Multilevel Calibration Weighting for Survey Data

Eli Ben-Michael[1], Avi Feller[2] and Erin Hartman[3]

[1] *Department of Statistics & Data Science and Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA. E-mail: ebenmichael@cmu.edu*
[2] *Goldman School of Public Policy and Department of Statistics, University of California, Berkeley, Berkeley, CA, USA. E-mail: afeller@berkeley.edu*
[3] *Departments of Political Science and Statistics, University of California, Berkeley, Berkeley, CA, USA. E-mail: ekhartman@berkeley.edu*

Abstract

In the November 2016 U.S. presidential election, many state-level public opinion polls, particularly in the Upper Midwest, incorrectly predicted the winning candidate. One leading explanation for this polling miss is that the precipitous decline in traditional polling response rates led to greater reliance on statistical methods to adjust for the corresponding bias—and that these methods failed to adjust for important interactions between key variables like educational attainment, race, and geographic region. Finding calibration weights that account for important interactions remains challenging with traditional survey methods: raking typically balances the margins alone, while post-stratification, which exactly balances all interactions, is only feasible for a small number of variables. In this paper, we propose multilevel calibration weighting, which enforces tight balance constraints for marginal balance and looser constraints for higher-order interactions. This incorporates some of the benefits of post-stratification while retaining the guarantees of raking. We then correct for the bias due to the relaxed constraints via a flexible outcome model; we call this approach "double regression with post-stratification." We use these tools to re-assess a large-scale survey of voter intention in the 2016 U.S. presidential election, finding meaningful gains from the proposed methods. The approach is available in the `multical` R package.

Keywords: public opinion, survey weighting, calibration, post-stratification

1 Introduction

Given the precipitous decline in response rates for traditional polling approaches and increased reliance on possibly nonrepresentative convenience samples, a pressing statistical question in modern public opinion research is how to find survey weights that appropriately adjust for higher-order interactions between key variables. Traditional approaches, like raking, can perform poorly with even a moderate number of characteristics, typically balancing marginal distributions while failing to balance higher-order interactions. By contrast, post-stratification, which in principle exactly balances all interactions, is only feasible for a small number of variables. And, while approaches like multilevel regression and post-stratification (MRP; Gelman and Little 1997) use outcome modeling to overcome this, they do not produce a single set of survey weights for all outcomes and can lead to unchecked extrapolation away from the data. Fortunately, recent research on modern survey calibration (e.g., Chen, Li, and Wu 2020; Guggemos and Tillé 2010) and on balancing weights for causal inference (e.g., Hirshberg and Wager 2021; Zubizarreta 2015) offer promising paths forward.

Building on these advances, we propose two principled approaches to account for higher-order interactions when estimating population quantities from non-probability samples. First, we propose *multilevel calibration* weighting, which exactly balances the first-order margins and *approximately* balances interactions, prioritizing balance in lower-order interactions over higher-order interactions. Thus, this approach incorporates some of the benefits of post-stratification while retaining the guarantees of the common-in-practice raking approach. And unlike outcome modeling approaches like MRP, multilevel calibration weights are estimated once and applied to all survey outcomes, an important practical constraint in many survey settings.

Political Analysis (2024)
vol. 32: 65–83
DOI: 10.1017/pan.2023.9

Published
31 March 2023

Corresponding author
Avi Feller

Edited by
Jeff Gill

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

In some cases, however, multilevel calibration weighting alone may be insufficient to achieve good covariate balance on all higher-order interactions, possibly leading to bias; or researchers might only be focused on a single outcome of interest. For this, we propose *double regression with post-stratification* (DRP), which combines multilevel calibration weights with outcome modeling, taking advantage of flexible modern prediction methods to estimate and correct for possible bias from imperfect balance. When the weights alone achieve good balance on higher-order interactions, the adjustment from the outcome model is minimal. When the higher-order imbalance is large, however, the bias correction will also be large and the combined estimator will rely more heavily on the outcome model. We characterize the numerical and statistical properties of both multilevel calibration weighting and the combined DRP estimator.

With these tools in hand, we consider the question of the failure of state-level polls in the 2016 U.S. presidential election. Kennedy *et al.* (2018) show that many 2016 surveys failed to accurately account for the shift in public opinion among white voters with no college education, particularly in the Midwestern region of the country, indicating that failing to adjust for the interaction between key variables of education, race, and geographic region resulted in substantial bias. We evaluate whether accounting for this higher-order interaction of race, education level, and geographic region can, retrospectively, improve public opinion estimates in the final publicly available preelection Pew poll. We show that the multilevel weights substantially improve balance in interactions relative to raking and *ad hoc* post-stratification and that further bias correction through DRP can meaningfully improve estimation.

Our proposed approach builds on two important advances in both modern survey methods and in causal inference. First, there has been a renewed push to find calibration weights that allow for *approximate* balance on covariates, rather than exact balance (Guggemos and Tillé 2010; Park and Fuller 2009; Zubizarreta 2015). Second, several recent approaches combine such weights with outcome modeling, extending classical generalized regression estimators in survey sampling and doubly robust estimation in causal inference (e.g., Chen *et al.* 2020; Hirshberg and Wager 2021); we view our proposed DRP approach as a particular implementation of such augmented balancing weights. We give more detailed reviews in Sections 2.2 and 3.2.

The paper proceeds as follows: Section 2 describes the notation and estimands, and formally describes various common survey weighting procedures such as raking and post-stratification. Section 3 characterizes the estimation error for arbitrary weighting estimators to motivate our multilevel calibration procedure, then describes the procedure. Section 4 proposes the DRP estimator and analyzes its numerical and statistical properties. Section 5 uses these procedures in the application. The methods we develop here are available in the `multical` R package.

1.1 2016 U.S. Presidential Election Polling

While national public opinion polls for the November 8, 2016 U.S. presidential election were, on average, some of the most accurate in recent public opinion polling, state-level polls were notable in their failure to accurately predict the winning candidate, particularly in the Upper Midwest. These state-level errors in turn led public opinion researchers to incorrectly predict the winner of the electoral college. Kennedy *et al.* (2018) attribute these errors to three main sources: (1) a late swing among undecided voters toward Trump, (2) failure to account for non-response related to education level, particularly among white voters, and (3) to a lesser degree, failure to properly predict the composition of the electorate.

While all three of these concerns are important for survey practitioners, our analysis focuses on addressing concern (2) by allowing for deep interactions among important covariates, including race, education level, and region. To isolate this concern, we combine two high-quality surveys from before and after the 2016 election. We begin with the October 16, 2016 Pew survey of 2,062

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

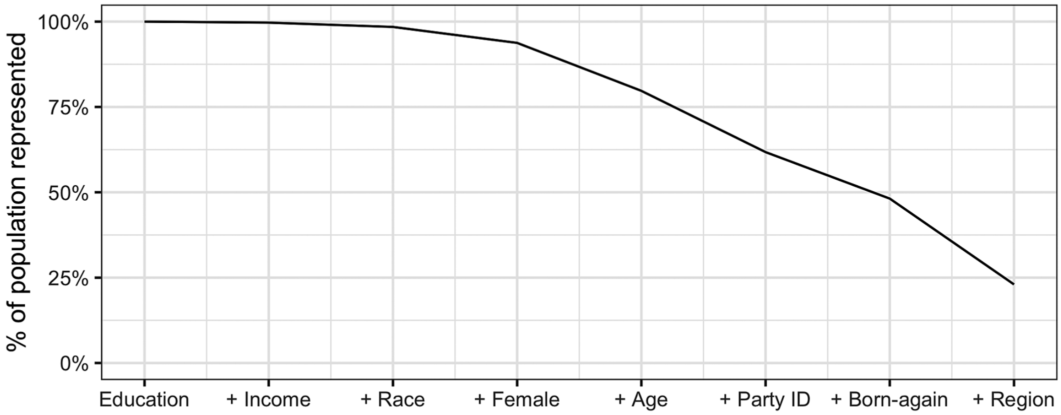


Figure 1. Percentage of the population that is represented in the survey, beginning with education and successively interacting with income, religion, race, a binary for self-reported female, age, party identification, born-again Christian status, and region.

respondents, the final public release of Pew's election polling (Pew Research Center 2016).[1] The primary outcome is respondents' "intent to vote" for each major party. We combine this with the $44,909$ respondents in the 2016 Congressional Cooperative Election Study (CCES) postelection survey, a large, high-quality public opinion survey that accurately captures electoral outcomes at the state level (see Ansolabehere and Schaffner 2017). Here, the primary outcome is respondents' "retrospective" vote for each major party candidate.[2]

The combined Pew and CCES observations form a "population" of size $N = 46,971$, where observations from the Pew survey are coded as respondents and observations from the CCES are coded as nonrespondents. Using this target, rather than the ground truth defined by the actual electoral outcomes, helps to address concern (3) above. Specifically, the CCES validates voters against Secretaries of State voter files, allowing us to use known voters for whom we have measured auxiliary covariates to define our target population.

Our goal is to adjust the respondent sample for possible non-response bias from higher-order interactions and assess whether the adjusted estimates are closer to the ground truth. Figure 1 shows the eight auxiliary variables we consider, measured in both the Pew and CCES surveys. All eight variables are coded as discrete, with the number of possible levels ranging from two to nine.[3] Ideally, we would adjust for all possible interactions of these variables, via post-stratification. This is infeasible, however; there are $12,347$ possible combinations, far greater than the $n = 2,062$ respondents in our survey. Figure 1 shows the percentage of the population that is represented in the survey as we progressively include—and fully interact—more covariates. With a single covariate, education (six levels), each cell has at least one respondent. When including all eight covariates, the non-empty cells in the sample represent less than a quarter of the population. This motivates our search for alternative adjustment methods that account for higher-order interactions in a parsimonious way, prioritizing adjustment for strong interactions.

2 Background and Setup

2.1 Notation and Estimands

We consider a finite population of N individuals indexed $i = 1, \dots, N$. We observe the outcome Y_i for units that respond to the survey. Define a binary variable R_i that denotes inclusion in the

[1] Since our survey is from mid-October, we cannot account for concern (1) above, a late break toward Trump among undecided voters, which may contribute to remaining residual bias.

[2] Data and code to replicate this analysis are available in Ben-Michael, Feller, and Hartman (2023).

[3] These are (i) education (six levels), (ii) income (nine levels), (iii) race (four levels), (iv) a binary for self-reported female (two levels), (v) age (four levels), (vi) party ID (three levels), (vii) born again Christian (two levels), and (viii) region (five levels).

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

survey, where $R_i = 1$ indicates that unit i responds; $n = \sum_i R_i$ is the total number of respondents. This response variable can include respondents to a probability sample or a convenience sample. These response variables R_i are the only random component of our setup, and all expectations, variances, and probability statements will be with respect to the randomness in the response variables. In addition, each individual is also associated with a set of d categorical covariates $X_{i1}, \ldots, X_{id}$, where the ℓth covariate is categorical with J_ℓ levels, so that the vector of covariates is $\boldsymbol{X}_i \in [J_1] \times \cdots \times [J_d]$.[4]

Rather than consider these variables individually, we rewrite the vector X_i as a single categorical covariate, the *cell* for unit i, $S_i \in [J]$, where $J \leq J_1 \times \cdots \times J_d$ is the number of unique levels in the target population.[5] While we primarily consider a fixed population size N and total number of distinct cells J, in Section C of the Supplementary Material, we extend this setup to an asymptotic framework where both the population size and the number of cells can grow. With these cells, we can summarize the covariate information. We denote $\boldsymbol{N}^{\mathcal{P}} \in \mathbb{N}^J$ as the *population count vector* with $N_s^{\mathcal{P}} = \sum_i \mathbb{1}\{S_i = s\}$, and $\boldsymbol{n}^{\mathcal{R}} \in \mathbb{N}^J$ as the *respondent count vector*, with $n_s^{\mathcal{R}} = \sum_i R_i \mathbb{1}\{S_i = s\}$. While the population count $N_s^{\mathcal{P}} > 0$ for every cell s, the respondent count $n_s^{\mathcal{R}}$ may be equal to zero. We assume that we have access to these cell counts for both the respondent sample and the population.

Finally, for each cell s, we consider a set of binary vectors $\boldsymbol{D}_s^{(k)}$ that denote the cell in terms of its kth order interactions, and collect the vectors into matrices $\boldsymbol{D}^{(k)} = [\boldsymbol{D}_1^{(k)} \ldots \boldsymbol{D}_J^{(k)}]'$, and into one combined $J \times J$ *design* matrix $\boldsymbol{D} = [\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(d)}]$. This is the usual design matrix for interaction terms in linear models, with the rows corresponding to unique cells, rather than units. It can be constructed using the `model.matrix` command in the `R` programming language. Figure A.1 in the Supplementary Material shows an example of $\boldsymbol{D}$ with three covariates from our running example: a binary for self-reported female, discretized age, and party identification.

Our goal is to estimate the *population* average outcome, which we can write as a cell-size weighted average of the within-cell averages, that is,

$$\mu \equiv \frac{1}{N}\sum_{i=1}^{N} Y_i = \sum_{s=1}^{J} \frac{N_s^{\mathcal{P}}}{N}\mu_s, \quad \text{where} \quad \mu_s \equiv \frac{1}{N_s^{\mathcal{P}}}\sum_{S_i=s} Y_i. \tag{1}$$

To estimate the population average, we rely on the average outcomes we observe within each cell. For cell s, the responder average is

$$\bar{Y}_s \equiv \frac{1}{n_s^{\mathcal{R}}}\sum_{S_i=s} R_i Y_i. \tag{2}$$

We invoke the assumption of missingness at random within cells, so that the cell responder averages are unbiased for the true cell averages (Rubin 1976).

ASSUMPTION 1 (Missing at random within cells) *For all cells* $s = 1, \ldots, J$, $\mathbb{E}\left[\bar{Y}_s \mid n_s^{\mathcal{R}}\right] = \mu_s$.

We denote the *propensity score* as $P(R_i = 1) \equiv \pi_i$ and the probability of responding conditional on being in cell s as $\pi(s) \equiv \frac{1}{N_s^{\mathcal{P}}}\sum_{S_i=s}\pi_i$. For a probability sample, π_i denotes the joint probability of both selection into the survey and responding. The analyst knows and controls the selection probabilities but does not know the probability of response given selection. For a convenience

[4] We focus on categorical covariates because it is common to only have population counts at this level, and continuous covariates are often coarsened. However, our method can be adapted for continuous covariates by incorporating more structure. For example, by considering a polynomial basis expansion to include higher-order moments, both marginally and jointly for interactions.

[5] Note that while there are at most $J_1 \times \cdots \times J_d$ levels, some may never appear in the target population and so we drop them from the definition of the cell.

sample, π_i is the unknown probability of inclusion in the sample. For both cases, the overall propensity score π_i is unknown. We assume that this probability is nonzero for all cells.

Assumption 2. *$\pi(s) > 0$ for all $s = 1, \dots, J$.*

These assumptions allow us to identify the overall population average using only the observed data. However, in order for Assumption 1 to be plausible, we will need the cells to be very fine-grained and have covariates that are quite predictive of nonresponse and the outcome. As we will see below, this creates a trade-off between identification and estimation: the former becomes more plausible with more fine-grained information, whereas the latter becomes more difficult (see also D'Amour *et al.* 2020).

2.2 Review: Raking and Post-Stratification

We first consider estimating μ by taking a weighted average of the respondents' outcomes, with weights $\hat{\gamma}_i$ for unit i. Because the cells are the most fine-grained information we have, we will restrict the weights to be constant within each cell. Researchers typically face a trade-off, discussed below, between feasibility of the cell-level estimator, known as post-stratification, and imposing modeling assumptions to address sparsity concerns, such as through calibration weighting. Our goal is to parsimoniously balance the feasibility advantage of calibration with the nonparametric bias-reduction of post-stratification. In particular, we leverage the fact that first-order coefficients typically explain more variation in outcome and response models than higher-order interactions (Cox 1984), and so mitigating bias can be accomplished by enforcing balance on these first-order terms and allowing approximate balance on higher-order interactions.

We start by denoting the estimated weight for cell s as $\hat{\gamma}(s)$ and estimate the population average μ via

$$\hat{\mu}(\hat{\gamma}) \equiv \frac{1}{N}\sum_{i=1}^{N} R_i \hat{\gamma}_i Y_i = \frac{1}{N}\sum_{s} n_s^{\mathcal{R}} \hat{\gamma}(s) \bar{Y}_s. \tag{3}$$

If the individual probabilities of responding were known, we could choose to weight cell S by the inverse of the propensity score, $\frac{1}{\pi(s)}$. Since the propensity score $\pi(s)$ is unknown, researchers can estimate $\pi(s)$ via $\hat{\pi}(s)$ and weight cell s by the inverse *estimated* propensity score $\hat{\gamma}(s) = \frac{1}{\hat{\pi}(s)}$.

Post-stratification weights estimate the propensity score as the proportion of the population in cell s that responded, $\frac{n_s^{\mathcal{R}}}{N_s^{\mathcal{P}}}$, leading to $\hat{\gamma}^{\text{ps}}(s) = \frac{N_s^{\mathcal{P}}}{n_s^{\mathcal{R}}}$. Under Assumption 1, these post-stratification weights give an unbiased estimator for the population average μ. However, in practice post-stratification faces feasibility constraints due to sparsity—this estimator is only defined if there is at least one responder within each cell, which is unlikely with even a moderate number of covariates. For example, in Figure 1, cells corresponding to nearly one quarter of the population are empty when post-stratifying on only five of our eight covariates.[6]

An alternative, *calibration*, chooses weights so that the weighted distribution of covariates exactly matches that of the full population, which identifies the population average under a modeling assumption called linear ignorability (Hartman, Hazlett, and Sterbenz 2021). A common implementation, *raking on margins*, matches the *marginal* distribution by solving a convex optimization problem that finds the minimally "disperse" weights that satisfy this balance constraint (Deming and Stephan 1940; Deville and Särndal 1992; Deville, Särndal, and Sautory 1993). For example, raking on the margins ensures that the percentage of female, the percentage

[6] Researchers can partly address this by coarsening covariates, although how to do so is not straightforward while still meeting Assumption 1, and the problem persists with even a moderate number of coarsened covariates. Researchers can also redefine the target population to only include those cells represented in the sample, however, this may greatly change the interpretation of the results depending on the response pattern.

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

of Democrats, Independents, and Republicans, and so forth for our eight covariates all match the target population; however, it would not ensure that the pairwise interactions are matched. Bias is mitigated so long as these higher-order interactions are not important to the outcome or response model. Calibration thus addresses feasibility concerns by focusing on balancing lower-order interactions, which are less likely to contain empty cells. For example, in Figure 1, focusing on three-way interactions and below would retain nearly the whole population.

Specifically, starting with baseline weights $q(s)$ for cell s, calibration finds weights that solve

$$\begin{aligned} \min_{\gamma} \quad & \sum_{s} n_s^{\mathcal{R}} f(\gamma(s), q(s)) \\ \text{subject to} \quad & \sum_{s} \boldsymbol{D}_s^{(1)} n_s^{\mathcal{R}} \gamma(s) = \sum_{s} \boldsymbol{D}_s^{(1)} N_s^{\mathcal{P}} \\ & L \leq \gamma(s) \leq U \quad \forall \ s = 1, \ldots, J, \end{aligned} \tag{4}$$

where the function $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^{+}$, in the first equation, is a measure of dissimilarity. Some choices include the scaled and squared distance $f(\gamma(s), q(s)) = \frac{1}{2q(s)}(\gamma(s) - q(s))^2$ and the KL divergence $f(\gamma(s), q(s)) = \gamma(s) \log \frac{\gamma(s)}{q(s)}$. Common choices for baseline weights, $q(s)$, include probability weights when they are known (e.g., Deville and Särndal 1992), or estimated inverse propensity weights. Below we will use uniform weights $q(s) = \frac{N}{n}$, for which the dissimilarity measures reduce to standard notions of the spread of the weights: the scaled and squared distance becomes the variance of the weights and the KL divergence becomes the entropy of the weights. As we will see below, the variance of the weights is directly related to the variance of the weighting *estimate*, and so is a natural choice of penalty.

The second equation encodes the calibration constraints, or the weighted survey moments that must exactly match the target population. In addition to the dispersion penalty in the objective, in the third equation, we constrain the weights to be between a lower bound $L = 0$ and an upper bound $U = \infty$, which restricts the normalized cell weights, $\frac{1}{N}\gamma(s)n_s^{\mathcal{R}}$ to be in the $J - 1$ simplex. This ensures that the imputed cell averages are in the convex hull of the respondents' values and so do not extrapolate, and that the resulting estimator $\hat{\mu}(\hat{\gamma})$ is between the minimum and maximum sample outcomes.[7]

Note that the general calibration procedure can also include exact constraints on higher-order interaction terms, and including all interactions recovers the post-stratification weights. We use the *raking* and *post-stratification* nomenclature to distinguish between calibration weighting with first-order margins and with all interactions. Several papers have proposed "soft" or "penalized" calibration approaches to relax the exact calibration constraint in Equation (4), allowing for approximate balance in some covariates (see, e.g., Rao and Singh 1997; Park and Fuller 2009; Guggemos and Tillé 2010; Gao, Yang, and Kim 2023). Our multilevel calibration approach below can be seen as adapting the soft calibration approach to full post-stratification.

Before turning to our proposal for balancing higher-order interactions, we briefly describe some additional approaches. Chen, Valliant, and Elliott (2019) and McConville *et al.* (2017) discuss model-assisted calibration approaches which rely on the LASSO for variable selection. Caughey and Hartman (2017) select higher-order interactions to balance using the LASSO. Hartman *et al.* (2021) provide a kernel balancing method for matching joint covariate distributions between non-probability samples and a target population. Linzer (2011) provides a latent class model for estimating cell probabilities and marginal effects in highly-stratified data.

Finally, in Section 4, we also discuss outcome modeling strategies as well as approaches that combine calibration weights and outcome modeling. A small number of papers have previously

[7] If we allow unbounded extrapolation and set $L = -\infty$ and $U = \infty$, the resulting estimator will be equivalent to linear regression weights from a linear regression of the outcome on first-order indicators $D^{(1)}$ (Ben-Michael *et al.* 2021).

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

explored this combination for non-probability samples. Closest to our setup is Chen *et al.* (2020), who combine inverse propensity score weights with a linear outcome model and show that the resulting estimator is doubly robust. Related examples include Yang, Kim, and Song (2020), who give high-dimensional results for a related setting; and Si *et al.* (2020), who combine weighting and outcome modeling in a Bayesian framework.

3 Multilevel Calibration: Approximate Post-Stratification

We now propose *multilevel calibration* weights, which bridge the gap between post-stratification and raking on the margins. First, we inspect the finite-sample estimation error and mean squared error (MSE) of the weighting estimator $\hat\mu(\hat\gamma)$ for a set of weights $\hat\gamma$ that are deterministic functions of the cell counts $n^{\mathcal{R}}$, differentiating the impact of imbalance in lower- and higher-order terms on the bias. Importantly, we note that first-order terms typically explain more variation in the outcome and response process than higher-order terms. We then use our decomposition to find weights that control the components of the MSE by *approximately* post-stratifying while maintaining raking on the margins, thus leveraging the advantages of both raking and post-stratification.

3.1 Estimation Error

We begin by inspecting the estimation error $\hat\mu(\gamma) - \mu$ for weights γ. Define the *residual* for unit i as $\varepsilon_i \equiv Y_i - \mu_{S_i}$, and the average respondent residual in cell s as $\bar\varepsilon_s = \frac{1}{n_s^{\mathcal{R}}}\sum_{S_i=s} R_i\varepsilon_i$. The estimation error decomposes into a term due to imbalance in the cell distributions and a term due to idiosyncratic variation within cells:

$$\hat\mu\left(\hat\gamma\right) - \mu = \underbrace{\frac{1}{N}\sum_s\left(n_s^{\mathcal{R}}\hat\gamma(s) - N_s^{\mathcal{P}}\right)\times\mu_s}_{\text{imbalance in cell distribution}} + \underbrace{\frac{1}{N}\sum_s n_s^{\mathcal{R}}\hat\gamma(s)\bar\varepsilon_s}_{\text{idiosyncratic error}}. \tag{5}$$

By Assumption 1, which states that outcomes are missing at random within cells, the idiosyncratic error will be zero on average, and so the bias will be due to imbalance in the cell distribution. By Hölder's inequality, we can see that the MSE, given the number of respondents in each cell, is

$$\begin{aligned}\mathbb{E}\left[\left(\hat\mu\left(\hat\gamma\right) - \mu\right)^2 \mid n^{\mathcal{R}}\right] &= \underbrace{\frac{1}{N^2}\left(\sum_s\left(n_s^{\mathcal{R}}\hat\gamma(s) - N_s^{\mathcal{P}}\right)\mu_s\right)^2}_{\text{bias}^2} + \underbrace{\sum_s\left(\frac{n_s^{\mathcal{R}}}{N}\right)^2\hat\gamma(s)^2\sigma_s^2}_{\text{variance}} \\ &\leq \frac{1}{N^2}\sum_s\mu_s^2\times\underbrace{\sum_s\left(n_s^{\mathcal{R}}\hat\gamma(s) - N_s^{\mathcal{P}}\right)^2}_{\text{imbalance in cell distribution}} + \sigma^2\underbrace{\sum_s\left(\frac{n_s^{\mathcal{R}}}{N}\right)^2\hat\gamma(s)^2}_{\text{noise}},\end{aligned} \tag{6}$$

where $\sigma_s^2 = \text{Var}(\bar Y_s \mid n^{\mathcal{R}})$ and $\sigma^2 = \max_s \sigma_s^2$. We therefore have two competing objectives if we want to control the MSE for any given realization of our survey. To minimize the bias, we want to find weights that control the imbalance between the true and weighted proportions within each cell. To minimize the variance, we want to find "diffuse" weights so that the sum of the squared weights is small.

The decomposition above holds for imbalance measures across all of the strata, without taking into account their multilevel structure. In practice, we expect cells that share features to have similar outcomes on average. We can therefore have finer-grained control by leveraging our representation of the cells into their first-order marginals $\boldsymbol{D}_s^{(1)}$ and interactions of order

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

k, $\boldsymbol{D}_s^{(k)}$. To do this, consider the infeasible population regression using the $\boldsymbol{D}_s^{(k)}$ representation as regressors,

$$\min_{\boldsymbol{\eta}} \sum_{i=1}^{N} \left(Y_i - \sum_{k=1}^{d} \boldsymbol{D}_{S_i}^{(k)} \cdot \boldsymbol{\eta}_k \right)^2 . \tag{7}$$

With the solution to this regression, $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^*, \ldots, \boldsymbol{\eta}_d^*)$, we can decompose the population average in cell s based on the interactions between the covariates, $\mu_s = \sum_{k=1}^{d} \boldsymbol{D}_s^{(k)} \cdot \boldsymbol{\eta}_k^*$.[8]

This decomposition in terms of the multilevel structure allows us to understand the role of imbalance in lower- and higher-order interactions on the bias. Plugging this decomposition into Equation (6), we see that the bias term in the conditional MSE further decomposes into the level of imbalance for the kth-order interactions weighted by the strength of the interactions:

$$\begin{aligned}
\mathbb{E}\left[(\hat{\mu}(\hat{\gamma}) - \mu)^2 \mid n^{\mathcal{R}}\right] &= \frac{1}{N^2} \left(\sum_{k=1}^{d} \boldsymbol{\eta}_k^* \cdot \sum_s \left(n_s^{\mathcal{R}} \hat{\gamma}(s) - N_s^{\mathcal{P}} \right) \boldsymbol{D}_s^{(k)} \right)^2 + \sum_s \left(\frac{n_s^{\mathcal{R}}}{N} \right)^2 \hat{\gamma}(s)^2 \sigma_s^2 \\
&\leq \frac{1}{N^2} \left(\sum_{k=1}^{d} \|\boldsymbol{\eta}_k^*\|_2 \left\| \sum_s \left(n_s^{\mathcal{R}} \hat{\gamma}(s) - N_s \right) \boldsymbol{D}_s^{(k)} \right\|_2 \right)^2 + \sigma^2 \sum_s \left(\frac{n_s^{\mathcal{R}}}{N} \right)^2 \hat{\gamma}(s)^2 .
\end{aligned} \tag{8}$$

Equation (8) formalizes the benefits of raking on the margins as in Equation (4). If there is an additive functional form with no influence from higher-order interaction terms—so $\boldsymbol{\eta}_k^* = 0$ for all $k \geq 2$—then raking will yield an unbiased estimator. Even if the "main effects" are stronger than any of the interaction terms and so the coefficients on the first-order terms, $\|\boldsymbol{\eta}_1^*\|_2$, are large relative to the coefficients for higher-order terms, raking can remove a large amount of the bias and so it is often seen as a good approximation (Mercer, Lau, and Kennedy 2018). However, ignoring higher-order interactions entirely can lead to bias. We therefore propose to find weights that prioritize main effects while still minimizing imbalance in interaction terms when feasible.

3.2 Multilevel Calibration

We now design a convex optimization problem that controls the conditional MSE on the right-hand side of Equation (8). To do this, we apply the ideas and approaches developed for approximate balancing weights (e.g., Hirshberg, Maleki, and Zubizarreta 2019; Ning, Sida, and Imai 2020; Wang and Zubizarreta 2020; Xu and Yang 2022; Zubizarreta 2015) to the problem of controlling for higher-order interactions, using our MSE decomposition as a guide. We find weights that control the imbalance in all interactions in order to control the bias, while penalizing the sum of the squared weights to control the variance. Specifically, we solve the following optimization problem:

$$\begin{aligned}
\min_{\gamma \in \mathbb{R}^J} \quad & \underbrace{\sum_{k=2}^{d} \frac{1}{\lambda_k} \left\| \sum_s \boldsymbol{D}_s^{(k)} n_s^{\mathcal{R}} \gamma(s) - \boldsymbol{D}_s^{(k)} N_s^{\mathcal{P}} \right\|_2^2}_{\text{approximate higher-order balance}} + \underbrace{\sum_s n_s^{\mathcal{R}} \left(\gamma(s) - \frac{N}{n} \right)^2}_{\text{variance penalty}} \\
\text{subject to} \quad & \underbrace{\sum_s \boldsymbol{D}_s^{(1)} n_s^{\mathcal{R}} \gamma(s) = \sum_s \boldsymbol{D}_s^{(1)} N_s^{\mathcal{P}}}_{\text{raking constraint}} \\
& L \leq \gamma(s) \leq U \quad \forall s = 1, \ldots J,
\end{aligned} \tag{9}$$

[8] As we describe above, if we consider up to three-way interactions of age, gender, and educational attainment, the coefficients capture the main effects and the pair-wise second- and third-order interactions in explaining vote choice.

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

where the λ_k are hyper-parameters. Note that the optimization problem is on the scale of the population counts rather than the population proportions. This follows the bound on the bias in Equation (8); categories and interactions with larger population counts $\boldsymbol{D}^{(k)'}\boldsymbol{N}^{\mathcal{P}}$ contribute more to the bias. With a common hyper-parameter, this means that higher-order interactions or smaller categories—which have lower population counts—will have less weight by design.

We can view this optimization problem as adding an additional objective optimizing for higher-order balance to the usual raking estimator in Equation (4) with the scaled squared dissimilarity function and uniform baseline weights. As with the raking estimator, the multilevel calibration weights are constrained to *exactly balance* first-order margins. Subject to this exact marginal constraint, the weights then minimize the imbalance in kth-order interactions for all $k = 2, \dots, d$. In this way, the multilevel calibration weights approximately post-stratify by optimizing for balance in higher-order interactions rather than requiring exact balance in all interactions as the post-stratification weights do. Following the bias–variance decomposition in Equations (6) and (8), this objective is penalized by the sum of the squared weights. As with the raking estimator, we can also replace the variance penalty with a different penalty function, including penalizing deviations from known sampling weights following Deville and Särndal (1992). A benefit of the variance penalty is that Equation (9) is a quadratic program (QP) that can be solved efficiently via first-order methods (Boyd *et al.* 2010). We use OSQP, an efficient first-order method developed by Stellato *et al.* (2020).

Variations on the measure of approximate higher-order balance are also possible. Equation (8) uses the Cauchy–Schwarz inequality to relate the sum of the squared imbalances in kth-order interactions to the bias. We can instead use Hölder's inequality to relate the *maximal* imbalance via the L^∞ norm. Using this measure in Equation (9) would find weights that minimize the imbalance in the worst-balanced kth-order interaction, related to the proposal from Wang and Zubizarreta (2020). We could also solve a variant of Equation (9) without the multilevel structure encoded by the $\boldsymbol{D}_s^{(k)}$ variables. This would treat cells as entirely distinct and perform no aggregation across cells while approximately post-stratifying. From our discussion in Section 3.1, this would ignore the potential bias gains from directly leveraging the multilevel structure. Finally, in Section B of the Supplementary Material, we inspect the Lagrangian dual of this optimization problem and show that the weights are a form of propensity score weights with a multilevel GLM propensity score model.

An important component of the optimization problem are the hyper-parameters λ_k for $k = 2, \dots, d$. They control the relative priority that balancing the higher-order interactions receives in the objective in an inverse relationship, and define a bias–variance trade-off. If λ_k is large, then the weights will be more regularized and the kth-order interaction terms will be less prioritized. In the limit as all $\lambda_k \to \infty$, no weight is placed on any interaction terms, and Equation (9) reduces to raking on the margins. Conversely, if λ_k is small, more importance will be placed on balancing kth-order interactions. For example, if $\lambda_2 = 0$, then the optimization problem will rake on margins *and* second-order interactions. As all $\lambda_k \to 0$, we recover post-stratification weights, if they exist.

The trade-off between bias and variance can be viewed as one between balance and effective sample size. Improving the balance decreases the bias, but comes at the expense of increasing variance by decreasing the effective sample size. In practice, we suggest explicitly tracing out this trade-off. For a sequence of potential hyper-parameter values $\lambda^{(1)}, \lambda^{(2)}, \dots$, set all of the hyper-parameters to be $\lambda_k = \lambda^{(j)}$. We can then look at the two components of the objective in Equation (9), plotting the level of imbalance $\sum_{k=2}^{d} \left\| \sum_s \boldsymbol{D}_s^{(k)} n_s^{\mathcal{R}} \gamma(s) - \boldsymbol{D}_s^{(k)} N_s^{\mathcal{P}} \right\|_2^2$ against the effective sample size $n^{\text{eff}} = \frac{\left(\sum_s n_s^{\mathcal{R}} \hat{\gamma}(s)\right)^2}{\sum_s n_s^{\mathcal{R}} \hat{\gamma}(s)^2}$. After fully understanding this trade-off, practitioners can choose a common λ somewhere along the curve. For example, in our analysis in Section 5, we choose λ to achieve 95% of the potential balance improvement in higher-order terms of $\lambda = 0$ relative to raking.

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

4 Double Regression with Post-Stratification

So far we have focused on multilevel calibration, with weights that exactly match the first-order marginals between the weighted sample and the full population, while approximately balancing higher-order interactions. This approach is independent of the outcomes and so we can use a single set of weights to estimate the population average for multiple different outcomes. However, in some cases, it may not be possible to achieve good covariate balance on higher-order interactions, meaning that our estimates may still be biased. We can address this by specializing to a particular outcome and by explicitly using outcome information to estimate and correct for the bias.

Outcome models for adjustment are widely used in the study of public opinion in Political Science. Multilevel regression with post-stratification (MRP), in particular, is often used to obtain subnational estimates of public opinion on issues such as climate change (e.g., Howe *et al.* 2015) and support for gay rights (e.g., Lax and Phillips 2009) and for the study of policy responsiveness (e.g., Tausanovitch and Warshaw 2014). We begin by reviewing outcome modeling and then propose DRP.

4.1 Using an Outcome Model for Bias Correction

A common alternative to the weighting approach above is to estimate the population average μ using an outcome regression model $m(x_i)$ to predict the outcome given the covariates, averaging over the predictions $\hat{m}(x_i)$. When observations are in discrete cells, this is equivalent to taking the modeled regression estimates of the cell averages, $\hat{\mu}_s$, and *post-stratifying* them to the population totals as (Gelman and Little 1997):

$$\hat{\mu}^{\text{omp}} = \frac{1}{N}\sum_i \hat{m}(x_i) = \frac{1}{N}\sum_s N_s^{\mathcal{P}}\hat{\mu}_s, \tag{10}$$

where $\hat{\mu}_s = 1/n_s \sum_{i:S_i=s}\hat{m}(x_i)$ is the cell-level average prediction, and where "omp" denotes *outcome modeling and post-stratification*. For example, we could obtain a prediction from our model of vote choice in each eight-way interacted cell in our running example, and obtain an overall estimate for vote choice by weighting these by the population proportion in each cell. By smoothing estimates across cells, outcome modeling gives estimates of $\hat{\mu}_s$ even for cells with no respondents, thus sidestepping the primary feasibility problem of post-stratification. We discuss particular choices of outcome regression model in Section 4.2.

Heuristically, for weights $\hat{\gamma}$, we can use an outcome regression model to estimate the bias (conditional on the cell counts) due to imbalance in higher-order interactions by taking the difference between the outcome regression model estimate for the population and a hypothetical estimate with population cell counts $n_s^{\mathcal{R}}\hat{\gamma}(s)$:

$$\widehat{\text{bias}} = \hat{\mu}^{\text{omp}} - \frac{1}{N}\sum_s n_s^{\mathcal{R}}\hat{\gamma}(s)\hat{\mu}_s = \frac{1}{N}\sum_s \hat{\mu}_s \times \left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}}\hat{\gamma}(s)\right). \tag{11}$$

This uses the outcome model to collapse the imbalance in the J cells into a single diagnostic. Our main proposal is to use this diagnostic to correct for any remaining bias from the multilevel calibration weights. We refer to the estimator as *DRP*, as it incorporates two forms of "regression"—a regression of the outcome $\hat{\mu}(s)$ and a regression of response $\hat{\gamma}(s)$ through the dual problem, as we discuss in Section B of the Supplementary Material. We construct the estimator using weights $\hat{\gamma}(s)$ and cell estimates $\hat{\mu}_s$ as

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

$$\begin{aligned}\hat{\mu}^{\text{drp}}(\hat{\gamma}) = \hat{\mu}(\hat{\gamma}) \quad & + \frac{1}{N}\sum_{s}\hat{\mu}_s \times \underbrace{\left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}}\hat{\gamma}(s)\right)}_{\text{imbalance in cell } s} \\ = \hat{\mu}^{\text{omp}} \quad & + \frac{1}{N}\sum_{s} n_s^{\mathcal{R}}\hat{\gamma}(s) \times \underbrace{(\bar{Y}_s - \hat{\mu}_s)}_{\text{error in cell } s} .\end{aligned} \tag{12}$$

The two lines in Equation (12) give two equivalent perspectives on how the DRP estimator adjusts for imbalance. The first line begins with the multilevel calibration estimate $\hat{\mu}(\hat{\gamma})$ and then adjusts for the estimate of the bias using the outcome model $\frac{1}{N}\sum_s \hat{\mu}_s(N_S^{\mathcal{P}} - n_s^{\mathcal{R}}\hat{\gamma}(s))$. If the population and re-weighted cell counts are substantially different in important cells, the adjustment from the DRP estimator will be large. On the other hand, if the population and re-weighted sample counts are close in all cells then $\hat{\mu}^{\text{drp}}(\hat{\gamma})$ will be close to $\hat{\mu}(\hat{\gamma})$. In the limiting case of post-stratification where all the counts are equal, the two estimators will be equivalent, $\hat{\mu}^{\text{drp}}(\hat{\gamma}^{\text{ps}}) = \hat{\mu}(\hat{\gamma}^{\text{ps}})$. The second line instead starts with the outcome regression estimate, $\hat{\mu}^{\text{omp}}$, and adjusts the estimate based on the error within each cell. If the outcome model has poor fit in cells that have large weight, then the adjustment will be large. This estimator is a special case of augmented approximate balancing weights estimators (e.g., Hirshberg and Wager 2021) and is closely related to generalized regression estimators (Cassel, Sarndal, and Wretman 1976), augmented IPW estimators (Chen *et al.* 2020), and bias-corrected matching estimators (Rubin 1976).

This DRP approach uses outcome information to reduce bias by adjusting for imbalance remaining after weighting. To see this, we can again inspect the estimation error. Analogous to Equation (5), the difference between the DRP estimator and the true population average is

$$\hat{\mu}^{\text{drp}}(\hat{\gamma}) - \mu = \frac{1}{N}\sum_{s}\underbrace{\left(n_s^{\mathcal{R}}\hat{\gamma}(s) - N_s^{\mathcal{P}}\right)}_{\text{imbalance in cell } s} \times \underbrace{(\hat{\mu}_s - \mu_s)}_{\text{error in cell } s} + \underbrace{\frac{1}{N}\sum_{s=1}^{s} n_s^{\mathcal{R}}\hat{\gamma}(s)\bar{\varepsilon}_s}_{\text{noise}} . \tag{13}$$

Comparing to Equation (5), where the estimation error depends solely on the imbalance and the true cell averages, we see that the estimation error for DRP depends on the *product* of the imbalance from the weights and the estimation error from the outcome model. Therefore, if the model is a reasonable predictor for the true cell averages, the estimation error will decrease.

In Section C of the Supplementary Material, we formalize this intuition via finite-population asymptotic theory. We find that as long as the modeled cell averages estimate the true cell averages well enough, the model and the calibration weights combine to ensure that the bias will be small enough to conduct normal-based asymptotic inference. Furthermore, the asymptotic variance will depend on the variance of the *residuals*ε_i, which we expect to have much lower variance than the raw outcomes. So asymptotically, the DRP estimator will also have lower variance than the oracle Horvitz–Thompson estimator that uses the true response probabilities, similar to other model-assisted estimators (Breidt and Opsomer 2017). We construct confidence intervals for the population total μ based on these asymptotic results. First, we start with a plug-in estimate for the variance,

$$\hat{V} = \frac{1}{N^2}\sum_{i=1}^{n} R_i \hat{\gamma}(S_i)^2 (Y_i - \hat{\mu}_{S_i})^2 . \tag{14}$$

We then construct approximate level α confidence intervals via $\hat{\mu}^{\text{drp}}(\hat{\gamma}) \pm z_{1-\alpha/2}\sqrt{\hat{V}}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution.

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

4.2 Choosing an Outcome Model for Bias Correction

The choice of outcome model is crucial for bias correction. As we discussed in Section 3, we often believe that the strata have important hierarchical structure where main effects and lower-order interactions are more predictive than higher-order interactions. We consider two broad classes of outcome model that accommodate this structure: multilevel regression outcome models, which explicitly regularize higher-order interactions; and tree-based regression models, which implicitly regularize higher-order interactions. Section C of the Supplementary Material gives technical conditions on the outcome regression model under the finite-population asymptotic framework above. We require these regularized models to estimate the true relationship sufficiently well—that is, the regression error of the cell-level estimates $\sum_s (\mu_s - \hat{\mu}_s)^2$ must decrease to zero at a particular rate as the population and sample size increases—depending on how the total number of cells and the population size relate.

4.2.1 Multilevel Outcome Model. We first consider multilevel models, which have a linear form as $\hat{\mu}_s^{\text{mr}} = \hat{\boldsymbol{\eta}}^{\text{mr}} \cdot \boldsymbol{D}_s$, where $\hat{\boldsymbol{\eta}}^{\text{mr}}$ are the estimated regression coefficients (Gelman and Little 1997; Ghitza and Gelman 2013). MRP directly post-stratifies these model estimates, using the coefficients to predict the value in the population:

$$\hat{\mu}^{\text{mrp}} = \hat{\boldsymbol{\eta}}^{\text{mr}} \cdot \frac{1}{N} \sum_s N_s^{\mathcal{P}} \boldsymbol{D}_s.$$

This is closely related to the multilevel calibration approach in Section 3.2. If we set $L = -\infty$ and $U = \infty$ in Equation (9)—and so allow for unbounded extrapolation—the resulting estimator will be equivalent to using the *maximum a posteriori* (MAP) estimate of $\hat{\boldsymbol{\eta}}^{\text{mr}}$, with regularization hyperparameters $\lambda^{(k)}$ (Ben-Michael *et al.* 2021). In contrast, the DRP estimator only uses the coefficients to adjust for any remaining imbalance after weighting,

$$\hat{\mu}^{\text{drp}}(\hat{\gamma}) = \hat{\mu}(\hat{\gamma}) + \hat{\boldsymbol{\eta}}^{\text{mr}} \cdot \left(\frac{1}{N} \sum_s \boldsymbol{D}_s \left(N_s^{\mathcal{P}} - n_s^{\mathcal{R}} \hat{\gamma}(s) \right) \right).$$

This performs bias correction. When we use the MAP estimate of a multilevel outcome model, the corresponding DRP estimator is itself a weighting estimator where the outcome regression model directly adjusts the weights:

$$\tilde{\gamma}(s) = \hat{\gamma}(s) + \left(\boldsymbol{N}^{\mathcal{P}} - \text{diag}(n^{\mathcal{R}})\hat{\gamma} \right)' \boldsymbol{D} \left(\boldsymbol{D}' \text{diag}(\boldsymbol{n}^{\mathcal{R}}) \boldsymbol{D} + \boldsymbol{Q} \right)^{-1} \boldsymbol{D}_s,$$

where $\boldsymbol{Q}$ is the prior covariance matrix associated with the multilevel model (Breidt and Opsomer 2017). While the multilevel calibration weights $\hat{\gamma}$ are constrained to be nonnegative, DRP weights allow for *extrapolation* outside the support of the data (Ben-Michael *et al.* 2021).

4.2.2 Trees and General Weighting Methods. More generally, we can consider an outcome regression model that smooths out the cell averages, using a weighting function between cells s and s', $W(s, s')$, to estimate the population cell average, $\hat{\mu}_s = \sum_{s'} W(s, s') n_{s'}^{\mathcal{R}} \bar{Y}_{s'}$. A multilevel model is a special case that smooths the cell averages by partially pooling together cells with the same lower-order features. In general, the DRP estimator is again a weighting estimator, with adjusted weights

$$\tilde{\gamma}(s) = \hat{\gamma}(s) + \sum_{s'} W(s, s')(N_{s'}^{\mathcal{P}} - n_{s'}^{\mathcal{R}} \hat{\gamma}(s')).$$

Here the weights are adjusted by a smoothed average of the imbalance in similar cells. In the extreme case, where the weight matrix is diagonal with elements $\frac{1}{n^{\mathcal{R}}}$, the DRP estimate reduces to the post-stratification estimate, as above.

One important special case is tree-based methods such as those considered by Montgomery and Olivella (2018) and Bisbee (2019). These methods estimate the outcome via bagged regression trees, such as random forests, gradient boosted trees, or Bayesian additive regression trees. These approaches can be viewed as data-adaptive weighting estimators where the weight for cell s and s', $W(s, s')$, is proportional to the fraction of trees where cells s and s' share a leaf node (Athey, Tibshirani, and Wager 2019). Therefore, the DRP estimator will smooth the weights by adjusting them to correct for the imbalance in cells that share many leaf nodes.

4.2.3 *Bias–Variance Trade-Off.* The key difference between OMP and DRP is what role the outcome model plays, and how one chooses the model to negotiate the bias–variance trade-off. Because outcome model-style estimators *only* use the outcome model, the performance of the outcome model completely determines the performance of the estimator. For example, in a multilevel model, we want to include higher-order interaction terms in order to reduce the bias. However, this can increase the variance to an unacceptable degree, so we choose a model with lower variance and higher bias.

In contrast, with DRP, the role of the model is to correct for any potential bias remaining after multilevel calibration. Because we can only *approximately* post-stratify, this bias-correction is key. It also means that DRP is less reliant on the outcome model, which only needs to adequately perform bias correction. Therefore, the bias–variance trade-off is different for DRP, and we prioritize bias over variance. By including higher-order interactions or deeper trees, the model will be able to adjust for any remaining imbalance in higher-order interactions after weighting.

5 2016 U.S. Presidential Election Polls

We now turn to evaluating the proposed estimators in the context of 2016 U.S. presidential polling, as described in Section 1.1. We begin by showing balance gains from the multilevel calibration procedure and inspecting how bias correction through DRP affects both the point estimates and confidence intervals. We then evaluate the performance of multilevel calibration and DRP when predicting state-specific vote counts from the national preelection survey of vote intention. In Section E of the Supplementary Material, we conduct simulations calibrated to our application and show that there are sizeable reductions in RMSE due to multilevel calibration over raking on the margins, and that bias reduction can provide large improvements.

We compute population cell counts $N_s^{\mathcal{P}}$ from the post-2016 election CCES poll, limiting to those who voted in the election as indicated by a flag for a verified voter from the Secretaries of State files, and weighting according to provided CCES survey weights. We consider the balance of three different weighting estimators. First, we rake on margins for eight variables measured in both surveys, equivalent to solving (9) with $\lambda_k \to \infty$ for $k \geq 2$ and $L = 0, U = \infty$. Next, we create post-stratification weights. Due to the number of empty cells, we limit to post-stratifying on four variables, collapsed into coarser cells.[9] Last, we balance up to sixth-order interaction terms, setting a common hyper-parameter $\lambda_k = \lambda$ for $k = 2, \dots, 6$ and $\lambda_k \to \infty$ for $k = 7, 8$.[10] To select λ, we solve (9) for a series of 40 potential values, equally spaced on the log scale, tracing out the bias–variance trade-off in Figure 2. On a 2020 Intel-based MacBook Pro, finding the weights across these 40 λ values takes 2.5 minutes, warm-starting the optimization for each value of lambda with the optimal weights at the previous value. We find that a value of $\lambda = 12.8$ achieves 95% of the potential imbalance reduction while having an effective sample size 30% larger than the least-regularized solution. We also consider bias-correcting the multilevel weights with DRP with (a) a fourth-order ridge regression model and (b) gradient boosted trees, both tuned with cross validation.

[9] We collapse income and age to three levels, education to a binary indicator for greater than a high school degree, and race to a binary indicator for white.

[10] Seventh- and eighth-order interactions are unlikely to be meaningful given the lower-order interactions, but including them substantially increases the memory and time complexity of solving (9).

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

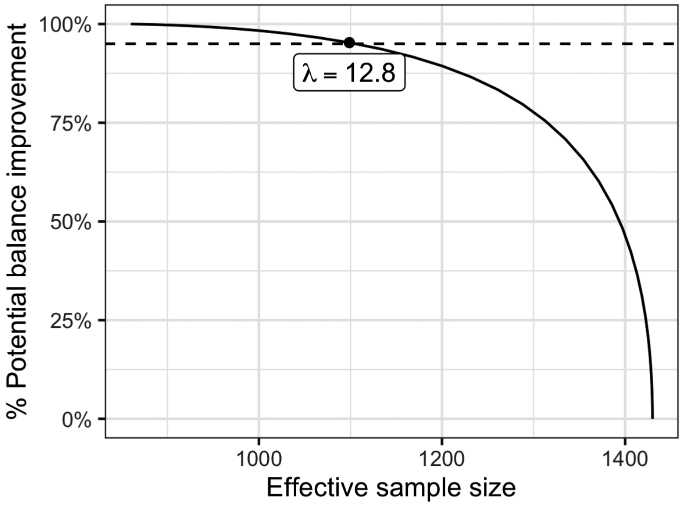


Figure 2. Difference between the re-weighted sample and the population, measured as the square root of the sum of squared imbalances for interactions $k = 1, \ldots, 6$, versus the effective sample size. Imbalance measures are scaled as the percent reduction in imbalance relative to raking on margins.

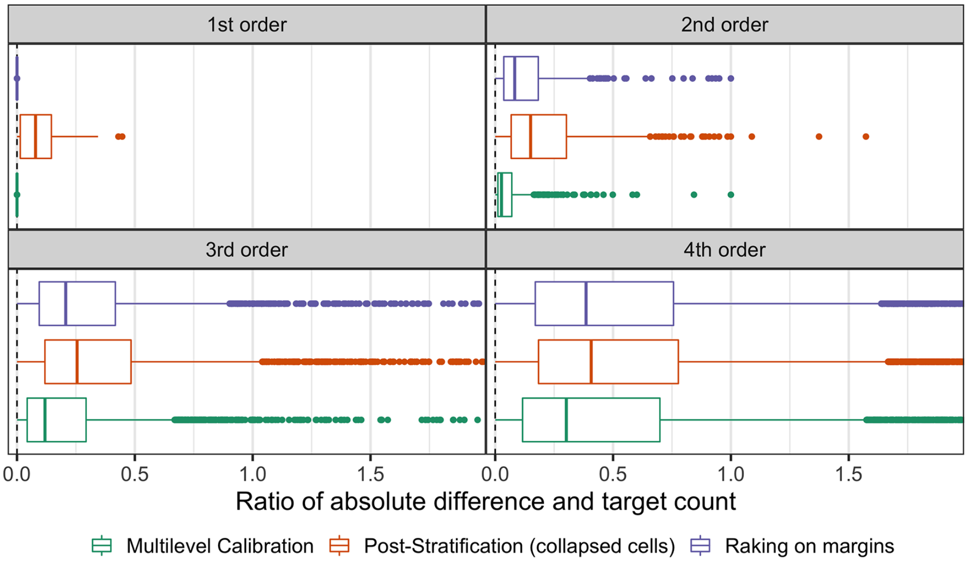
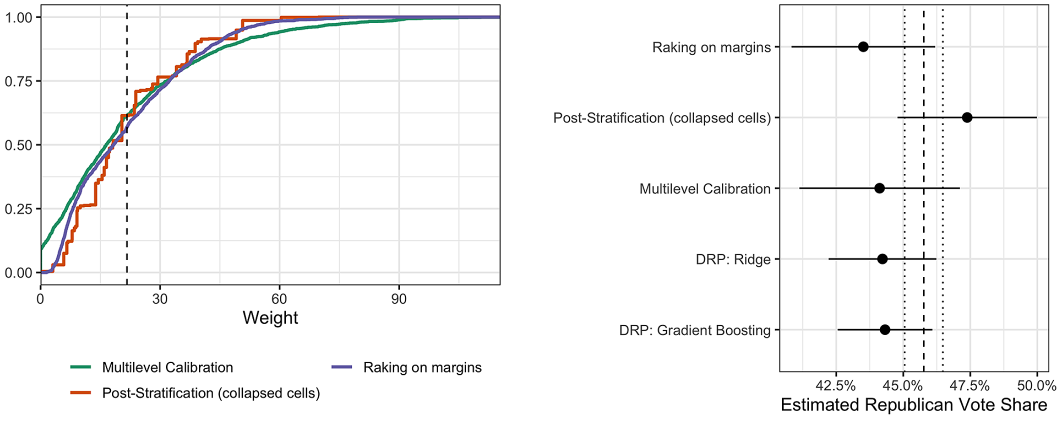


Figure 3. Distribution of covariate imbalance for interactions up to order 4, measured as the difference between the weighted and target counts, divided by the target count.

Figure 3 shows the imbalance when weighting by these three approaches for interactions up to order 4. To place the balance on the same scale, we divide the difference between the re-weighted sample and the population in the jth interaction of order k by the population count, $\frac{\left|\sum_s D_{sj}^{(k)}(n_s^{\mathcal{R}}\hat{\gamma}(s)-N^{\mathcal{P}})\right|}{\sum_s D_{sj}^{(k)} N_s^{\mathcal{P}}}$. By design, both the raking and multilevel calibration weights exactly balance first-order margins; however, post-stratifying on a limited set of collapsed cells does not guarantee balance on the margins of the uncollapsed cells, due to missing values. The multilevel calibration weights achieve significantly better balance on second-order interactions than do the raking weights or the post-stratification weights. For higher-order interactions, these gains are still visible but less stark, as it becomes more difficult to achieve good balance.

This improvement in balance comes at some cost to variance. Figure 4a shows the empirical CDF of the respondent weights for the three approaches. The multilevel calibration weights that balance higher-order interactions have a greater proportion of large weights, with a longer tail

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press



(a) Empirical CDF of weights

(b) Predictions of Republican vote share

Figure 4. (a) Distribution of weights. Dashed line indicates a uniform adjustment $\frac{N}{n}$. (b) Point estimates and approximate 95% confidence intervals. Thick dashed line is the weighted CCES estimate, and thinner dashed lines indicate the lower and upper 95% confidence limits.

than raking or collapsed post-stratification. These large weights lead to a smaller effective sample size. The multilevel calibration weights yield an effective sample size of 1,099 for a design effect of 1.87, while raking and the collapsed post-stratification weights have effective sample sizes of 1,431 and 1,482, respectively.

Figure 4b plots the point estimates and approximate 95% confidence intervals for the multilevel calibration and DRP approaches, along with the estimated Republican vote share from the weighted CCES sample. Confidence intervals are computed as $\hat{\mu}^{\text{drp}}(\hat{\gamma}) \pm z_{1-0.025}\sqrt{\hat{V}}$, with the standard error estimate from Equation (14). The different weights result in different predictions of the vote share, ranging from a point estimate of 42.5% for raking to 47.5% for post-stratification. Additionally, the somewhat smaller effective sample size for multilevel calibration manifests itself in the standard error, leading to slightly larger confidence intervals. The DRP estimators, bias correcting with either ridge regression or gradient boosted trees, have similar point estimates to multilevel calibration alone. This indicates that the remaining imbalances in higher-order interactions after weighting in Figure 3 do not lead to large estimated biases. However, by including an outcome model the DRP estimators significantly reduce the standard errors.

To empirically validate the role of balancing higher-order interactions, we use the national preelection Pew survey to predict Republican vote share within each state. The preelection survey was designed as a *national* survey and so there are substantial differences between the sample and the state-level totals. For each state, we compute the population count vector $N^{\mathcal{P}}$ from the weighted CCES, subsetted to the state of interest. Here, we use a common $\lambda = 1$. We impute the Republican vote share for that state via weighting alone and DRP with gradient boosted trees, balancing interactions up to order 6; we also include OMP estimates using gradient boosted trees for the outcome model. We consider both restricting the sample respondents to be in the same region as the state and including all sample respondents. Figure 5 shows the absolute bias and RMSE across the 50 states as the order increases from raking on first-order margins to approximately balancing sixth-order interactions. There are substantial gains to bias-correction through DRP when raking on the margins in terms of both bias and variance. Balancing higher-order interactions also improves estimation over raking alone. And, while the relative improvement of DRP over multilevel calibration diminishes as we balance higher-order interactions, these gains are still apparent, though small. This indicates that the additional bias-correction from gradient-boosted trees has less impact than balancing higher-order interactions does. Finally, while not restricting respondents by region results in lower bias across the 50 states, the higher RMSE indicates the estimates of state vote share are poor but averaging out. In Figure A.2 in the

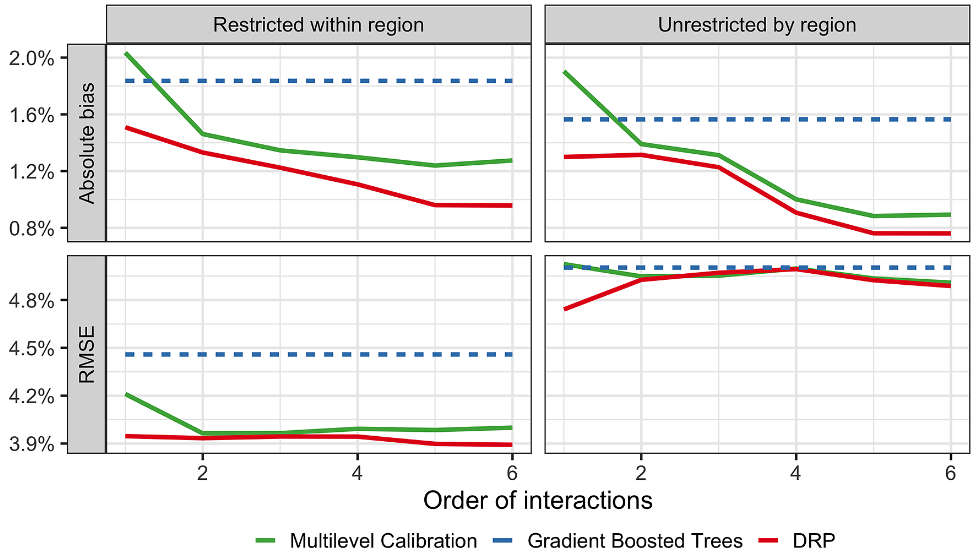


Figure 5. Absolute bias and MSE when imputing Republican vote share in 50 states from the national Pew survey, using multilevel calibration weights and DRP with gradient boosted trees, balancing margins up to sixth-order interactions, restricting to respondents in the same region and unrestricted by region. Blue dashed lines show the bias and RMSE for an OMP using gradient boosted trees for comparison.

Supplementary Material, we show the individual DRP estimates of state-level Republican vote share balancing up to sixth-order interactions and using gradient boosted trees, with respondents restricted to the same region. We see that DRP is biased in the negative direction, somewhat underestimating Republican vote share in the majority of states.

6 Discussion

As recent public opinion polling has shown, differential non-response across groups defined by fine-grained higher-order interactions of covariates can lead to substantial bias. While, ideally, we would address such nonresponse by post-stratifying on all interactions of important covariates simultaneously, the cost of collecting the necessary sample size is prohibitive, especially with low response rates. In practice, analysts circumvent this via *ad hoc* approaches, such as only adjusting for first-order marginal characteristics or collapsing cells together.

In this paper, we provide two alternative approaches, *multilevel calibration weighting* and *DRP*, which provide principled ways to combine fine-grained calibration weighting and modern machine-learning prediction techniques. The multilevel calibration weights improve on existing practice by approximately post-stratifying in a data-driven way, while at least ensuring exact raking on first-order margins. DRP then takes advantage of flexible regression methods to further adjust for differences in fine-grained cells in a parsimonious way. For groups where the weights successfully adjust for differences in response rates, the DRP estimate is driven by the weights; for groups that remain over- or under-represented, DRP instead relies on a flexible regression model to estimate and adjust for remaining non-response bias. Through theoretical, numerical, and simulation results, we find that these approaches can significantly improve estimation. Specifically, adjusting for higher-order interactions with multilevel calibration has much lower bias than ignoring them by only raking on the first-order margins. Incorporating flexible outcome estimators such as multilevel regression or tree-based approaches in our DRP estimator further improves upon weighting alone.

However, our proposal is certainly not a panacea, and important questions remain. First, while we choose the value of the hyper-parameters by tracing out the bias–variance trade-off, it might be preferable to select them via data-adaptive measures. For example, Wang and Zubizarreta

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

(2020) propose a cross-validation style approach that takes advantage of the Lagrangian dual formulation. It may be possible to use such approaches in this setting.

Second, the key assumption is that outcomes are missing at random within cells. While we never expect this to be entirely true, it allows us to make progress on estimation, and with granular enough groups, we may hope that this assumption is approximately true. It is important then to characterize how our conclusions would change if this assumption is violated, and the response and the outcome are correlated even within cells. Hartman and Huang (2022) discuss this form of *sensitivity analysis* for survey weights that is readily adaptable to this context. In a similar vein, some covariates (or their interactions) may be irrelevant to the response probability, in which case enforcing balance on them would lead to decreased precision with no reduction in bias. To avoid this, researchers can combine our proposals with a covariate selection procedure (e.g., Egami and Hartman 2021) that can reduce the number of covariates to balance.

Third, with many higher-order interactions, it is difficult to find good information on population targets. We may have to combine various data sources collected in different manners, or estimate unknown cells in the target population (Kuriwaki *et al.* 2021), and uncertainty in the population targets can also lead to increased variance (see Caughey *et al.* (2020) for a recent review). Fourth, during the survey process, we can obtain very detailed auxiliary information on survey respondents that we cannot obtain for the population, even marginally. Incorporating this sort of auxiliary information into the estimation procedure will be important to future work.

Finally, non-response bias is far from the only difficulty with modern surveys. We therefore view multilevel calibration and DRP as only one part of the analyst's toolkit, supplementing design and data considerations.

Acknowledgments

We would like to thank the editor and reviewers, Peng Ding, Yair Ghitza, Mark Handcock, Luke Keele, Shiro Kuriwaki, Drew Linzer, Luke Miratrix, Doug Rivers, Jonathan Robinson, Jas Sekhon, Betsy Sinclair, Brandon Stewart, and participants at BigSurv2020, AsianPolmeth VIII, and Polmeth 2021 for useful discussion and comments.

Data Availability Statement

Replication code for this article is available in Ben-Michael *et al.* (2023) at https://doi.org/10.7910/DVN/J7BSXQ.

Funding

This research was supported in part by the Hellman Family Fund at UC Berkeley and by the Institute of Education Sciences, U.S. Department of Education, Grant R305D200010. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Supplementary Material

For supplementary material accompanying this paper, please visit http://doi.org/10.1017/pan.2023.9

References

Ansolabehere, S., and B. F. Schaffner. 2017. "CCES Common Content, 2016." Version V4. https://doi.org/10.7910/DVN/GDF6Z0.

Athey, S., J. Tibshirani, and S. Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1179–1203. https://doi.org/10.1214/18-AOS1709; arXiv:1610.01271.

Ben-Michael, E., A. Feller, and E. Hartman. 2023. Replication Data for: Multilevel Calibration Weighting for Survey Data. Version V1. https://doi.org/10.7910/DVN/J7BSXQ.

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

Ben-Michael, E., A. Feller, and J. Rothstein. 2021. “The Augmented Synthetic Control Method.” *Journal of the American Statistical Association* 116: 1789–1803. https://doi.org/10.1080/01621459.2021.1929245

Bisbee, J. 2019. “BARP: Improving Mister P Using Bayesian Additive Regression Trees.” *American Political Science Review* 113 (4): 1060–1065.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2010. “Distributed optimization and statistical learning via the alternating direction method of multipliers.” *Foundations and Trends in Machine Learning* 3 (1): 1–122. https://doi.org/10.1561/2200000016

Breidt, F. J., and J. D. Opsomer. 2017. “Model-Assisted Survey Estimation with Modern Prediction Techniques.” *Statistical Science* 32 (2): 190–205. https://doi.org/10.1214/16-STS589.

Cassel, C. M., C.-E. Sarndal, and J. H. Wretman. 1976. “Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations.” *Biometrika* 63 (3): 615–620.

Caughey, D., A. J. Berinsky, S. Chatfield, E. Hartman, E. Schickler, and J. S. Sekhon. 2020. *Target Estimation and Adjustment Weighting for Survey Nonresponse and Sampling Bias*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108879217

Caughey, D., and E. Hartman. 2017. “Target Selection as Variable Selection: Using the Lasso to Select Auxiliary Vectors for the Construction of Survey Weights.” Available at SSRN 3494436.

Chen, J. K. T., R. L. Valliant, and M. R. Elliott. 2019. “Calibrating Non-probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68 (3): 657–681.

Chen, Y., P. Li, and C. Wu. 2020. “Doubly Robust Inference with Nonprobability Survey Samples.” *Journal of the American Statistical Association* 115 (532): 2011–2021.

Cox, D. R. 1984. “Interaction.” *International Statistical Review/Revue Internationale de Statistique* 52 (1): 1–24.

D’Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon. 2020. “ *Overlap in Observational Studies with High-Dimensional Covariates*.” *Journal of Econometrics* 221: 644–654.

Deming, W. E., and F. F. Stephan. 1940. “On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals Are Known.” *The Annals of Mathematical Statistics* 11 (4): 427–444. https://doi.org/10.1214/aoms/1177731829 arXiv:1306.3979v1.

Deville, J. C., and C. E. Särndal. 1992. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87 (418): 376–382. https://doi.org/10.1080/01621459.1992.10475217.

Deville, J. C., C. E. Särndal, and O. Sautory. 1993. “Generalized Raking Procedures in Survey Sampling.” *Journal of the American Statistical Association* 88 (423): 1013–1020. https://doi.org/10.1080/01621459.1993.10476369

Egami, N., and E. Hartman. 2021. “Covariate Selection for Generalizing Experimental Results: Application to a Large-Scale Development Program in Uganda.” *Journal of the Royal Statistical Society. Series A: Statistics in Society* 184 (4): 1524–1548. https://doi.org/10.1111/rssa.12734; arXiv:1909.02669.

Gao, C., S. Yang, and J. K. Kim. 2023. “Soft Calibration for Selection Bias Problems Under Mixed-Effects Models.” *Biometrika*. https://doi.org/10.1093/biomet/asad016.

Gelman, A., and T. C. Little. 1997. “Poststratification into Many Categories Using Hierarchical Logistic Regression.” *Survey Methodology* 23 (2): 127–135.

Ghitza, Y., and A. Gelman. 2013. “Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups.” *American Journal of Political Science* 57 (3): 762–776. https://doi.org/10.1111/ajps.12004

Guggemos, F., and Y. Tillé. 2010. “Penalized Calibration in Survey Sampling: Design-Based Estimation Assisted by Mixed Models.” *Journal of Statistical Planning and Inference* 140 (11): 3199–3212. https://doi.org/10.1016/j.jspi.2010.04.010

Hartman, E., C. Hazlett, and C. Sterbenz. 2021. “Kpop: A Kernel Balancing Approach for Reducing Specification Assumptions in Survey Weighting.” Preprint, arXiv:2107.08075 [stat.ME].

Hartman, E., and M. Huang. 2022. “Sensitivity Analysis for Survey Weights.” Preprint, arXiv:2206.07119 [stat.ME].

Hirshberg, D., and S. Wager. 2021. “Augmented Minimax Linear Estimation.” *Annals of Statistics* 49 (6): 3206–3227. https://doi.org/10.1214/21-AOS2080.

Hirshberg, D. A., A. Maleki, and J. Zubizarreta. 2019. “Minimax Linear Estimation of the Retargeted Mean.” Preprint, arXiv:1901.10296v1.

Howe, P. D., M. Mildenberger, J. R. Marlon, and A. Leiserowitz. 2015. “Geographic Variation in Opinions on Climate Change at State and Local Scales in the USA.” *Nature Climate Change* 5 (6): 596–603.

Kennedy, C., et al. 2018. “An Evaluation of the 2016 Election Polls in the United States.” *Public Opinion Quarterly* 82 (1): 1–33. https://doi.org/10.1093/poq/nfx047

Kuriwaki, S., S. Ansolabehere, A. Dagonel, and S. Yamauchi. 2021. “The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level.” https://doi.org/10.31219/osf.io/mk9e6.

Lax, J. R., and J. H. Phillips. 2009. “How Should we Estimate Public Opinion in the States?” *American Journal of Political Science* 53 (1): 107–121.

Linzer, D. A. 2011. “Reliable Inference in Highly Stratified Contingency Tables: Using Latent Class Models as Density Estimators.” *Political Analysis* 19: 173–187.

https://doi.org/10.1017/pan.2023.9 Published online by Cambridge University Press

McConville, K. S., F. J. Breidt, T. C. Lee, and G. G. Moisen. 2017. “Model-Assisted Survey Regression Estimation with the Lasso.” *Journal of Survey Statistics and Methodology* 5 (2): 131–158.

Mercer, A., A. Lau, and C. Kennedy. 2018. For Weighting Online Opt-in Samples, What Matters Most: Pew Research Center.

Montgomery, J. M., and S. Olivella. 2018. “Tree-Based Models for Political Science Data.” *American Journal of Political Science* 62 (3): 729–744. https://doi.org/10.1111/ajps.12361.

Ning, Y., P. Sida, and K. Imai. 2020. “Robust Estimation of Causal Effects Via a High-Dimensional Covariate Balancing Propensity Score.” *Biometrika* 107 (3): 533–554. https://doi.org/10.1093/biomet/asaa020; arXiv:1812.08683.

Park, M., and W. A. Fuller. 2009. “The Mixed Model for Survey Regression Estimation.” *Journal of Statistical Planning and Inference* 139 (4): 1320–1331. https://doi.org/10.1016/j.jspi.2008.02.021.

Pew Research Center . 2016. “As Election Nears, Voters Divided over Democracy and ‘Respect’.” Technical report, Pew Research Center, October 2016.

Rao, J. N. K., and A. C. Singh. 1997. “A Ridge-Shrinkage Method for Range-Restricted Weight Calibration in Survey Sampling.” In *JSM Proceedings of the Section on Survey Research Methods*, 57–85. Alexandria, VA: American Statistical Association.

Rubin, D. B. 1976. “Inference and Missing Data.” *Biometrika* 63 (3): 581–592.

Si, Y., R. Trangucci, J. S. Gabry, and A. Gelman. 2020. “Bayesian Hierarchical Weighting Adjustment and Survey Inference.” *Survey Methodology* 46 (2): 181–214.

Stellato, B., G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. 2020. “OSQP: An Operator Splitting Solver for Quadratic Programs.” *Mathematical Programming Computation* 12 (4): 637–672. https://doi.org/10.1007/s12532-020-00179-2.

Tausanovitch, C., and C. Warshaw. 2014. “Representation in Municipal Government.” *American Political Science Review* 108 (3): 605–641.

Wang, Y., and J. R. Zubizarreta. 2020. “Minimal Dispersion Approximately Balancing Weights: Asymptotic Properties and Practical Considerations.” *Biometrika* 107 (1): 93–105. https://doi.org/10.1093/biomet/asz050; arXiv:1705.00998.

Xu, Y., and E. Yang. 2022. “Hierarchically Regularized Entropy Balancing.” *Political Analysis*: 1–8. https://doi.org/10.1017/pan.2022.12.

Yang, S., J. K. Kim, and R. Song. 2020. “Doubly Robust Inference when Combining Probability and Non-probability Samples with High Dimensional Data.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (2): 445–465.

Zubizarreta, J. R. 2015. “Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data.” *Journal of the American Statistical Association* 110 (511): 910–922. https://doi.org/10.1080/01621459.2015.1023805.