the displayed equation implies for arbitrary $\epsilon > 0$ there is a large constant C^* such that for large enough n with probability $\geq 1 - \epsilon$, $\|K(\beta_*)\| \leq C^* / \sqrt{n}$ and

$$\inf \big\{ \|K(\beta)\| : \; (2 \, C_1 \, C^*) / (c_2 \, \lambda_* \, \sqrt{n}) \leq \|\beta - \beta_*\| \leq r_1 \big\} \; \geq \; 2C^* / \sqrt{n}$$

Therefore, with high probability $\|K(\beta)\|$ has a minimum on $B(r_1, \beta_*)$ that lies within a ball of radius $2C_2 \, C^* / (c_3 \, \lambda_* \, \sqrt{n})$ around β_*. By the Implicit Function Theorem, this minimum is unique and is actually a root of $K(\beta)$. The Lemma is proved. □

We next prove analogues of Theorems 1 and 2 in the present setting, again by essentially the same proofs with small modifications.

Theorem 3 *Assume (A.2)-(A.3), (A.4′) and $G \equiv G_{rak}$, and define*

$$\tilde{M} \, = \, \tilde{A} \, = \, \int \mathbf{z}^{\otimes 2} \, \tilde{\nu}(d\mathbf{z}) \quad , \qquad \mathbf{m} \, \equiv \, \int \mathbf{z} \, \tilde{\eta}(d\mathbf{z})$$

and denote by $\tilde{B}_{22}$ the lower-right $(p+1) \times (p+1)$ sub-matrix of $\tilde{B}$. Then as $n, N \to \infty$,

$$\sqrt{n} \, \big(\hat{\beta} \, - \, \beta_* \big) \, + \, \tilde{M}^{-1} \, \sqrt{n} \, K(\beta_*) \, \xrightarrow{P} \, 0 \tag{38}$$

$$\sqrt{n} \, \big(\hat{\beta} \, - \, \beta_* \big) \, \xrightarrow{\mathcal{D}} \, \mathcal{N}\Big(0, \; \tilde{M}^{-1} \, B_{22} \, \tilde{M}^{-1} \Big) \tag{39}$$

and with $\hat{w}_i \equiv \exp(\hat{\beta}' \, \mathbf{X}_i)$,

$$\sqrt{n} \, \Big[\, \frac{1}{N} \, \sum_{i \in \mathcal{S}} \, \Big(\hat{w}_i \, - \, w_i^o \, e^{\beta_*' \mathbf{X}_i} \Big) \, Y_i \, - \, \tilde{\mathbf{m}}' \, \tilde{M}^{-1} \, K(\beta_*) \Big] \, \xrightarrow{P} \, 0 \tag{40}$$

$$\sqrt{n} \, \big(\frac{1}{N} \, \sum_{i \in \mathcal{S}} \, \hat{w}_i \, Y_i \, - \, \int e^{\mathbf{z}' \beta_*} \, \tilde{\eta}(d\mathbf{z}) \big) \, \xrightarrow{\mathcal{D}} \, \mathcal{N}\Big(0 \, , \, \begin{pmatrix} 1 \\ -\tilde{M}^{-1} \, \tilde{\mathbf{m}} \end{pmatrix}^{tr} \tilde{B} \, \begin{pmatrix} 1 \\ -\tilde{M}^{-1} \, \tilde{\mathbf{m}} \end{pmatrix} \Big) \tag{41}$$

Proof. As in the proof of Theorem 1, equation (37) with $\hat{\beta}$ replacing β implies (38) by using (36), and (39) follows immediately because $\sqrt{n}\,K(\beta_*)$ is the subvector of coordinates 2 thropugh $p+2$ on the left-hand side of (A.4.4).

First-order Taylor expansion of the summands of $N^{-1}\sum_{i\in\mathcal{S}}\hat{w}_i\,Y_i - \int e^{\mathbf{z}'\beta_*}\,\tilde{\eta}(d\mathbf{z})$ using $\hat{w}_i = w_i^o\cdot(G')^{-1}(\hat{\beta}'\mathbf{X}_i)$ yields

$$\frac{1}{N}\sum_{i\in\mathcal{S}}\hat{w}_i\,Y_i - \int e^{\mathbf{z}'\beta_*}\,\tilde{\eta}(d\mathbf{z}) \;=\; \frac{1}{N}\sum_{i\in\mathcal{S}} w_i^o\,e^{\beta_*'\mathbf{X}_i}\,Y_i - \int e^{\mathbf{z}'\beta_*}\,\tilde{\eta}(d\mathbf{z}) + \frac{1}{N}\sum_{i\in\mathcal{S}} w_i^o\left(e^{\hat{\beta}'\mathbf{X}_i}-e^{\hat{\beta}'\mathbf{X}_i}\right)Y_i$$

$$= \;\frac{1}{\sqrt{n}}\left\{T_n \;+\; \frac{1}{N}\sum_{i\in\mathcal{S}} w_i^o\,e^{\beta_*'\mathbf{X}_i}\,Y_i\,\mathbf{X}_i'\,\sqrt{n}(\hat{\beta}-\beta_*) \;+\; o_P(1)\right\}$$

where T_n is the first coordinate of the left-hand side of (A.4.4). Therefore, by (38),

$$\frac{\sqrt{n}}{N}\sum_{i\in\mathcal{S}}\hat{w}_iY_i - \int e^{\mathbf{z}'\beta_*}\,\tilde{\eta}(d\mathbf{z}) \;=\; T_n - \Big(\frac{1}{N}\sum_{i\in\mathcal{S}} w_i^o\,e^{\beta_*'\mathbf{X}_i}\,Y_i\,\mathbf{X}_i\Big)'\,\tilde{M}^{-1}\,\sqrt{n}\,K(\beta_*) + o_P(1) \quad (42)$$

Now by (A.4.4), together with (A.4′) and the definition of $\tilde{\mathbf{m}}$,

$$\frac{\sqrt{n}}{N}\sum_{i\in\mathcal{S}}\hat{w}_i\,Y_i \;-\; \frac{\sqrt{n}}{N}\sum_{i\in\mathcal{S}} w_i^o\,e^{\beta_*'\,\mathbf{X}_i}\,Y_i \;+\; \tilde{\mathbf{m}}'\,\tilde{M}^{-1}\,\sqrt{n}\,K(\beta_*) \;=\; o_P(1)$$

which is the same statement as (40). Finally, the right-hand side of (42) is

$$\begin{pmatrix} 1 \\ -\tilde{M}\,\tilde{\mathbf{m}}\end{pmatrix}'\begin{pmatrix} T_n \\ \sqrt{n}\,K(\beta_*)\end{pmatrix} \;+\; o_P(1) \;\xrightarrow{\;\mathcal{D}\;}\; \mathcal{N}\Big(0\,,\;\begin{pmatrix} 1 \\ -\tilde{M}^{-1}\,\tilde{\mathbf{m}}\end{pmatrix}^{tr}\tilde{B}\begin{pmatrix} 1 \\ -\tilde{M}^{-1}\,\tilde{\mathbf{m}}\end{pmatrix}\Big)$$

where the convergence in distribution follows immediately from (A.4.4). □

Because the parametric model for weight-ratios w_i^*/w_i^o assumsd in (A.4) is misspecified but is still the working model in the setting of Theorem 3, the large-sample limit of the raked survey estimator $N^{-1}\sum_{i\in\mathcal{S}}\hat{w}_i\,Y_i$ has no generally predictable relation to the target $\bar{Y}$ of estimation. Nevertheless, as is true also in the misspecified-model theory of M-estimators, there is a robustified form of estimator for the variance of the raking-calibrated survey estimator that can under some circumstances be consistent.

Theorem 4 *Assume (A.2)-(A.3), (A.4′), and $G\equiv G_{rak}$, and assume in addition that sampling is Poisson and the 'quasirandomization' model for nonresponse holds, i.e., the random variables $R_i = I_{[i\in\mathcal{S}]}$ are independent random variables for all $i\in\mathcal{U}$, with correct weights and base-weights restricted by:*

$$E(R_i) = \pi_i^* = 1/w_i^* \quad , \qquad 0 < c_3 \le w_i^o \le C_3 < \infty \quad (43)$$

Assume also that

(A.6′) $B^\circ \equiv lim_{n\to\infty}\,(n/N^2)\,\sum_{i\in\mathcal{U}}\,\left(e^{\beta'_*\mathbf{X}_i}\,w_i^o/w_i^*\right)^2\,(w_i^*-1)\left(\begin{array}{c} Y_i-\int\,e^{\beta'_*\mathbf{z}}\,\tilde{\eta}(d\mathbf{z}) \\ \mathbf{X}_i-\int\,e^{\beta'_*\mathbf{z}}\,\mathbf{z}\,\tilde{\nu}(d\mathbf{z}) \end{array}\right)^{\otimes 2}$ *exists.*

Then the limiting variance matrix $\tilde{B}$ *in (A.4.4) is equal to the limit in (A.6′), and the asymptotic variance (multiplied by n) of the survey estimator* $N^{-1}\,\sum_{i\in\mathcal{S}}\,\hat{w}_i\,Y_i$ *is*

$$\left(\begin{array}{c} 1 \\ -\tilde{M}^{-1}\,\tilde{\mathbf{m}} \end{array}\right)'\,B^\circ\left(\begin{array}{c} 1 \\ -\tilde{M}^{-1}\,\tilde{\mathbf{m}} \end{array}\right) \tag{44}$$

where

$$\int\,e^{\beta'_*\mathbf{z}}\,\tilde{\eta}(d\mathbf{z})\;=\;P\text{-}lim_{n\to\infty}\,N^{-1}\sum_{i\in\mathcal{S}}\hat{w}_i\,Y_i\;\;,\quad\int\,e^{\beta'_*\mathbf{z}}\,\mathbf{z}\,\tilde{\nu}(d\mathbf{z})\;=\;P\text{-}lim_{n\to\infty}\,N^{-1}\sum_{i\in\mathcal{S}}\hat{w}_i\,\mathbf{X}_i$$

$$\tilde{M}\;=\;P\text{-}lim_{n\to\infty}\,N^{-1}\sum_{i\in\mathcal{S}}\hat{w}_i\,\mathbf{X}_i^{\otimes 2}\;\;,\qquad\tilde{\mathbf{m}}\;=\;P\text{-}lim_{n\to\infty}\,N^{-1}\sum_{i\in\mathcal{S}}\hat{w}_i\,\mathbf{X}_i\,Y_i$$

This Theorem is proved with steps very similar to those used in Theorem 2. Consistency of estimation variance (44) will generally not be possible in Theorem 4 unless there is some auxiliary source enabling consistent estimation of the true weights w_i^*. There would not need to be any established rate for that consistency, which leaves open the possiblility that in some settings a very highly parameterized 'machine-learning' type model might be used. Such consistent estimation of weights would generally be sufficient to establish consistency of estimation of survey-weighted averages but not asymptotic normality of those estimates, but in the context of Theorem 4 consistently estimated weights would enable consistent estimation of the variances of the survey-weighted estimates.

C Discussion

The usual caveats about Missing at Random (MAR) or other assumptions enabling valid inference in the presence of nonresponse have been avoided here through (A.3)-(A.4), in which the design-based 'asymptotic correctness' of weights applied to survey estimation of (functions of $\mathbf{X}_i$ and) Y_i is assumed. MAR or a similar assumption about conditional independence of Y_i and response-indicator R_i given $\mathbf{X}_i$ would be the usual way to assume the correctness of weighting to compensate for nonresponse, when not only R_i but also $\mathbf{X}_i, Y_i$ are random. But we are not aware of any published design-based presentation of assumptions guaranteeing design-consistent survey inferences after calibration in the presence of nonresponse. In that sense, the assumption-set presented here contributes toward a design-based understanding of missing data due to nonresponse.

Appendix References

Deville, J. and Särndal, C.-E. (1992), Calibration Estimators in Survey Sampling, *Jour. Amer. Statist. Assoc.* **87**, 376-382.

Deville, J., Särndal, C.-E. and Sautory, O. (1993), Generalized Raking Procedures in Survey Sampling, *Jour. Amer. Statist. Assoc.* **88**, 1013-1020.

Fay, R. and Train, G. (1995), Aspects of survey and model-based post-censal estimation of income and poverty characteristics for states and counties, *Proc. Amer. Statist. Assoc., Govt. Statist. Section*, 154-159.
`https://cps.ipums.org/cps/resources/repwt/FayTrain95.pdf`

Krewski, D. and Rao, J. (1981), Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods, *Annals of Statistics* **9**, 1010-1019.

Wolter, K. (2007), **Variance Estimation**, Springer.

Pollard, D. (1984), **Convergence of Stochastic Processes**, Springer-Verlag.

Wolter, K. (2007), **Variance Estimation**, 2nd ed., Springer.