Calibrating Small-Area MRP Estimates to Known Population Quantities

William Marble[*] Josh Clinton[†]

August 25, 2023

Abstract

Researchers commonly use multilevel regression and poststratification (MRP) to estimate opinion in small geographies and to adjust unrepresentative samples. Yet, MRP estimates are still subject to error, whether due to modeling choices or non-ignorable survey error caused by declining response rates and differential nonresponse. We show how auxiliary quantities whose marginal distributions are known at the geography of interest — such as election outcomes — can be leveraged to account for this error. We propose jointly modeling the geographic-level correlation between the auxiliary calibration variables and responses to survey items of interest, whose truth is unknown. We then use the estimated correlations to adjust the estimates of target variables. T his procedure relies on the assumption that the correlations observed in the sample are the same as in the populatio. To validate the approach, we use a pre-election poll to examine vote intention in three statewide races in Michigan in 2022. We calibrate estimates to one election and examine errors for the remaining races. We find that calibration decreases the average error by two-thirds — suggesting that the method has the potential to greatly increase the accuracy and value of MRP estimates. We then apply the method to study the distribution of partisan animus across the country and show that uncalibrated estimates overstate animus, consistent with survey respondents being strong partisans. We also find that animus is most common in politically homogenous counties, not those that are closely contested.

[*] Director of Data Science, Program for Opinion Research and Election Studies, University of Pennsylvania. Email: marblew@upenn.edu

[†] Abby and Jon Winkelried Chair, Professor of Political Science, Department of Political Science, Vanderbilt University. Email: josh.clinton@vanderbilt.edu

Surveys are fundamental to the measurement of public opinion and behavior. Political scientists use surveys to study democratic legitimacy, accountability and representation. Economists rely on surveys to gauge consumer confidence and economic behavior. And as the recent pandemic revealed, public health scholars and policymakers rely on surveys to measure the incidence of infection and vaccination. However, high-quality surveys have become more difficult and expensive to conduct. Random samples of individuals are harder than ever to recruit, fueling concern that survey respondents may be different from the populations they are meant to represent.

In recent years, researchers have developed flexible methods to adjust survey estimates to account for non-representative samples. While specific methods vary, the core idea in most work is to account for discrepancies between observable sociodemographic characteristics of survey respondents and known population benchlines. One particularly successful paradigm for such adjustment is multilevel regression and poststratification, or MRP (Gelman and Little, 1997; Park, Gelman and Bafumi, 2004). This method models opinion using individual-level and geographic-level predictors to improve survey estimates in small populations that were not intentionally sampled. It combines regression modeling, which reduces sampling variability, with poststratification, which ensures that inferences take into account sociodemographic discrepancies between the sample and population.

MRP and other survey adjustment methods fundamentally rely on an ignorability assumption: conditional on the covariates included in the model, representation in the survey is conditionally independent of the outcome being studied. This assumption justifies using estimates from a survey as stand-ins for opinion in subgroups of the population as a whole. This assumption is more plausible when the regression model includes fine-grained covariates that are highly predictive of the outcome. As such, much recent research in the MRP paradigm has focused on specifying complex interactions in the regression (e.g. Ghitza and Gelman, 2013; Broniecki, Leemann and Wuest, 2022; Goplerud, 2023; Bisbee, 2019) and on extending the number of poststratification variables that can be included (Leemann and Wasserfallen, 2017). Nonetheless, given challenges with surveys in the contemporary environment, respondents may be different in unobservable, nonignorable ways even after flexible modeling. Moreover, as the number of variables increases, bias is likely to be reduced at the cost of increased variance. This variability can limit the accuracy of MRP in estimating opinion in small geographies and subgroups (Buttice and Highton, 2013; Warshaw and Rodden, 2012).

In many cases, however, auxiliary ground-truth data for certain outcomes is available. In this paper, we propose a method to use known auxiliary data — such as election results — to improve MRP-based estimates of public opinion in small geographies and subgroups. The goal is to generate accurate measures of public opinion on issues for which no ground-truth baseline data is available, such as policy attitudes. Our approach is to use the errors in survey-based estimates for the auxiliary variables, which have ground-truth data, to update inferences for the variables of interest.

The core insight is that because attitudes are correlated across issues, the presence of known errors in some measures enables calibration of estimates on other measures. The calibration strategy we propose is principled and data-driven. We first fit a joint model for the survey questions of interest and for the auxiliary variables, generating an estimate of the geographic correlation between the outcomes. After estimation, we use poststratification to generate model-based estimates for the auxiliary variables, which can be compared to the ground-truth data. The observed error in the auxiliary variables, along with the correlation between auxiliary variables and the variables of interest, form the basis of the adjustment. The adjustments are larger for variables that are more highly correlated with the auxiliary variables, holding all else constant. The method provides a systematic way to adjust for survey error that cannot be accounted for using traditional sociodemographic weighting targets. It also provides a way to incorporate ground-truth data that is observed at geographies much smaller than those sampled in the survey — improving estimates of public opinion within those geographies.

The method relies on the assumption that the survey-based estimates of correlations between the auxiliary variables and the variable of interest accurately capture the population-level correlations. This assumption may not be valid. For example, survey respondents tend to be more politically knowledgeable than the population, so their attitudes may be more correlated. Still, standard practice ignores the correlation between outcome variables, implicitly assuming that observable errors on auxiliary variables are completely uninformative about errors on other variables. Relative to this baseline, our approach is likely to improve inferences even when the assumption does not hold exactly.

We demonstrate the utility of this approach using two examples. First, we perform a validation exercise using pre-election polling for the 2022 midterm election in Michigan. We jointly model self-reported vote choice for Governor, Secretary of State, and Proposition 3, a ballot proposition on reproductive rights, alongside several other questions of interest. We apply our approach to calibrate the survey to Governor results at the county level. This adjustment greatly improves the accuracy of county-level estimates for Secretary of State and Proposition 3, reducing error by around two-thirds. Additionally calibrating the estimates to Secretary of State results further improves the accuracy of Proposition 3 estimates, but only modestly. The error reduction occurs through reduction in both bias and variance of estimates.

After validating the method using outcomes with ground-truth data, we then examine the impact of calibration on other quantities of interest, such as partisanship in the county. The calibrated partisan estimates are quite different that the raw survey averages, and exhibit some notable differences compared to the baseline MRP results.

Next, we then show how the calibration can be used more generally. We study partisan animus, which we operationalize as survey respondents who say that out-partisans are "a danger to our country" who "must be defeated at any cost." Using self-reported 2020 presidential vote as the auxiliary variable for calibration, we produce county-level estimates of animus toward Democrats and Republicans and show how they relate to political geography. First, we find that calibration slightly reduces population-level estimates of animus — consistent with survey respondents being hyperpartisan. Second, we show that calibration substantially affects the distribution of *county-level* animus. Third, we find that there is more anti-Democratic animus than anti-Republican animus. Finally, we find that animus is especially high in places dominated by one party. Rather than animus being most common in places with strong partisan competition, it is most prevalent in politically homogenous areas.

Substantively, our method will help researchers generate improved estimates of public opinion using MRP. The ability of MRP to estimates of public opinion in small subsets of the population has led to its wide adoption across subfields of political science and related disciplines. These estimates have been invaluable for assessing foundational questions of representation and accountability. They allow scholars to measure the connection between district opinion, elite actions, and lawmaking outcomes. The method has been used to study public opinion within states, congressional districts, and cities (Gelman, 2009; Lax and Phillips, 2012; Clinton, 2006; Levendusky, Pope and Jackman, 2008; Tausanovitch and Warshaw, 2013; Simonovits and Payson, 2023; Toshkov, 2015; Lipps and Schraff, 2019). Other work extends the analysis further, to study subgroup opinion and behavior within electoral districts (Ghitza and Gelman, 2013; Kuriwaki et al., 2023) or to generate time series estimates of public opinion (Caughey and Warshaw, 2022). Our method enables researchers to generate estimates of subgroup issue opinion that account for observed election results or other ground-truth data, thereby adjusting for non-ignorable survey error.

Methodologically, our paper builds on methods for calibrating model-based inferences to known population quantities. Research in the MRP paradigm commonly uses a correction, sometimes called the "logit shift," to account for discrepancies between known geographic aggregates and model-based estimates (Ghitza and Gelman, 2013; Rosenman, McCartan and Olivella, 2023). This method can improve estimates for items for which ground-truth data is available, but not at the level of aggregation that is of primary interest. For example, Kuriwaki et al. (2023) apply the logit shift to their model-based estimates of election results within congressional districts. This adjustment helps improve inference on their ultimate quantity of interest — the distibution of vote choice by racial group within each congressional district. We extend this technique by proposing a method to use the calibration error to update separate survey items for which no ground-truth data — at any level of aggregation — is available.

Our work is also related to a different strand of the MRP literature, which focuses on expanding the number of variables that are available for poststratification. Traditionally, researchers conduct poststratification by obtaining information on joint distribution of demographic variables from the Census. However, this requirement is limiting, as there are many important predictors of public opinion (such as partisanship) whose population joint distribution with other variables is not known. Thus, researchers have proposed methods to combine administrative data with data from surveys, voter files, and other sources to generate synthetic poststratification tables (Leemann and Wasserfallen, 2017). These augmented poststratification tables enable researchers to fit more flexible and accurate models of survey responses. Our approach sidesteps the need to estimate this joint distribution separately. The data requirement for our estimator is that the marginal distribution of the auxiliary variables is known at some level of aggregation. We use the survey to pool information between outcomes, using the observed error in auxiliary variables to inform the estimated error in other outcomes.[1]

The paper proceeds as follows. In Section 1, we provide an overview of MRP and describe existing calibration methods. In Section 2, we present our method, which generalizes calibration methods to the case of multiple outcomes, some of which have no ground truth data available. This section also contains details on implementation and computation. We then turn to two empirical applications. Section 3 briefly describes the survey data we use. Section 4 presents our validation exercise using 2022 pre-election polling in Michigan. In Section 5, we apply our method to the study of partisan animus. Finally, we conclude in Section 6.

[1] This approach also shares similarities with James-Stein estimators, as it models correlation between outcomes — even after adjusting for individual- and geographic-level controls — to improve inference across outcomes. While we focus on how to use the model for calibration, even without any calibration jointly modeling outcomes is likely to reduce total error.

1 Small-Area Estimation Using Multilevel Regression and Poststratification

Our method extends the standard MRP modeling approach to estimating opinion in small geographies. In this section, we provide a brief conceptual overview of MRP, without delving into specific modeling choices. We then review one popular approach to calibrating the results to some known population margins, such as election outcomes. This method ensures that model-predicted outcomes are consistent with the ground truth data that is available, but leaves estimates on other oufcomes unchanged. Thus, in the next section, we extend this basic model to include multiple outcomes and propose a principled method for adjusting all of the estimates to account for known population margins — even those outcomes for which we do not have any ground-truth data.

1.1 MRP Overview

MRP commonly involves three steps. In the first step, the researcher estimates an individual-level regression model of an outcome of interest (e.g., opinion, behavior) using survey data. The regression should include individual-level and geographic-level covariates that help explain the outcome and whose joint distribution is known for the geography of interest. Second, the researcher uses the resulting statistical model to predict the expected outcome for each combination of covariates in the known joint distribution (i.e., each poststratification cell). Third, the researcher produces an aggregate estimate for the geography of interest by taking a weighted averages of the expected outcome in each of the poststratification cells. The weights in this step are proportional to the number of people in each cell in the population of interest.[2]

To motivate and explain how we can leverage the known truth of some outcomes to account for non-ignorable non-response in outcomes whose truths are unknown, it is useful to make the process more precise. Suppose we are interested in public opinion (or behavior) on J outcomes. Commonly, the list of opinions and behavior of interest includes vote choice, policy opinions, overall ideology, and various forms of political participation and engagement. For expositional simplicity, assume that the outcomes are binary and public opinion can therefore be summarized using the percentage of the population who hold that opinion. Denote this population-level quantity of outcome j in the geography of interest by θ^j.[3] For

[2] We will often refer to the subgroup of interest as a "geography," following work employing MRP to estimate opinion in sub-national political units. It is important to note, however, that MRP can be applied just as easily to other subgroups, such as those defined by race or age, so long as their joint distribution with other variables in the regression is known.

[3] In our notation, superscripts index issues and subscripts indicate groups or individuals.

example, θ^j might represent the share of the public who vote in an election, or who vote early, or who vote for the Democrat.

We can partition the population into a set of cells $\mathcal{C}$, with varying sizes N_c. Each cell c is defined by a combination of demographic variables and geographies. For example, $\mathcal{C}$ might represent every combination of age, race, gender, education, and county. We assume that the the cell sizes in the population N_c are known from Census data or other sources.[4] Given this decomposition, the total population is simply the sum of the population in the cells: $N \equiv \sum_c N_c$. The combination of cell-level covariates and the cell sizes N_c form the poststratification table.

Additionally, because the cells are defined in part by geography, we can define a list of geographies g, which are made up of the set of cells $\mathcal{C}_g$ belonging to each geography. For example, $\mathcal{C}_{\text{Calif.}}$ would represent the subset of poststratification cells that are in California.

Given this setup, overall opinion in the population is simply the weighted average of opinion in each cell:

$$\theta^j \equiv \frac{\sum N_c \theta_c^j}{\sum N_c}. \tag{1}$$

Similarly, opinion in any geography g is the weighted average of opinion in all cells belonging to geography g: $\theta_g^j \equiv \frac{\sum_{c \in \mathcal{C}_g} N_c \theta_c^j}{\sum_{c \in \mathcal{C}_g} N_c}$.

MRP begins by specifying a statistical model for outcome j using the covariates that define the postratification table. This model is constrained by the necessity of using covariates whose joint distribution are known (or can be estimated). The model usually also contains geographic predictors to account for variation in opinions across geographies, conditional on individual-level covariates. The researcher estimates the model using survey data. Then, using the estimated model, the researcher predicts the probability of responses to each item for each cell in the poststratification table, thereby generating a set of estimates $\hat{\theta}_c^j$.

After generating estimates of opinion within each cell, the next step is to poststratify those estimates to the (sub)population of interest using the cell sizes N_c. An estimate of population-level opinion can be generated by summing over all cells in the population:

$$\hat{\theta}^j = \frac{\sum N_c \hat{\theta}_c^j}{\sum N_c}. \tag{2}$$

Estimates for particular geographic areas can be generated by summing over the cells within

[4] As discussed in the introduction, when researchers do not have information on the full joint distribution, there are several methods available to estimate the joint distribution. Key ideas in the literature include combining marginal distributions by assuming independence, using survey data to estimate the joint distribution, or combining these two approaches (Leemann and Wasserfallen, 2017).

a particular geographic: $\hat{\theta}^j_c = \frac{\sum_{c \in \mathcal{C}_g} N_c \hat{\theta}^j_c}{\sum_{c \in \mathcal{C}_g} N_c}$. Because the joint distribution of the variables within each geography is assumed known (via knowledge of N_c), this poststratification step ensures that the estimated opinion for each geography reflects the joint distribution of the demographics in that geography.

This procedure enables researchers to estimate opinion and behavior in small levels of geography that were not intentionally sampled. For example, in a study of representation, researchers might want to know whether policy outcomes accord with public opinion (e.g Lax and Phillips, 2012; Tausanovitch and Warshaw, 2014). However, even very large surveys are unlikely to have enough survey respondents in any given state or congressional district to generate reliable estimates of public opinion using traditional methods. MRP provides a solution by sharing information across units of geography and demographic groups — via the regression step — and accounting for differences in population composition across geographies — via the poststratification step.

Researchers often use multilevel regression models to predict opinion within each cells, but there are increasingly many estimators available to generate these predictions.[5] While our proposed method will leverage the structure of multilevel regression models, we defer discussion of modeling details until the next section. For the remainder of this section, it suffices to assume we have estimates of cell-level opinion $\hat{\theta}^j_c$.

1.2 Adjusting Estimates to Account for Known Population Margins

MRP naturally ensures that inferences take into account discrepancies in demographic variables between the survey and the target population, via the poststratification step. The quality of inference in MRP therefore depends critically on the quality of the model-based estimates of opinion within each cell. Methodological research has thus focused on proposing flexible outcome models with rich covariate data. Yet, error may remain if survey participation is correlated with attitudes, even after adjusting for covariates.

For some outcomes, researchers can empirically assess error by comparing model-based estimates to ground truth data, usually measured at some geographic aggregate. Election results are the canonical example. While there is no ground-truth data on the relationship between demographics and vote choice, we do observe ground-truth election outcomes. Even if predicting these aggregates is not the main point of an analysis, discrepancies between a model's estimates and the ground-truth data are evidence of error. Adjusting for this error

[5] Analysts can choose from any model that can generate predictions for every cell in the postratification table. As the name MRP reveals, early research proposed using multilevel regression to model the outcome (Park, Gelman and Bafumi, 2004). Other flexible machine learning methods have also been proposed in the literature (Bisbee, 2019; Broniecki, Leemann and Wuest, 2022; Ornstein, 2020). These methods may generate better predictions in-sample at the expense of less interpretability (though see Goplerud, 2023).

can improve inferences on other aggregates — for example, estimates of vote choice by race (Kuriwaki et al., 2023). There are several approaches to incorporating this information.

First, in a classical survey weighting context, the known population margins can be used as targets in generating survey weights.[6] These techniques will ensure that the weighted sample match the ground-truth data exactly. The drawback of these methods is that, unlike poststratification, they can account for only a small number of interactions between weighting variables. They require at least one survey respondent in every weighting cell — meaning relatively few variables can be included to define the weighting cells. This requirement is problematic if researchers want to study outcomes in small geographies, as there typically will be few or no respondents in small levels of geography — meaning there are no feasible survey weights to match ground-truth data measured at this level of aggregation.[7]

A second approach is to include the known geographic aggregates as predictors in the regression step of MRP. For example, individual vote choice could be predicted as a function of individual-level covariates and the election results in a respondent's state. Indeed, it is best practice to include such geographic-level predictors in MRP (Buttice and Highton, 2013), as their inclusion will typically improve the model fit by sharing information across states. However, if there is nonignorable error in the survey, better model fit does not necessarily improve predictions of ground-truth data.

We focus on a third approach, also in the MRP paradigm. This adjustment adds a geography-specific intercept shift to the modeled probabilities in each cell of the poststratification table. The value of the intercept shift is chosen to ensure that the model-implied vote shares exactly match the known population vote shares in each geographic unit (Ghitza and Gelman, 2013, 769). This method approximates the posterior distribution of cell-level probabilities after conditioning on the ground-truth data Rosenman, McCartan and Olivella (2023). Because the intercept adjustment is applied on the logit scale, this method has sometimes been called the "logit shift."

To make the logit shift precise, suppose we observe the population-level outcome θ_g^j for some geography g. Given model estimates $\hat{\theta}_g^j$, we can calibrate results by adding an intercept shift on the logit scale to the predicted probabilities of cells in geography g to ensure that

[6] For example, researchers could generate weights via raking that target election outcomes along with demographic variables. This method is used by the National Exit Poll to ensure the exit poll vote share margins match the outcomes. Increasingly many polling firms are also including respondents' vote in the last election as a weighting target.

[7] Recent research has proposed methods to generate weights that match first-order margins exactly — as in traditional raking — while approximately matching higher-order interactions (e.g. Ben-Michael, Feller and Hartman, 2023). These methods are a compromise between raking on marginal variables alone and full poststratification, thereby reducing bias in estimates relative to raking on margins alone. Yet, even these more flexible methods require some representation of every level of a weighting variable, making it infeasible to weight to (say) county-level election results.

the model estimates match the actual results.

Recall that the set of poststratification cells in geography g is denoted $\mathcal{C}_g$. Formally, the logit shift parameter for geography g on outcome j is the value of δ_g^j that solves the equation

$$\theta_g^j = \underbrace{\frac{1}{N_g} \sum_{c \in \mathcal{C}_g} N_c \overbrace{\text{logit}^{-1}\left(\text{logit}(\hat{\theta}_c^j) + \delta_{g[c]}^j\right)}^{\text{updated estimate for cell } c}}_{\text{updated estimate for geography } g}, \tag{3}$$

where $N_g \equiv \sum_{c \in \mathcal{C}_g} N_c$ is the population of geogaphy g and the notation $g[c]$ indicates the geography for cell c. This expression equates the true population outcome, θ_g^j, with the model-implied outcome after adjusting by an offset δ_g^j.

After solving for δ_g, new cell-level probabilities are generated by applying the geographic offset to the existing probabilities:

$$\tilde{\theta}_c^j = \text{logit}^{-1}\left(\text{logit}(\hat{\theta}_c^j) + \delta_{g[c]}^j\right). \tag{4}$$

The updated probabilities can then be used in downstream analysis. When these updated estimates are poststratified to the geographic level, the estimates match the ground-truth data by construction. When poststratified to some other subgroup, the results will differ from the uncalibrated MRP results whenever the logit shift parameter varies according to the demographic makeup of the geographies.

In the racially polarized voting example, analysts could generate poststratification-based estimates of voting by racial group within a state after applying the logit shift. If the value of the logit shift parameter varies according to the racial composition of the county, then the calibrated estimates of vote share by race will differ from the uncalibrated estimates.

The logit shift is an intuitively appealing method for incorporating known population margins into the MRP workflow. It provides a minimal adjustment that does not distinguish between different individuals within the same geography — consistent with the aggregate nature of the data available. Additionally, the logit shift adjusts each geography independently, so an update to one county is not affected by updates to other counties. Finally, it is flexible enough to incorporate observations of ground-truth data at very fine levels of geography.[8]

Of course, this calibration procedure only improves estimates under certain assumptions. First, the shift is uniform across poststratification cells — because it the precise source of

[8] For example, a report by the Democratic data firm Catalist on the 2020 presidential election uses this method to calibrate estimates to precinct-level results (Ghitza and Robinson, 2021).

the error is unknown, when applying the shift it must be applied uniformly across cells (i.e., every poststratification cell c in g is shifted by δ_g). If the true cell-level error is not uniform (on the logit scale) within geographies, it is possible the calibration will actually increase the error for some cells, even as it decreases the average error across cells. This could occur if the model overestimates the outcome probability for some cells and underestimates it for others. Second, the level of aggregation at which the calibration is performed influences its accuracy. Logit shift calibration performs better when geographies are more homogenous in their true probabilities across cells. Given patterns of residential segregation along political and racial lines in the U.S., this implies that the logit shift will work better when applied at smaller geographies (e.g. Rodden, 2019).[9]

2 Updating Estimates on Multiple Issues

In principle, discrepancies between model-based predictions and ground-truth data should reveal important information about the nature of the survey-based error across a range of outcomes. The logit shift provides a way to update estimates on a single outcome in light of the discrepancies between model-based estimates and ground-truth data *for the same outcome*. Yet, existing work does not consider how to incorporate the information contained in the logit shift parameters to update inferences on *other* outcomes for which ground-truth data are not available.

We propose an extension to the logit shift correction to multiple outcomes. We assume that we observe ground truth information about a subset of the outcomes and can use the logit shift to calibrate the model-based estimates for these outcomes. To update the model estimates for the other outcomes, we use the survey to estimate the correlation in opinions across issues within counties. We then use that correlation to estimate a logit shift to apply to those outcomes for which we do not have calibration data. In so doing, we treat the logit shift that results from calibrating the outcomes to ground-truth data as the *realized* geographic error for the calibrated outcome, which we then use to adjust outcomes whose truth is unknown based on the correlations in outcomes reflected in the county-level intercepts.

To fix ideas, consider the task of estimating subgroup opinion on two policy attitudes: increasing top income taxes and providing local tax incentives to new businesses. Suppose we have a survey measuring respondents' preferences on these two issues, along with their vote choice in the last presidential election. We can use the logit shift to calibrate the vote choice outcome.

[9] For more discussion on these points, refer to Section 3 of Rosenman, McCartan and Olivella (2023).

The discrepancy between the survey-based prediction of vote shares and the actual election outcome should also be informative about errors in survey-based estimates of opinion on income taxes and business incentives. But attitudes on these two issues are likely not equally correlated with presidential vote choice. Democratic and Republican presidential candidates tend to be sharply differentiated on the issue of income tax policy, so there is likely to be a high correlation between respondents' opinions on that issue and their vote choice. In contrast, there is little partisan disagreement over local business incentives (Jensen et al., 2020), meaning that attitudes on this issue are likely less correlated with vote choice. The survey error in presidential vote choice is therefore likely to be more similar to the survey error in income tax preferences than to the survey error in local business incentives. Our method formalizes this intuition and provides a principled, data-driven method for estimating this correlation and adjusting survey estimates in light of calibration errors on auxiliary variables with ground-truth data.

The proposed method proceeds as follows. First, we specify a joint model for all outcomes, including both the outcomes of interest (for which there is no ground-truth data) and auxiliary variables (for which there is ground-truth data). The model estimates county-level random effects for each of the outcomes as well as the correlation in these random effects across outcomes. Next, we calibrate each of the auxiliary variables independently using the logit shift method outlined above. Rather than viewing the logit shift as an ad-hoc adjustment, we instead interpret the sum of logit shift parameter and the estimated random effect as the *realized* random effects for the calibrated outcomes. This interpretation suggests a principled way to update the uncalibrated variables: we condition on the realized random effects and compute the conditional expectation of the random effects for the uncalibrated variables. Given joint normality of the random effects, this conditional expectation is a simple linear function of the calibrated logit shifts and the covariance between random effects across outcomes. Finally, we generate updated estimates using the *predicted* random effects for the uncalibrated outcomes.

2.1 Multivariate Outcome Model

The basic setup and notation follow those in the previous section. We are interested in public opinion on J issues, with the estimate for issue j denoted θ^j. To estimate opinion as a function of the covariates that define the poststratification cells, we fit a multivariate logistic regression model to binary measures based on the survey responses. This results in a measure of respondent i's response related to the binary outcome j being given by $y_i^j \in \{0, 1\}$. To model the observed responses, we use a multivariate response model for the survey options where response probabilities are a function individual-level demographics and geographic-level characteristics. In addition, we include a random intercept for each geography g to account for remaining between-geography variation conditional on the statistical controls being used. The basic model is therefore:

$$
\begin{aligned}
y_i^j &\sim \text{Bernoulli}(\theta_i^j) \\
\theta_i^j &= \text{logit}^{-1}\left(\alpha_{g[i]}^j + \beta^j X_i + \gamma^j Z_{g[i]}\right).
\end{aligned}
\tag{5}
$$

In Equation 5, X_i is a vector of individual-level covariates (including an intercept), $Z_{g[i]}$ is a vector of county-level covariates, and $\alpha_{g[i]}$ is a random intercept that varies at the county level.[10] The county random effects have a hierarchical structure, which facilitates partial pooling across counties (Gelman and Hill, 2007).

To model dependency across outcomes, we model the correlation in county intercepts across different outcomes. Specifically, we specify a hierarchical multivariate normal distribution for the county random intercepts:

$$
(\gamma_g^1, \gamma_g^2, \dots, \gamma_g^J) \sim \text{MultivariateNormal}(\mathbf{0}, \mathbf{\Sigma}) \tag{6}
$$

The covariance matrix $\mathbf{\Sigma}$, which is estimated from the data, encodes information about the correlation in outcomes within geography, after accounting for the other terms in the model.[11] We will use this covariance matrix to compute the predicted logit shifts for outcomes without ground-truth data.

2.2 Multivariate Logit Shift

We leverage the model structure above to provide an re-interpretation of the logit shift as the residual between the estimated county random effect α_c^j and the *realized* county random effect. To motivate this interpretation, substitute the model for θ_c^j given in Equation 5 into the expression for updating cell-level predicted probabilities (Equation 4):

$$
\begin{aligned}
\theta_c^{j*} &= \text{logit}^{-1}\left(\text{logit}(\hat{\theta}_c^j) + \delta_{g[c]}^j\right) \\
&= (\alpha_{g[c]}^j + \delta_{g[c]}^j) + \beta^j X_c + \gamma^j Z_{g[c]}
\end{aligned}
$$

[10] In our application below, we include additional random effects for some demographic groups. For expositional clarity, we omit those here.

[11] Even without calibrating to ground-truth data, the joint outcome model is likely to improve overall inferences due to the regularization provided by this hierarchical structure, relative to fitting separate models for each outcome.

The updated probabilities are constructed such that the model-implied geographic aggregates match the observed ground-truth geographic aggregates. Thus, one way to interpret the logit shift parameters δ_g^j is as the *residual* between the estimated county-level intercept, α_g^j, and the realized county-level intercept. The payoff of viewing the logit shift from this perspective is that it suggests a method for predicting the logit shift parameters for the unobserved outcomes. Given the parametric structure of the random effects, we can condition on the *realized* random effects for the calibrated outcomes to compute the conditional expectation of the unobserved random effects.

In the model specified above, the random effects have a convenient multivariate normal distribution across outcomes. Collecting the random effects and logit shift parameters into vectors for observed outcomes (subscripted $\mathbf{o}$) and unobserved outcomes (subscripted $\mathbf{u}$), this conditional expectation has the following simple linear form (Eaton, 2007, 116):

$$E[\alpha_c^{\mathbf{u}} + \delta_c^{\mathbf{u}} \mid \alpha_c^{\mathbf{o}}, \delta_c^{\mathbf{o}}] = E[\alpha_c^{\mathbf{u}} + \delta_c^{\mathbf{u}}] + \Sigma_{\mathbf{uo}}\Sigma_{\mathbf{oo}}^{-1}\left((\alpha_c^{\mathbf{o}} + \delta_c^{\mathbf{o}}) - E[\alpha_c^{\mathbf{o}} + \delta_c^{\mathbf{o}}]\right), \tag{7}$$

where the expectations are taken over the joint distribution of the variables. In this equation $\Sigma_{\mathbf{uo}}$ is the cross-covariance in random effects between the unobserved and observed elements and $\Sigma_{\mathbf{oo}}^{-1}$ is the inverse of the covariance of the observed elements. This covariance matrix is estimated from the survey data when estimating Equation 5.

Recalling that the unconditional expectation of the random effects is 0, we arrive at an estimator for $\delta_c^{\mathbf{u}}$, the logit shifts for outcomes without calibration data:

$$\hat{\delta}_c^{\mathbf{u}} = \Sigma_{\mathbf{uo}}\Sigma_{\mathbf{oo}}^{-1}\delta_c^{\mathbf{o}}. \tag{8}$$

This estimator predicts the logit shift for each outcome without calibration data as a linear combination of the logit shifts for the outcomes with calibration data. The weights in this linear combination depend on the covariance in the geographic random effects across outcomes, with higher weight going to the logit shift on outcomes with higher covariance.

To see the result more clearly, consider the case of observing a single outcome. In this case, the predicted logit shift for unobserved outcome j', conditional on the observed logit shift for outcome j, reduces to to $\hat{\delta}_c^j = \rho_{jj'}\frac{\sigma_{j'}}{\sigma_j}\delta_c^j$, where $\rho_{jj'}$ is the correlation in random effects between outcomes j and j' and σ_j, $\sigma_{j'}$ are the standard deviations of the respective random effects. The correlation coefficient controls the magnitude of the shift, while the ratio of standard deviations simply corrects for the different scales between outcomes j and j'. If there is a perfect correlation in county-level random effects between outcomes j and j' — i.e., $\rho_{jj'} = 1$ — then the logit shift for the unobserved outcome will be essentially the same as the logit shift for the observed outcome (up to a difference in scale). If there is no correlation in county-level random effects — i.e., $\rho_{jj'} = 0$ — then the logit shift for the observed outcome contains no information about the logit shift for the unobserved outcome.

2.3 Assumptions and Limitations

This method relies on the important assumption that our survey-based estimate of $\mathbf{\Sigma}$ reflects the population-level correlation. Put differently, we assume that the survey might contain error in the proportion of people who hold some opinion, but we assume that the survey accurately captures the correlation between opinions. This assumption need not hold: after all, the whole point of the method is to adjust for surveys that generate incorrect estimates of known population quantities. Survey respondents tend to be highly educated and more politically interested than the public writ large, meaning their attitudes may be more correlated across issues than those of the general public (Marble and Tyler, 2022; Freeder, Lenz and Turney, 2018).

However, the alternative to this assumption is to not update estimates at all. By not updating estimates, researchers are implicitly assuming that there is no correlation at all between issues — which is almost certainly an even worse assumption.[12]

Another limitation of the method to note is that we use the discrepancy between model-based and ground-truth estimates to update the *expectation* of the county-level random effects, but we do not incorporate this information to update the *covariance* between them. An interesting extension would be to take the estimated covariance matrix $\mathbf{\Sigma}$, update it using the observed discrepancies to generate a "calibrated" $\tilde{\mathbf{\Sigma}}$, then use the updated covariance matrix to predict the unobserved logit shiffts in Equation 8. This would partially correct for the assumption that the correlation in the survey and population are equal.

2.4 Estimation

Analysts have several options available to implement this method. One possibility is to use plug-in estimators for Σ after estimating Equation (5). For example, in a frequentist framework one could use the maximum likelihood estimator for Σ, or in a Bayesian framework one could use the posterior mean.

In our applications below, we adopt a fully Bayesian framework. As a result, for each draw of the posterior distribution of the outcome model being fit, we calculate the logit shifts for the calibrated outcomes and then use them to predict the implied logit shifts for the uncalibrated outcomes. By generating estimates for each iteration, we are then able to directly account for the uncertainty are arises in the estimation of Σ.

[12] Future work could probe this assumption more formally. For example, it would be useful to compare the correlation in vote choices in public opinion polling with ground-truth data gleaned from cast-vote records.

Assuming we have M draws from the posterior distribution of the parameters, we apply the method to each draw (m) separately. Specifically, for each draw (m) we calculate the following quantities:

- $\hat{\theta}_c^{j(m)}$ and $\hat{\theta}_g^{j(m)}$: cell-level and geographic-level estimates of opinion on each issue j for all cells c and geographics g
- $\delta_g^{j\mathbf{o}(m)}$: the logit shift parameters for the observed outcomes, which ensures model-based predictions for these outcomes the ground-truth data
- $\delta_g^{j\mathbf{u}(m)}$: the estimated logit shift parameters for the unobserved outcomes, calculated from Equation 8 using the covariance parameters from this draw, $\Sigma^{(m)}$.
- $\tilde{\theta}_g c^{j(m)}$: updated cell-level probabilities
- Finally, we generate updated subgroup estimates by poststratifying $\tilde{\theta}_c^{j(m)}$ to the population of interest

By doing these steps once for each draw from the posterior, we generate a calibrated posterior distribution for the quantities of interest.

3 Data

For our empirical applications, we rely on data from an original survey conducted via SurveyMonkey throughout the fall of 2022.[13] Respondents were recruited via a river sample, whereby respondents who completed any survey on the SurveyMonkey platform during this period were invited to take our another survey. Respondents who opted to participate were then asked a survey we wrote which asked a series of questions related to the 2022 midterm elections as well as standard demographics and the county they live in.

We use several variables to model opinion based on conventional weighting practices and theories about the nature of survey non-response. While it is certainly possible to extend the covariates being used — or to employ recent methods to determine the optimal set of covariates using cross-validation or other methods — we rely on typical covariates to demonstrate the contributions of our method relative to the most commonly used approach.

To model individual-level variation in survey responses to question j from individual i we create a series of indicator variables from self-reported demographics that include: age (computed based on their reported birth year), race, gender, education level, and county.

[13] Because no sensitive information was collected, the University of Pennsylvania IRB deemed the survey exempt from IRB review (Protocol #852133).

We recoded the continuous age variable into discrete bins containing the ages of: 18-29, 30-39, 40-49, 50-64, 65-74, 75+. To measure race we used indicators for: Black, White, Hispanic/Latino, Asian/Pacific Islander, and a residual Other. While there is continuing conversations about the relationship between racial and ethnic categories, we follow the practices used by the National Exit Poll and we classify all people who indicate Hispanic/Latino ethnicity as Hispanic/Latino regardless of their race. We use four indicators to measure education attainment based on: high school degree or less, some college or an associates degree, college graduate, and postgraduate degree.

To account for unmodeled variation related to geography — and also to pool information across geographies — we map respondents into counties based on their residential zip code. Although the most zip codes are located within a single county, when zip codes span multiple counties we assign respondents to the county containing the highest fraction of zip code residents based on population tables (Missouri Census Data Center, 2022). So doing inevitably misclassifies some respondents, but the magnitude and impact of matching errors is trivial — over 97% of people can be uniquely classified into a single county and only 3% of the respondents live in a zip code spanning more than one county.

To generate the poststratification table, we use microdata from the 2020 American Community Survey 5-year data files. For each county, we count the number of people in every possible demographic combination using: age × race × gender × education for a total of $6 \times 5 \times 2 \times 4 = 240$ unique cells within each county. There is some error in measuring the number of county residents in each of the 240 demographic groupings because the smallest level of geography available in the Census microdata is a unit called a Public Use Microdata Area (PUMA) consisting of non-overlapping partitions with no fewer than 100,000 people. To map between PUMAs and counties, we use the share of the population within each PUMA that lives in each county according to (Missouri Census Data Center, 2022) to weight the Census microdata and generate the required population estimates. This assumes that respondents' demographics are independent of the county they live in conditional on the PUMA they live in. While this process introduces some error into the county-level estimates — especially in counties with populations less than 100,000 — the magnitude of this error is likely small relative to other sources of error.

4 Validation: 2022 Michigan Elections

To evaluate the performance of the method we propose we examine pre-election polling in Michigan leading up to the 2022 midterm election. The 2022 Michigan election was interesting because it involved three statewide election contests: the gubernatorial election between incumbent Democrat Gretchen Whitmer and Republican candidate, election-denier, and TV newscaster Tudor Dixon; the Secretary of State election between incumbent Democrat Jocelyn Benson and election-denier Republican Kristina Karamo (who was subsequently elected to Chair the Michigan GOP), and a proposed constitutional amendment to protect reproductive rights and the legal access to abortion (Proposition 3). Going into the election, the RealClearPolitics polling average had Whitmer ahead by 1.0 percentage point and the state was projected by RealClearPolitics as a "GOP Pick Up" based on the fact that past polling errors in Michigan favored Democrats by 2.2 percentage points on average.[14] In actuality, Whitmer won re-election by 10.6 percentage points and the Democrats gained control of statewide government for the first time since 1983.

Michigan is an appropriate and excellent case to examine not only because it is a state where the pre-election polling averages were again mistaken in 2022 after making similarly large misses in both 2020 (Clinton et al., 2021) and 2016 (Kennedy et al., 2018), but the fact that there were three statewide races being held means that we know the actual county-level outcomes for three different outcome measures. Knowing the actual county-level vote margins for multiple races allows us to not only assess the accuracy of calibrated estimates compared to known outcomes, but we can it also quantify the increase in accuracy when calibrating using multiple measures. Michigan is also a good state to use because there are no state-sourced measures of partisanship and the imputed measures of partisanship used by commercial voter file companies are notoriously limited – as Clinon and Trussler (2022) show, fewer than 50% of respondents to the 2020 National Exit Poll self-identified with the party they were imputed as being most likely to belong to. Large errors in the imputation of partisanship in voter files means it is extremely problematic to use voter-file based measures of partisanship to poststratify — which is one reason why pre-election polls continue to perform relatively poorly in Michigan.[15]

To demonstrate and evaluate the calibration we propose, we jointly modeling vote choice for the three elections alongside several other outcomes of possible interest (e.g., partisanship). We first generate uncalibrated MRP estimates by poststratifying based on the included demographic variables and the joint distribution of county demographics constructed using the ACS microlevel data as described above. We then use the certified gubernatorial results to calibrate the MRP estimates to match the observed county-level results and we use the

[14] https://www.realclearpolitics.com/epolls/2022/governor/mi/michigan_governor_dixon_vs_whitmer-7545.html

[15] Using voter-file measures rather than Census-based measures to create the poststratification table also creates issues because of known issues in the imputation of other measures (e.g., non-state sourced measures for race and ethnicity) as well as the need to rely on the accuracy of the voter file for characterizing the electorate.

implied adjustment to adjust the county-level MRP estimates for the other outcomes. Comparing the calibrated and uncalibrated estimates to the known outcome for the other two elections (i.e., the Secretary of State results and the vote on Proposition 3) identifies the effect of the proposed calibration shift. As we show, the calibration results in a substantial reduction in error: the average absolute error in county-level estimates for Secretary of State and the abortion proposition are reduced by 65% and 59%, respectively.

To conduct our analysis we use the 2,504 SurveyMonkey respondents living in Michigan who answered all demographic and outcome variable questions. The outcomes of interest that we jointly model include the three vote choice questions mentioned above as well as responses to other questions: (1) Biden job approval; (2) whether Joe Biden legitimately won the 2020 election; (3) whether elections are being conducted fairly; and (4) partisan identification.[16] We use several measures to show how the approach can be used to not only characterize the geographic distribution and concentration of outcomes of possible interest (e.g., opinions related to election outcomes), but also to illustrate how we can use the approach to model characteristics that might be instrumentally useful in adjusting future surveys (e.g., partisanship).

We model the outcomes as a function of both respondent-level and county-level predictors. The individual-level predictors we use include variables whose joint distribution is known for each county from Census data. These include: age, race, gender, education, with interactions between race × education and also nonwhite × education × age.[17] To account for other sources of between-county variation we also include county-level predictors for: the share of the county that is nonwhite, the share that is Hispanic/Latino, the share with a college degree, county median income, and 2020 two-party vote share for Biden. Finally, we also include county-level random intercepts that we allow to be correlated across outcomes and which provide the basis for our proposed adjustment method.

Because we estimate the model in a Bayesian framework, the model is completed by the assignment of priors for each of the estimated parameters. All coefficients are assumed to have a very diffuse $\text{Normal}(0, 5)$ prior. All county-level random effects are assumed to have multivariate normal priors centered at 0 — i.e., we assume no error on average — and we decompose the covariance matrix for the county-level random effects into a correlation matrix and a vector of standard deviations. We assume that the random effect correlations have

[16] Currently, the model can only accommodate binary outcomes, so we split party ID outcome into three binary outcomes (Democratic, Republican, Independent).

[17] We model age as a linear function of the category (treating the lowest category, 18-29, as 1, the second category, 30-39, as 2, and so on) as well as a random intercept for each age group (Ghitza and Gelman, 2013). This has the effect of partially pooling the age effects toward the regression line. The interactions between demographic variables are modeled as random intercepts that are independent across outcomes.

Table 1: Correlation of County Intercepts Across Outcomes, Michigan 2022

	Governor	Secretary	Abortion	Biden	Biden	Fair	Democratic	Independent	Republican
		of State	Prop.	Legitimate	Approval	Elections	PID	PID	PID
Governor	1.00								
Secretary of State	0.69	1.00							
Abortion Prop.	0.69	0.69	1.00						
Biden Legitimate	0.68	0.67	0.69	1.00					
Biden Approval	0.68	0.66	0.68	0.69	1.00				
Fair Elections	0.60	0.59	0.61	0.66	0.61	1.00			
Democratic PID	0.61	0.59	0.61	0.61	0.61	0.56	1.00		
Independent PID	-0.06	-0.01	-0.04	-0.06	-0.06	-0.10	-0.23	1.00	
Republican PID	-0.49	-0.52	-0.51	-0.47	-0.48	-0.41	-0.38	-0.22	1

Notes: Posterior mean correlations between county random intercepts across survey outcomes.

LKJ(1) priors (Lewandowski, Kurowicka and Joe, 2009) which means that all correlation matrices have an equal prior probability. The standard deviation for all random effects are given by half-Student-t priors with 3 degrees of freedom, location parameter of 0, and scale parameter of 2.5.[18] We fit the model using the `R` package `brms` (Bürkner, 2017), which provides an interface to the Stan modeling language (Carpenter et al., 2017).[19]

4.1 Validation Results

To begin, we report the correlations between county-level random intercepts across outcomes in Table 1. This correlation matrix should be interpreted as the residual correlation in responses within counties (on the logit scale) *after accounting for county-level demographics used in the regression specification.*[20] These correlations are the primary information used to calibrate the estimates for responses lacking known distributions.

As Table 1 reveals, there is a relatively high correlation between the county intercepts across the jointly modeled outcomes. These high correlations are to be expected given the well-known impact of partisanship on opinions and the nature of contemporary politics; the high correlations reveal that responses are associated with one another within counties even after accounting for associations that are related to the other statistical controls we employ. The sole exception to this pattern is Independent party identification, which is also substantively meaningful as it reveals that the percentage of respondents who self-identify as an independent within a county is largely unrelated to other opinions in that county.

[18] See Stan Development Team (2023) for details about the parameterization of the Student-t distribution.

[19] We run 4 chains for 600 post-warmup iterations each, for a total of 2,400 draws from the posterior distribution. All chains appear to have converged to the posterior distribution based on Gelman-Rubin $\hat{R}$ statistics. Hamiltonian Monte Carlo diagnostics also provide no indication of poor sampling (Betancourt, 2016).

[20] Recall that this includes: the share of the county that is nonwhite, the share that is Hispanic/Latino, the share with a college degree, county median income, and 2020 two-party vote share for Biden.

Despite the high average correlations across items within counties, substantively interesting and meaningful variation occurs. For example, the residual geographic correlation between governor vote choice and belief that Biden won legitimately is just below 0.7, but the correlation between governor vote choice and Republican party identification is about -0.5. This difference indicates that calibrating to the known county-level gubernatorial election results will result in a smaller calibration update for the party identification estimate than the update to our estimate of beliefs whether Biden won legitimately. In short, gubernatorial vote choice carries more information about — i.e., is more correlated with — opinions about Biden's legitimacy than it does about party identification. This finding is perfectly consistent with the particulars involved; the Republican candidate for Governor, Tudor Dixon, was a self-proclaimed election-denier and some Republicans expressed dissatisfaction with her selection as their party's nominee (Ulloa, 2022).

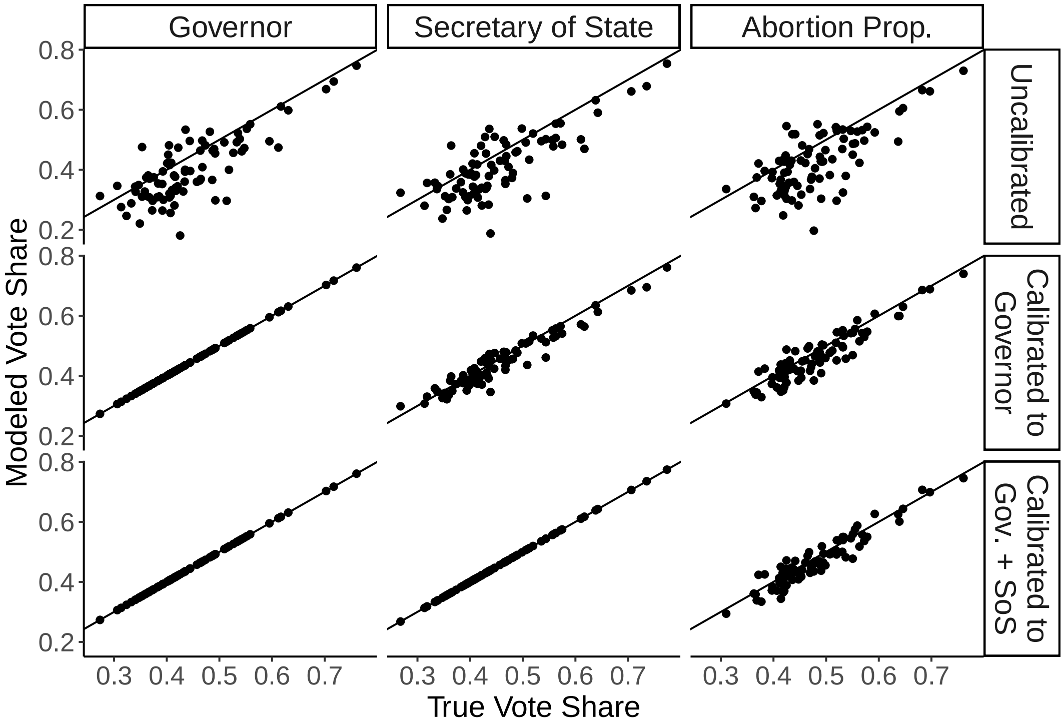
Using the correlations summarized in Table 1, we are able to calibrate the estimates using outcomes whose county-level outcomes are known. Because we have three vote-based outcomes for which the truth is known, we can examine the impact of calibrating to a single known outcome as well as multiple outcomes. Figure 1 plots the true county-level vote share on the x-axis (measured using the percentage in support of the Democrat candidate or pro-choice outcome for Proposition 3) against: the MRP modeled vote share absent any calibration (first row); MRP estimates calibrated using just the county-level gubernatorial results (second row), and MRP calibrated using both the gubernatorial results and Secretary of State results.[21]

The results in the top row indicate that the uncalibrated MRP model exhibits systematic errors in predicting county-level vote share. Consistent with the pattern of pre-election polls over-predicting Republican performance in Michigan, the results of a standard MRP model underestimates Democratic vote share in the races for Governor and Secretary of State as well as the support for the pro-choice position in Proposition 3. Moreover, even with the county-level covariates and intercept shifts, there is clear evidence of heteroskedasticity in the accuracy of the estimates; there is more variation in polling errors in less-Democratic counties compared to more-Democratic counties. The errors that occur in less-Democratic counties is also sometimes sizable and in excess of 10 percentage points in terms of the Democratic percentage, which means the error is more than 20 percentage points on the margin.

Next, consider the second row, which adjusts the estimates to the the known results of the

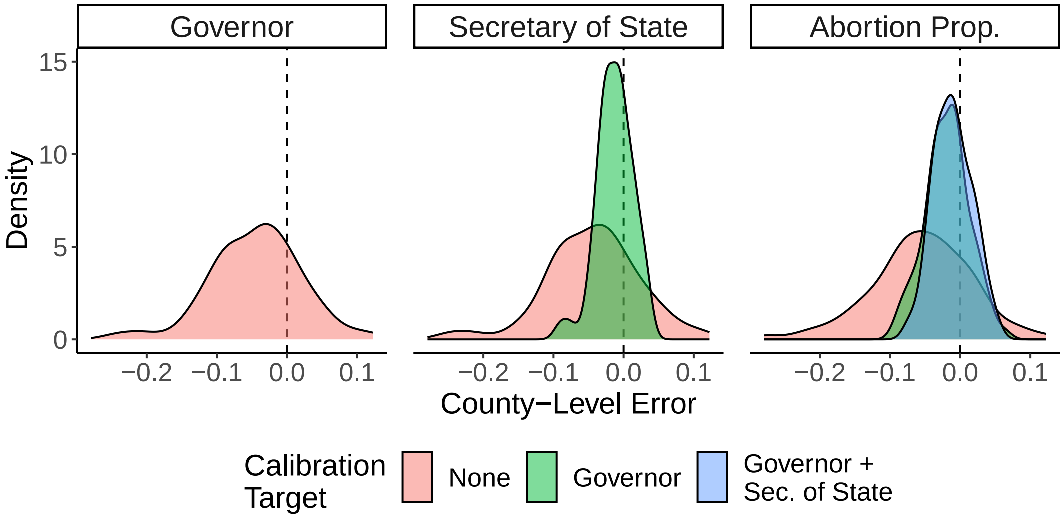
[21] The modeled vote share is the posterior mean of the poststratification estimates. For calibrated estimates, calibration is performed separately for each draw from the posterior, then the results are averaged together to form the point estimate.

Figure 1: County-Level MRP Results, Michigan 2022 Elections



Notes: The x-axis shows the true county-level Democratic/pro-choice vote share and the y-axis shows model-based estimates. The top row shows uncalibrated MRP estimates, the middle row shows estimates calibrated to the Governor race, and the bottom row shows estimates calibrated to Governor and Secretary of State races.

Figure 2: Distribution of County-Level Errors, Michigan 2022 Elections



Notes: Positive (negative) values indicate an overestimate (underestimate) of Democrat/pro-choice support.

gubernatorial election using our proposed method. The results show a dramatic improvement in the accuracy of the model-based estimates. As expected, the county-level vote shares for the gubernatorial election perfectly match the actual outcomes post-calibration. This result merely confirms that the logit-shift calibration occurred successfully: by construction, all points in this panel must fall on the 45-degree line. The relationships plotted in the center panel and center-right panel indicate that predictions for Secretary of State and the abortion proposition are substantially improved through calibration to the Governor results. Informally, compared to the uncalibrated results graphed in the top row, the points are much closer to the 45-degree line for both races and the sizable hetereogeneity in the uncalibrated county-level estimates is greatly minimized post-calibration.

As noted, one benefit of focusing on Michigan is that we know the actual results for all three elections. As a result, we can also examine the benefits from calibrating based on two outcome measures. The final row of Figure 1 investigates the impact of calibration to both Governor and Secretary of State results on predicted support for Proposition 3. As the panel in the lower-right reveals, the doubly-calibrated county-level estimated support for Proposition 3 improves only slightly relative to the estimates calibrated using only the gubernatorial results. In part, this is likely due to the similarity of the relationship between the relationship of these two elections and support for Proposition 3 and the similarly of the county-level gubernatorial and Secretary of State opinion. The correlation of county intercepts between the governor's race and Proposition 3 was nearly 0.7 — the same correlation as the correlation between the county intercepts for the Secretary of State race and Proposition 3. The double-calibration we perform consequently adds very little information over the information provided by calibrating based on a single race. Still, there is a slight increase in accuracy from calibrating to both elections.

To better characterize the increased accuracy of the calibrated estimates, Figure 2 plots the distribution of errors across calibration methods. The increase in accuracy is immediately obvious — as is the limited gain from calibrating using both the Governor and Secretary of State races. It is also worth noting that even with the improved accuracy post-calibration, several counties have errors in excess of 5 percentage points in terms of support for Proposition 3. This is notable because 5 percentage points of error in the level reflects 10 percentage points of error in the margin.

Table 2 shows these error reductions more formally by calculating the mean signed error, the mean absolute error, the root mean squared error, and the range of errors across counties. Each panel consists of three rows: the first row shows the error for the uncalibrated, standard MRP county-level estimates; the second row reports the results after calibrating to the Governor race; and the third row reports the consequences of calibrating to both the

Table 2: County-Level Errors, Michigan 2022 Elections

Race	Calibration Target	Mean Signed Error	Mean Abs. Error	Root Mean Sq. Error	Min. Error	Max. Error
Governor	None	−4.4	6.1	7.9	−24.5	12.2
	Governor	-	-	-	-	-
	Governor + Sec. of State	-	-	-	-	-
Secretary of State	None	−4.4	6.3	8.0	−25.1	11.7
	Governor	−1.2	2.2	2.8	−9.3	3.6
	Governor + Sec. of State	-	-	-	-	-
Abortion Proposition	None	−5.5	7.0	8.9	−28.0	12.1
	Governor	−1.8	2.9	3.6	−9.2	6.3
	Governor + Sec. of State	−1.1	2.4	2.9	−7.3	5.2

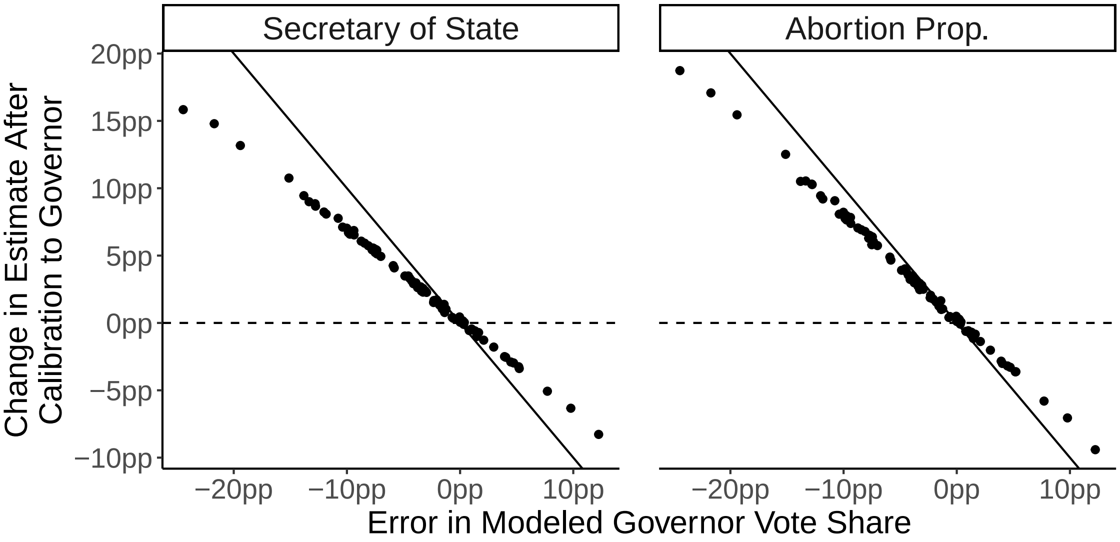
Notes: Entries in the table show the county-level error associated with uncalibrated and calibrated MRP vote share estimates for the Democrat/pro-choice outcome. The first row for each race is the error when using uncalibrated MRP; the second row is the error when calibrating to the governor outcomes; the third row is the error when calibrating to both the governor and secretary of state outcomes.

Governor and Secretary of State race.

Calibrating the estimates to match the outcome of the Governor race reduces the mean signed error in the Secretary of State race from 4.4 percentage points in the Republican direction to just 1.2 percentage points — a reduction of 73%. The mean absolute error and root mean squared error are both reduced by nearly two-thirds. The range of errors is also dramatically reduced, from $[-25.1, 11.7]$ to $[-9.3, 3.6]$.

The results for the abortion proposition are similar, and show that there are some modest additional gains from calibrating to multiple election outcomes. The mean absolute error for the abortion proposition is reduced from 7.0 to 2.9 percentage points when calibrating to the Governor results — a reduction of 59%. When jointly calibrating to both Governor and Secretary of State, the mean absolute error is further reduced to 2.4 percentage points. The maximum and minimum error are also greatly reduced in magnitude by calibrating to both races. In particular, the range of errors is decreased from $[-28.0, 12.1]$ to $[-7.3, 5.2]$ —

Figure 3: Change in County-Level Estimates from Calibration to Governor Results



Notes: The y-axis plots the difference between calibrated and uncalibrated county-level estimates for Secretary of State and the abortion proposition. The x-axis shows the county-level error in the Governor's race before calibration where positive values indicate overestimating the support for Democrat Incumbent Gretchen Whitmer.

meaning that the range of survey errors is reduced by 69%.[22]

Finally, to visualize the nature of the calibration, Figure 3 plots the difference between the calibrated and uncalibrated estimates against the baseline error in Governor estimates.[23] Points above 0 on the y-axis indicate that the estimates of Democratic/Yes vote share increased after calibration; points to the right of 0 on the x-axis indicate that the baseline MRP model overestimated Democratic vote share in the Governor race. If the estimates were adjusted by exactly the same amount as the error in the auxiliary variable, then they would lie along the 45-degree line. That larger calibrations were made in counties where the support for Whitmer was underestimated (producing errors < 0) means that most of the polling error was occurring in communities where her support was estimated to be the smallest based on the observed data and statistical adjustments being done; the magitude of the calibration was larger those areas relative to areas where her support was overestimated (and where the error > 0)/ That we see a lower slope in the left-panel characterizing the calibrations for the Secretary of State race than we do in the right panel plotting the calibrations for Proposition 3 highlights the fact that the correlation between vote choice for Governor and vote choice for Secretary of State was larger than the correlation between vote

[22] From a range of 40.1 in the uncalibrated estimates to a range of 12.5 in the doubly-calibrated estimates.

[23] Figure 9 in the Appendix plots the calibration effect against the raw county-level Governor results.

choice for Governor and support for Proposition 3.

4.2 Effect of Calibration on Party ID Estimates

In addition to the three outcomes related to known election outcomes, we also jointly estimated the opinions on several items for which the true distribution is unknown. This was intentional as the primary advantage of the calibration we propose is to use the association between items reported in Table 1 to calibrate estimates whose truth is known to adjust estimates whose truth is unknown. While it is impossible to confirm the effects of this calibration on the accuracy of those estimates, the validation results suggest that the gains are non-trivial so long as the underlying assumptions hold. Recall that the assumption required to make the calibration adjustment is that the residual geographic correlation between outcomes is the same in the population as in the survey sample.

One such measure we examine here is self-reported partisanship. In some states, researchers have access to ground-truth data on partisanship from voter files. Michigan, however, does not have partisan voter registration, and primaries in Michigan are open to all voters.[24] These features of election administration make surveys vital tools for measuring partisanship at the substate level. Pollsters increasingly rely on partisanship in weighting procedures, but the lack of ground-truth data in some states limits the applicability of this approach method. Our method presents a possibility for improving estimates of partisanship by jointly estimating partisanship with sociodemographic variables, geographic variables, and substate calibration data.

The association between partisanship and voting behavior is well-known and well-documented. As such, we can be confident that the errors made in estimating county-level election results are likely also related to errors made in estimating county-level partisanship.[25]

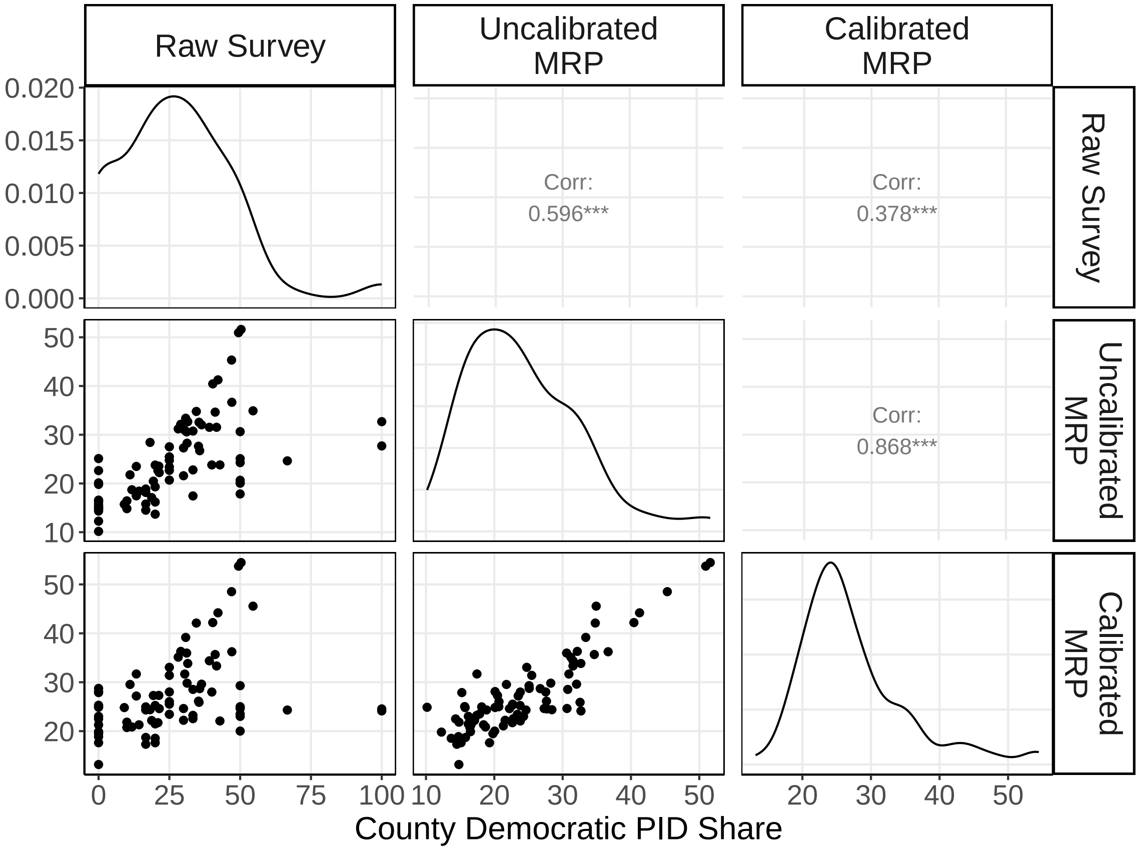
Figure 4 reports the raw, estimated, and calibrated estimates for the percentage of Democrats in each county to highlight the consequences of the calibration we propose. The left-most column reports the distribution of self-identified Democrats in each county according to our survey data with standard weighting techniques applied.[26] The coarseness and skew of the estimates due to the small sample sizes in particular counties is clearly visible

[24] `https://www.ncsl.org/elections-and-campaigns/state-primary-election-types`

[25] Note that it is also possible that the errors made in predicting vote share are a result of within-party differences (e.g., the partisans who respond vote differently than the partisans who do not). Unfortunately, it is impossible to account for such patterns because it is impossible to verify this pattern without conducting a non-response analysis to compare the opinions of partisans who do and do not respond. That said, most survey work finds very little variation in partisan voting behavior on statewide races between Democrats and Republicans.

[26] We generate these weights by raking to the Michigan-specific marginal distributions of the following variables: age, White/non-White indicator, gender, college educated, interactions between White × college, and 2020 presidential vote.

Figure 4: Democratic Party ID in Raw Survey Data, Uncalibrated MRP, and Calibrated MRP



(e.g., the number of counties containing no Democrats in the raw survey data). The second row reports the results of estimating the standard MRP model without calibration. Although modestly correlated with the raw survey estimates ($r = 0.596$), as would be expected given that the raw survey is the basis for the MRP estimation, the distribution of county partisanship changes when using the MRP estimates. In particular, the county-level estimates are pulled closer to the overall average — eliminating the cluster of counties with no Democrats.

Calibrating the MRP estimates using the results of the Governor and Secretary of State elections reveals additional effects. While the calibrated MRP estimates are highly correlated with the uncalibrated MRP estimates ($r = 0.868$) — suggesting that the county-level adjustments largely preserve the rank ordering of the estimated percentages — the calibrated estimates are relatively weakly related to the raw survey estimates ($r = 0.378$). The effects of the calibration are also evident in the plotted distribution of county-level partisanship and the scatterplots showing the relationship between the various estimates. As the bottom-right panel reveals, the calibrated MRP estimates are far less dispersed than either the uncalibrated MRP (center panel) or the raw survey averages (top-left panel). While the calibrated and uncalibrated estimates are clearly related (bottom center panel), the calibration reduces the variation in estimated county-level partisanship.

The changes we document reveal that there are limits to what we can extract from survey data alone — especially given small sample sizes and the difficulty of modeling outcomes using only predictors for which the full population joint distribution is known. Jointly estimating outcomes and calibrating the estimates using known ground-truth data, however, suggests a way to improve the accuracy of estimates, as long as the correlations across variables in the population are similar to those measured in the survey.

5 Application: Partisan Sorting and Animus

Having demonstrated the utility of employing the calibration shift we propose using pre-election polling in Michigan, we now turn to a substantive application related to the political geography of affective polarization in the United States. The causes and consequences of polarization have rightly received tremendous attention as scholars have sought to understand partisan divisions throughout the polity (Drutman, 2020). Much work focuses on how individuals are increasingly likely to hold positive feelings toward same-party partisans and negative or even hostile feelings toward those who belong to the opposite party (see Iyengar et al., 2019, for a recent review). At the same time, there has been increasing geographic polarization in the U.S. — with rural areas increasingly dominated by Republicans and urban areas increasingly dominated by Democrats (Rodden, 2019; Bishop, 2009; Brown and Enos, 2021).

These two trends raise the questions about the geographic distribution of affective polarization. Is antipathy toward the outparty highest in narrowly divided areas, or is it more common in homogenous, politically segregated communities?

Ex ante, either relationship seems plausible. The lack of day-to-day contact with outpartisans could generate stereotypes and affective polarization (Santoro and Broockman, 2022; Mutz and Mondak, 2006). On the other hand, partisan animosity occurs in places that are evenly split and where electoral contests are highly contested. In such areas, candidates may rely on demonizing opponents in an attempt to mobilize supporters. Or, perhaps, outparty antipathy is the result of a mismatch between local preferences and statewide political sentiment (Cramer, 2016).

To investigate the relationship between partisan animosity and geographic context, we use the same MRP calibration model to generate county-level estimates of partisan animosity. We then show how animosity varies across political contexts, as measured by vote share in the 2020 presidential election.

Table 3: Responses to: "Which comes closest to your view, even if none is exactly right?"

Response	*N*	%
Republicans are a danger to our country and must be defeated at any cost	3,204	21.9
Republicans have mistaken ideas, but they are not a danger to our country	638	4.4
Both parties have mistaken ideas	6,898	47.2
Democrats have mistaken ideas, but they are not a danger to our country	989	6.8
Democrats are a danger to our country and must be defeated at any cost	2,888	19.8

We analyze 9,788 respondents collected nationwide using the SurveyMonkey sample described in the previous section. To measure partisan animus we use a Pew Research Center question asking individuals to express how concerned they are about the danger posed by the parties. We use self-reported presidential vote in 2020 as the auxiliary variable. We apply our calibration method at the county level.

Table 3 reports the unweighted distribution of survey responses and reveals that while the modal respondent believes that "both parties have mistaken ideas," a sizable minority — around 40% — believe that either the Democrats or Republicans are "a danger to our country" who "must be defeated at any cost."

To create the county-level estimates we create indicator variables for whether individuals select "Republicans are a danger" or "Democrats are a danger." We jointly model these survey outcomes with self-reported 2020 presidential vote. We using the same statistical specification we use in the Michigan validation case.[27] We then calibrate the resulting county-level MRP estimates using self-reported 2020 vote choice to ensure that the county-level MRP estimates match the 2020 certified election results. We update the animus measures using the method outlined in Section 2. Finally, we poststratify the calibrated results to estimate the fraction of each county who believe that Democrats/Republicans are a danger to the country.[28]

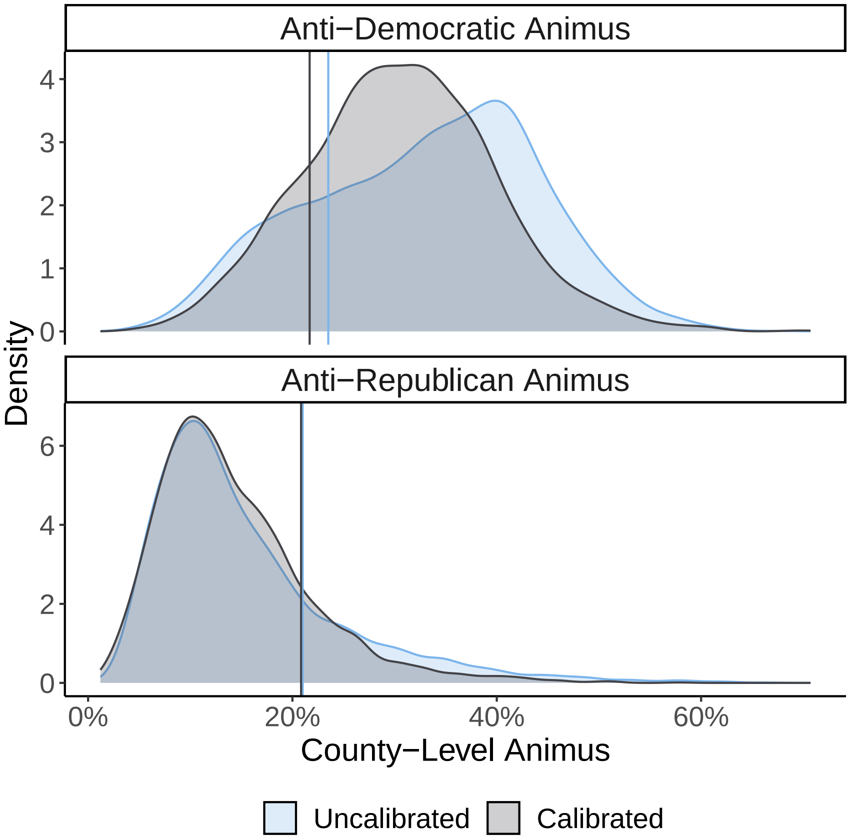
5.1 County-Level Distribution of Partisan Animus

To begin, Figure 5 plots the distribution of partisan animus by county — defined as the share of the county who say that Democrats or Republicans are a danger. We present both

[27] We include individual-level covariates related to age, race, gender, education, interactions between race × education and nonwhite × education × age. We include state- and county-level covariates related to the racial and ethnic composition of the county, share with a college degree, median income, 2020 two-party vote share for Biden, and random intercepts.

[28] Technically the errors for the two partisan questions should be related as they are constructed from the same question — if a respondent chooses one response they cannot choose the other. We ignore this complication in the analysis that follows.

Figure 5: Distribution of Calibrated and Uncalibrated County-Level Partisan Animus Estimates



Notes: Densities show calibrated and uncalibrated estimates of county-level animus. Vertical lines shows national average (i.e., weighted by county population).

the uncalibrated MRP estimates and the estimates after calibration to the 2020 county-level election results.

First, methodologically, the calibration procedure shifts the county-level distribution of anti-Democratic animus leftward. Our uncalibrated estimates suggest that in the median county, 33% of the population think Democrats are a danger to the country, compared to an estimate of 30% after calibration. The difference is even more dramatic when considering the tail of the distribution. Before calibration, we estimate that there are 921 counties where more than 40% of the population expresses anti-Democratic animus; after calibration, this number is reduced to 442. Calibration has less effect on the county-level estimates of anti-Republican animus, though it does lead to slightly lower estimates at the top of the distribution. Generally, this finding is consistent with research suggesting that contemporary survey respondents may be more partisan than non-respondents (see, for example Bailey, 2018; Clinon and Trussler, 2022).

Table 4: National Estimates of Partisan Animus

Type	Anti-Democratic Animus	Anti-Republican Animus
Uncalibrated	23.5%	21%
Calibrated	21.7	20.8

Second, we find that there are many more counties with high levels of anti-Democratic animus than there are counties with anti-Republican animus. As just discussed, in the median county, 30% of the population say that Democrats are a danger to the country. Compare this to just 14% of the population saying that Republicans are a danger to the country in the median county. This finding is consistent with Republican domination of smaller, rural counties, which are more numerous but less populous (Rodden, 2019; Brown and Enos, 2021; Bishop, 2009).

Table 4 reports the nationwide share of anti-Democratic and anti-Republican animus using both the calibrated and uncalibrated measures. Again consistent with the spatial distribution of partisans, the population averages in Table 4 are nearly identical, despite the differenecs in county-level distributions seen previously. Calibration to 2020 presidential vote reduces the estimate of anti-Democratic animus by about 2 percentage points, but barely affects the estimate of anti-Republican animus.

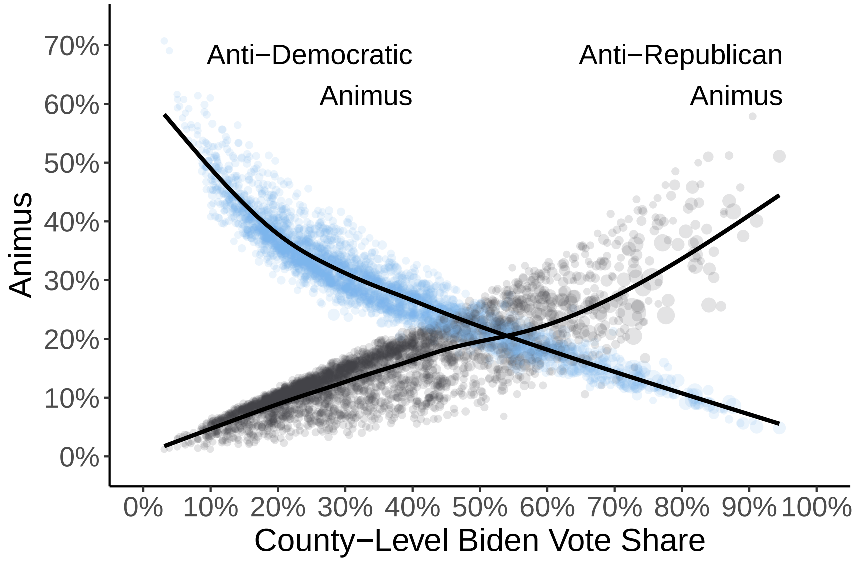
5.2 Relationship Between Animus and Count

The wide variation in county-level animus raises the question of what predicts animus. We begin to investigate this question in Figure 8, which plots partisan animus against the level of support for Biden in each county. This figure plots each county twice: blue dots represent the share of the county expressing anti-Democratic animus and the grey dots represent the share expressing anti-Republican animus.

There is a clear relationship between animus and election outcomes, as one would expect. It is also clear that the relationship is non-linear, with animus disproportionately high in partisan strongholds. Counties with higher Biden vote shares for Biden are certainly more likely to contain more individuals who express anti-Republican animus, but the smoothed regression line also reveals that the percentage is even larger in counties with the highest vote share for Biden. Similarly, counties that had very high Trump vote shares have extremely high levels of anti-Democratic animus.

Put differently, in areas most dominated by Biden supporters, we find even higher levels of political animus than we would expect based on the relationship between vote share and

Figure 6: County-Level Animus by 2020 Election Results



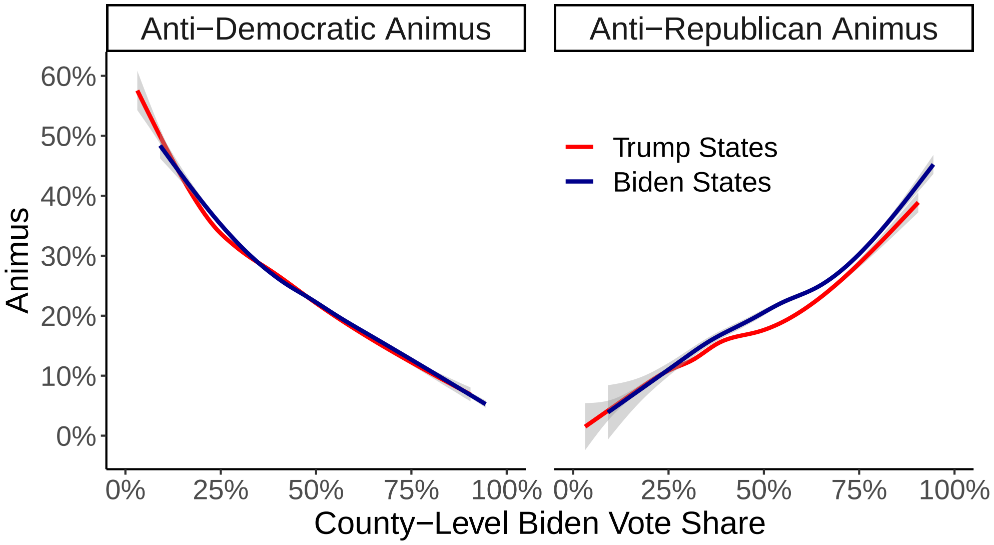
Notes: The y-axis shows the average partisan animus within a county and the x-axis shows Biden's two-party vote share in the county in 2020. Each county is plotted twice. Blue points show anti-Democratic animus and grey points show anti-Republican animus. Points are sized in proportion to county population. The smoothed regression lines are weighted by county population. Animus estimates are calibrated to 2020 presidential election results.

anti-Republican animus in places with less overwhelming support. A similar pattern emerges among Republicans, as the highest level of anti-Democratic animus is found in counties where Biden's support was less than 25% and the ratio of animus to vote share in these counties is greater here than elsewhere.

To compare the relative levels of animus holding the support for Biden fixed, Figure 6 plots the population-weighed average of animus estimates for both parties by certified Biden vote share. Consistent with work suggesting asymmetric polarization, there are generally higher levels of anti-Democratic animosity than anti-Republican animosity. In counties that split 50-50 for Trump and Biden, there is a higher level of anti-Democratic animus. The most pro-Trump counties, on the left-hand side of the plot, show average anti-Democratic animus levels reaching over 50%. In contrast, the most pro-Biden counties have average anti-Republican animus levels around 40%.

Another possible way that political geography may be related to partisan animus relates to the position of counties relative to statewide political power. Not every county with the same vote presidential vote share is similarly situated in terms of their political position within the polity. That difference may impact the level of perceived political danger posed by the opposition party. For example, are Republican-dominated rural areas in states that elect

Figure 7: Population Weighted Average of Estimated County-Level Animus by 2020 Biden Vote and State Result



Notes: The y-axis shows the average partisan animus within a county and the x-axis shows Biden's two-party vote share in the county in 2020. The smoothed regression lines are weighted by county population. Blue lines are states that Biden won and red lines are states that Trump won. The animus estimates are calibrated to 2020 presidential election results.

Democrats (e.g., NY, CA) more antagonistic towards Democrats than Republican-dominated rural areas in states that elect Republicans (e.g., TN, TX)? Or does the average level of anti-Republican animus in Democrat-dominated urban areas in Democrat-leaning states differ from the average level in Democrat-dominated urban areas in Republican-dominated states?

Figure 7 begins this investigation by plotting the population weighted smoothed county averages as a function of Biden vote share based on whether the state was won by Biden or Trump.

There is little graphical evidence based on the population-weighted smoothed averages in the average amount of anti-Democrat animus based on whether the county is located in a state where Biden or Trump won. The relationship between vote-share and the percentage of the populace who think Democrats pose a danger to the country depends far more on the level of Biden support in the county than whether Biden won or lost the state's Electoral College votes. However, cross-sectionally, extremely pro-Trump counties in Trump-supporting states (on the far left of the plot) exhibit especially high levels of anti-Democratic animus. But overall, the picture emerges that feelings toward Democrats in Republican-dominated counties are similar regardless of whether Democrats are winning or losing at the state level; being in the political majority or minority statewide has no obvious association with the level of animus expressed.

Among Democrats, however, a different pattern emerges. The deviation in the two plotted smoothed population averages reveals that there is more in anti-Republican animus in states won by Biden than in states won by Trump, holding fixed the Biden's vote share in the county. Anti-Republican animus is therefore highest in counties where Biden received the most support and in states that supported Biden. Those living in places surrounded by individuals with similar views and in which those views are also held statewide are likely to have the highest levels of animus – suggesting that perceptions of danger are more likely to occur among political majorities rather than political minorities.

To explore the association between local and statewide support and partisan animus in more detail, we estimate a regression of the percentage of residents expressing animus toward the other party as a function of county-level demographics, Biden's certified margin in the county, whether Biden won the state, and an interaction between margin and winning the state. The relationship with Biden margin captures the relationships noted earlier — there is more animus in counties with more partisans. As such, there should be a strong, positive relationship between Biden margin and anti-Republican animus and a strong negative relationship between Biden margin and anti-Democrat animus. The interaction between margin and the winner of the state allows the effect of margin to vary depending on whether the county is part of the political majority or minority in the state. If belonging to the political majority in the state increases partisan animus we should see a positive interaction effect between Biden margin and Biden winning; if belonging to the political minority in the state increases partisan animus we should observe a negative interaction. To account for the estimation uncertainty, we estimate the regression for every draw from the posterior for the calibrated county-level animous measures and we report the average and standard deviation of the resulting vector of coefficients to summarize the estimate and its' precision respectively.

Table 5 reports the county-level results for a weighted linear regressions using the posterior of estimated county margins and weighting each county estimate by its' population. As the coefficient estimates reveal, both the political context of the county and the relationship between the political leanings of the state matter. Controlling for county-level characteristics slightly affects the estimated relationships, but the substantive findings are unchanged.

The large coefficient on margin can be interpreted as revealing the average percentage of supporters who express animus toward the opposition party. In specification (1), for example, in states won by Trump (so that *Biden State* = 0), for every 100 Trump voters there are roughly 24 who think that Democrats pose a danger to the country. The interaction effect with the outcome in the state reveals the impact of living in a county that belongs

Table 5: Relationship Between Partisan Animus and Electoral Competition

	Anti-Democratic Animus		Anti-Republican Animus	
	(1)	(2)	(3)	(4)
Biden Margin	−23.8*	−26.2*	15.7*	24.9*
	(1.9)	(3.1)	(1.7)	(2.9)
Biden State	0.2	0.6	2.1*	0.5
	(1.0)	(1.1)	(0.9)	(1.0)
Biden Margin × Biden State	3.2*	2.7*	4.7*	6.2*
	(1.1)	(1.0)	(1.3)	(1.2)
% Nonwhite		1.8		−8.4*
		(1.4)		(1.6)
% Hispanic/Latino		−0.2		−1.9
		(1.1)		(1.1)
% College		0.9		0.1
		(1.7)		(1.8)
Median Income		−0.4		0.6
		(1.1)		(1.4)
Constant	22.4*	21.5*	18.4*	21.7*
	(1.0)	(1.4)	(0.8)	(1.3)
Observations	3,111	3,111	3,111	3,111

Notes: Outcomes are the estimated share of residents in a county who express partisan animosity. Posterior standard deviations are in parentheses. Asterisks indicate that the central 95% credible interval excludes 0.

to the political majority or minority in the state. In state won by Biden, for example, the relationship between anti-Democratic animus and county vote share is slightly smaller — $-23.8 + 3.2 = -20.6$ — suggesting that out of 100 Trump voters there are about 21 who express the opinion that Democrats are dangerous.[29]

A similar pattern emerges among Democrats. The effect of living in a state won by Biden is to increase the coefficient of anti-Republican sentiment from 15.6 to 20.0 — meaning that an additional 5 voters out of 100 Biden voters would think that Republicans posed a danger to the country.

Characterizing the geography of partisan animus reveals several important conclusions. There is a clear association with vote-share, but the fact that the relationship is non-linear suggests that higher-concentrations of supporters in an area are associated with even

[29] The interaction may seem difficult to rectify with the similarity that seems to occur in Figure 7, but the effect is being driven by the difference observed in counties with very low support for Biden where we observe some deviations. Given the extreme values of these observations, these will be counties with a high degree of leverage in the regression.

higher levels of animus than would be expected based on the association we observe in less-concentrated areas. Moreover, the relationship also depends on the political orientation of the county vis-a-vis statewide political orientations – the highest level of animus occurs in counties located in states sharing the same political orientation. These two features mean that political animus is not a reaction of political minorities to political majorities, but it is instead most concentrated in areas with the largest concentration of likely political supporters. Areas with the highest concentration of like-minded individuals living in a state where those views are also most likely to prevail are the areas where the highest percentages of individuals believe that the other party is a danger to the country and must be stopped at any cost. While it is beyond our ability to diagnose why these views are most concentrated in these areas, our characterization of these associations highlights the adverse consequences of "political bubbles" as the strongest expression of hatred towards others occurs not in areas that are closely-contested and where the parties are evenly matched, but in areas where a single party dominates.

6 Conclusion and Implications

Researchers across social sciences have come to rely on multilevel regression and poststratification to estimate opinion in small subgroups and geographies. It provides a principled way of adjusting non-representative survey data and enables estimation of opinion in populations that were not intentionally sampled. Still, even with flexible regression modeling, there still may be error in MRP estimates. However, researchers sometimes have access to ground-truth data for some outcomes in particular geographic areas, such as election results, that provide information about the nature of this error.

We propose a principled and data-driven method to exploit discrepancies between ground-truth and survey-based estimates on auxiliary outcomes to update inference on issues for which no ground-truth data exists. The method relies on the intuition that highly correlated outcomes should have similar error. We estimate the correlations across outcomes in a multivariate model, then propose an error correction that relies on this estimate correlation. The basis of the adjustment is the "logit shift" calibration sometimes used in studies of elections (Ghitza and Gelman, 2013; Rosenman, McCartan and Olivella, 2023). The method can be used to generate high-quality estimates of issue opinion — or other outcomes with no known ground truth — at small levels of geography.

We report a validation exercise using a pre-election poll for the 2022 Michigan midterm elections. This is arguably a difficult test, as pre-election polls had a significant amount of error, as with other Michigan polls since 2016. We examine vote intention in three statewide elections being help simultaneously. We measure the accuracy of our county-level estimates before and after calibration to one of the other races.

Calibrating the results using gubernatorial voter share greatly improves the accuracy of the estimates for Secretary of State and Proposition 3, which guaranteed access to an abortion. Calibrating our estimates to county-level Governor results reduced error by two-thirds in the other two races. Calibrating to both Governor and Secretary of State provides additional modest gains in accuracy for Proposition 3.

Using the Michigan data, we also show how calibration affects estimates of other outcomes whose truth is unknown. We generate county-level estimates of partisanship after calibrating to all three elections. Our calibrated estimates of partisanship have much lower variability than the uncalibrated estimates, generated using either dissaggregation of the survey or traditional MRP. Given the strong correlation between party identification and voting, our results suggest that the calibrated estimates of partisanship are likely more accurate. These partisanship estimates could be studied in their own right, or could be used as the basis of poststratification in subsequent analyses.

We apply our method to the study of partisan animus. We model responses to a question asking whether the out-party is a "danger to the country" alongside self-reported 2020 presidential vote, which we use to calibrate MRP estimates. First, we show that calibration reduces the share of counties with very high levels of anti-Democratic animus — consistent with survey respondents being more partisan than the population as a whole. Second, there are many more communities with high levels of anti-Democratic animus than with anti-Republican animus. In the median county in terms of anti-Democratic animus, 30% of the population believes that Democrats are a danger to the country. In contrast, only 14% of the median county in terms of anti-Republican sentiment believes Republicans are a danger. Similarly, the most anti-Democratic counties have levels of animus far higher than the most anti-Republican counties. This pattern reflects the concentration of Democrats in a small number of counties. Next, we show that partisan animus is most common in strongholds — with a nonlinear relationship between county-level animus and 2020 voting patterns. Rather than animus being highest in closely contested areas, instead it is highest in politically homogenous areas.

In sum, our calibration method can improve small-area estimates of attitudes and behavior, so long as some ground-truth data is available on a related outcome. The calibration we perform relies on assumptions that may not always hold — namely, that correlation across issues in surveys are representative of the correlations in the population as a whole. Even when this assumption does not hold perfectly, the procedure is still likely to decrease error.

References

Bailey, Michael A. 2018. "Designing Surveys to Account for Endogenous Non-Response." Georgetown University Typescript.

Ben-Michael, Eli, Avi Feller and Erin Hartman. 2023. "Multilevel Calibration Weighting for Survey Data." *Political Analysis* pp. 1–19.

Betancourt, Michael. 2016. Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. Technical report.

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

Bishop, Bill. 2009. *The Big Sort: Why the Clustering of Like-Minded American is Tearing Us Apart.* Mariner Books.

Broniecki, Philipp, Lucas Leemann and Reto Wuest. 2022. "Improved Multilevel Regression with Poststratification through Machine Learning (autoMrP)." *Journal of Politics* 84(1):597–601.

Brown, J.R. and Ryan D. Enos. 2021. "The meaurement of partisan sorting for 180 million voters." *Nature Human Behavior* 5(8):998–1008.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80.

Buttice, Matthew K. and Benjamin Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21(4):449–467.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1).

Caughey, Devin and Christopher Warshaw. 2022. *Dynamic Democracy Public Opinion, Elections, and Policymaking in the American States.* Chicago, IL: University of Chicago Press.

Clinon, Joshua D., John S. Lapinski and Marc J. Trussler. 2022. "Reluctant Republicans, Eager Democrats? Partisan Nonresponse and the Accuracy of 2020 Presidential Preelection Telephone Polls." *Public Opinion Quarterly* 86(2):247–269.

Clinton, Joshua D. 2006. "Representation in Congress: Constituents and Roll Calls in the 106th House." *Journal of Politics* 68(2):397–409.

Clinton, Joshua D., J. Agiesta, M. Brenan, C. Burge, M. Connelly, A. Edward-Levy, B. Fraga, E. Guskin, D. Hillygus, C. Jackson, J. Jones, S. Keeter, K. Khanna, J. Lapinski, L. Saad, D. Shaw, A. Smith, D. Wilson, and C. Wlezien. 2021. "Task force on 2020

pre-election polling: An evaluation of 2020 general election polls.”.
URL: *www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR-Task-Force-on-2020-Pre-Election-Polling_Report-FNL.pdf*

Cramer, Katherine J. 2016. *The Politics of Resentment.* Chicago: University of Chicago Press.

Drutman, Lee. 2020. “How Hatred Came to Dominate American Politics.”.
URL: *https://fivethirtyeight.com/features/how-hatred-negative-partisanship-came-to-dominate-american-politics/*

Eaton, Morris L. 2007. *Multivariate Statistics: A Vector Space Approach.* Vol. 53 of *Lecture Notes, Monograph Series* Institute of Mathematical Statistics.

Freeder, Sean, Gabriel S. Lenz and Shad Turney. 2018. “The Importance of Knowing “What Goes with What”: Reinterpreting the Evidence on Policy Attitude Stability.” *Journal of Politics* 81(1).

Gelman, Andrew. 2009. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do.* Princeton: Princeton University Press.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

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

Ghitza, Yair and Andrew Gelman. 2013. “Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups.” *American Journal of Political Science* 57(3):762–776.

Ghitza, Yair and Jonathan Robinson. 2021. “What Happened in 2020.” *Catalist* .
URL: *https://catalist.us/wh-national/*

Goplerud, Max. 2023. “Re-Evaluating Machine Learning for MRP Given the Comparable Performance of (Deep) Hierarchical Models.” *American Political Science Review* pp. 1–8.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J. Westwood. 2019. “The origins and consequences of affective polarization in the United States.” *Annual Review of Political Science* 22:129–146.

Jensen, Amalie, William Marble, Kenneth Scheve and Matthew J. Slaughter. 2020. “City Limits to Partisan Polarization in the American Public.” *Political Science Research and Methods* 9(2):223–241.

Kennedy, Courtney, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers, Lydia Saad, G Evans Witt and Christopher Wlezien. 2018. “An Evaluation of the 2016 Election Polls in the United States.” *Public Opinion Quarterly* 82(1):1–33.
URL: *https://doi.org/10.1093/poq/nfx047*

Kuriwaki, Shiro, Stephen Ansolabehere, Angelo Dagonel and Soichiro Yamauchi. 2023. "The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level." *American Political Science Review* .

Lax, Jeffrey R. and Justin H. Phillips. 2012. "The Democratic Deficit in the States." *American Journal of Political Science* 56:148–166.

Leemann, Lucas and Fabio Wasserfallen. 2017. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation." *American Journal of Political Science* 61(4):1003–1022.

Levendusky, Matthew S., Jeremy C. Pope and Simon D. Jackman. 2008. "Measuring District-Level Partisanship with Implications for the Analysis of U.S. Elections." *Journal of Politics* 70(3):736–53.

Lewandowski, Daniel, Dorota Kurowicka and Harry Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100(9):1989–2001.

Lipps, Jana and Dominik Schraff. 2019. "Estimating subnational preferences across the European Union." *Political Science Research and Methods* 9(1):197–205.

Marble, William and Matthew Tyler. 2022. "The Structure of Political Choices: Distinguishing Between Constraint and Multidimensionality." *Political Analysis* 20(3):328–345.

Missouri Census Data Center. 2022. "Geocorr 2022: Geographic Correspondence Engine.".
URL: *https://mcdc.missouri.edu/applications/geocorr2022.html*

Mutz, Diana C. and Jeffrey R. Mondak. 2006. "The Workplace as a Context for Cross-Cutting Political Discourse." *Journal of Politics* 68(1):140–155.

Ornstein, Joseph T. 2020. "Stacked Regression and Poststratification." *Political Analysis* 28(2):293–301.

Park, Daniel, Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–385.

Rodden, Jonathan. 2019. *Why Cities Lose: The Deep Roots of the Urban-Rural Political Divide*. New York: Basic Books.

Rosenman, Evan T. R., Cory McCartan and Santiago Olivella. 2023. "Recalibration Of Predicted Probabilities Using the 'Logit Shift: Why Does It Work, and When Can It Be Expected to Work Well?" *Political Analysis* pp. 1–11.

Santoro, Erik and David E. Broockman. 2022. "The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments." *Science Advances* 8(25).

Simonovits, Gábor and Julia Payson. 2023. "Locally Controlled Minimum Wages Are No Closer to Public Preferences." *Quarterly Journal of Political Science* .

Stan Development Team. 2023. "Stan Reference Manual.".
URL: *https://mc-stan.org/docs/reference-manual/index.html*

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *Journal of Politics* 75(2):330–342.

Tausanovitch, Chris and Christopher Warshaw. 2014. "Representation in Municipal Government." *American Political Science Review* 108(03):605–641.

Toshkov, Dimiter. 2015. "Exploring the Performance of Multilevel Modeling and Poststratification with Eurobarometer Data." *Political Analysis* 23(3):455–460.

Ulloa, Jazmine. 2022. "A G.O.P. Test in Michigan: Is Trump a Help or a Hindrance?" *New York Times* .
URL: *https://www.nytimes.com/2022/10/01/us/politics/tudor-dixon-trump-michigan.html*

Warshaw, Christopher and Jonathan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *Journal of Politics* 74(1):203–219.

A Additional Figures and Tables

Figure 8: Estimated County-Level Animus by 2020 Biden Vote Share

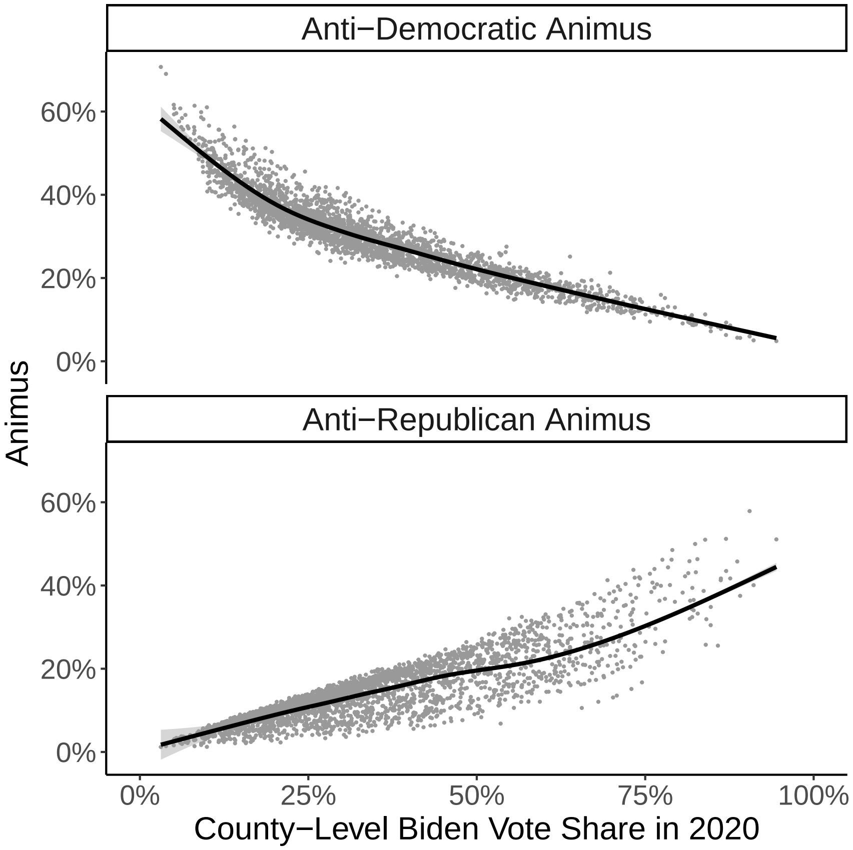
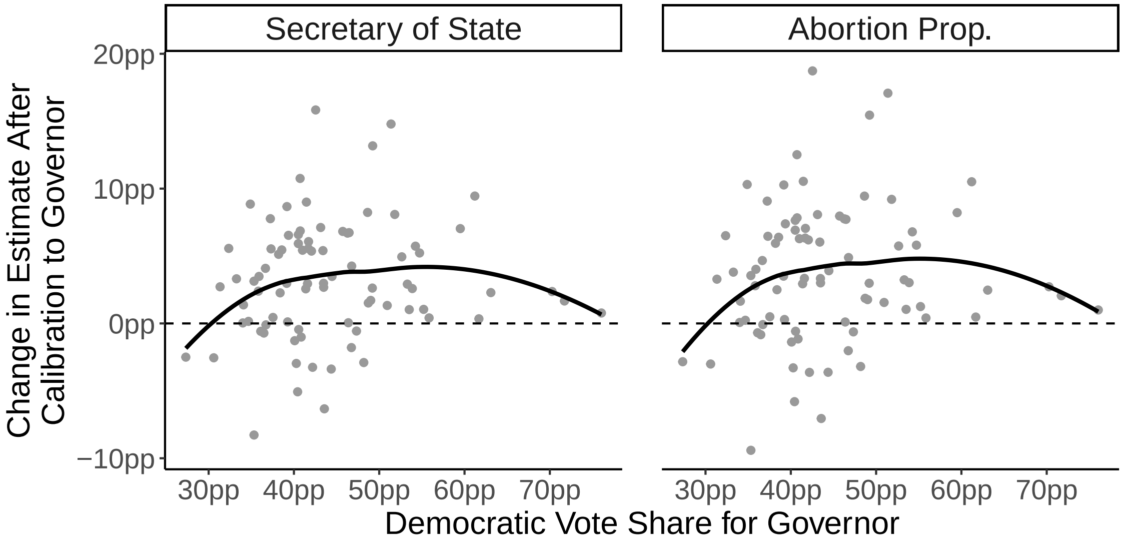
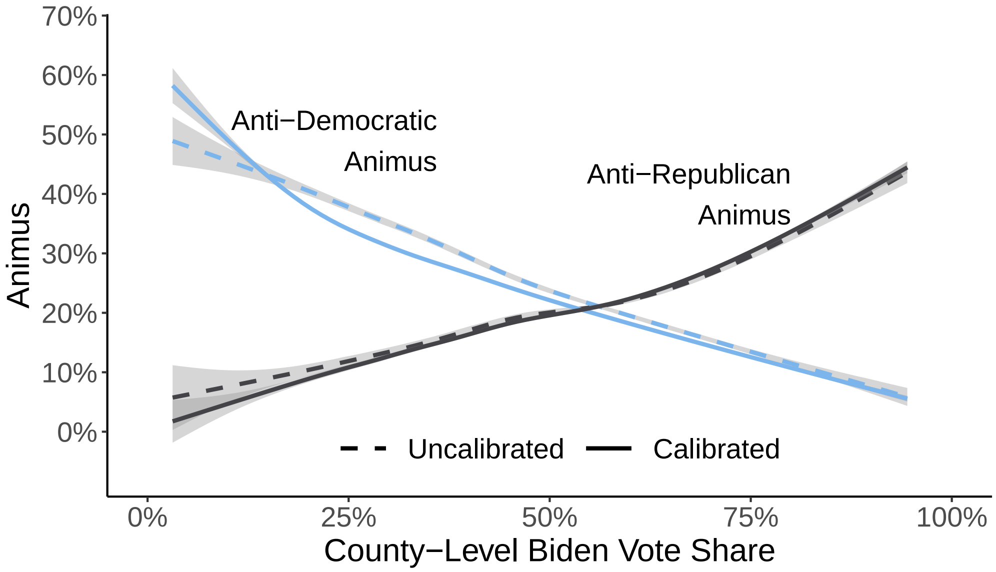


Figure 9: Change in County-Level Estimates from Calibration to Governor Results



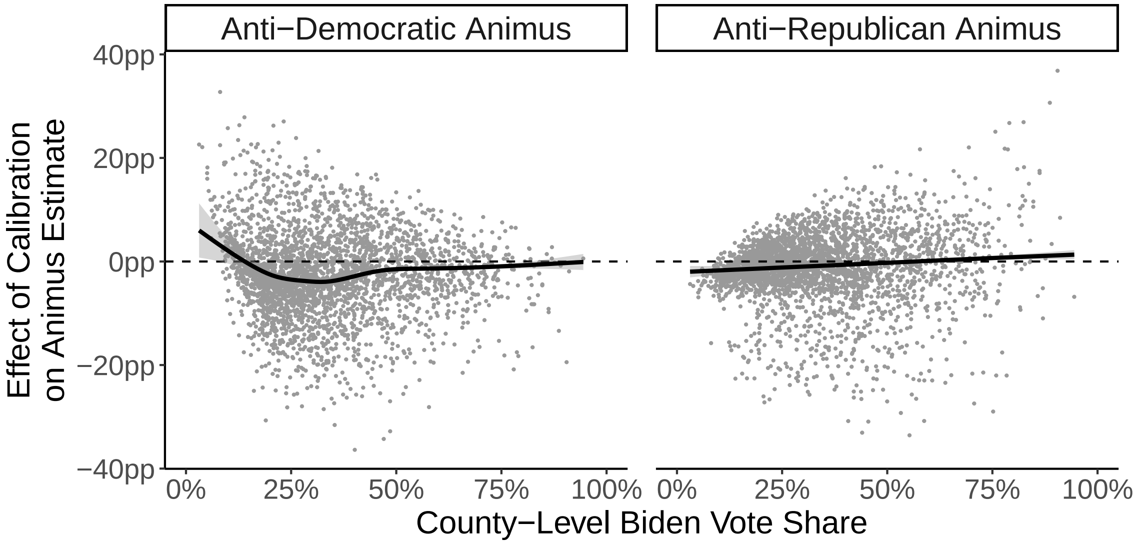
Notes: The y-axis plots the difference between calibrated and uncalibrated county-level estimates for Michigan Secretary of State and the abortion proposition. The x-axis shows the county-level vote share for Governor.

Figure 10: Animus by 2020 Election Results, Calibrated and Uncalibrated Estimates



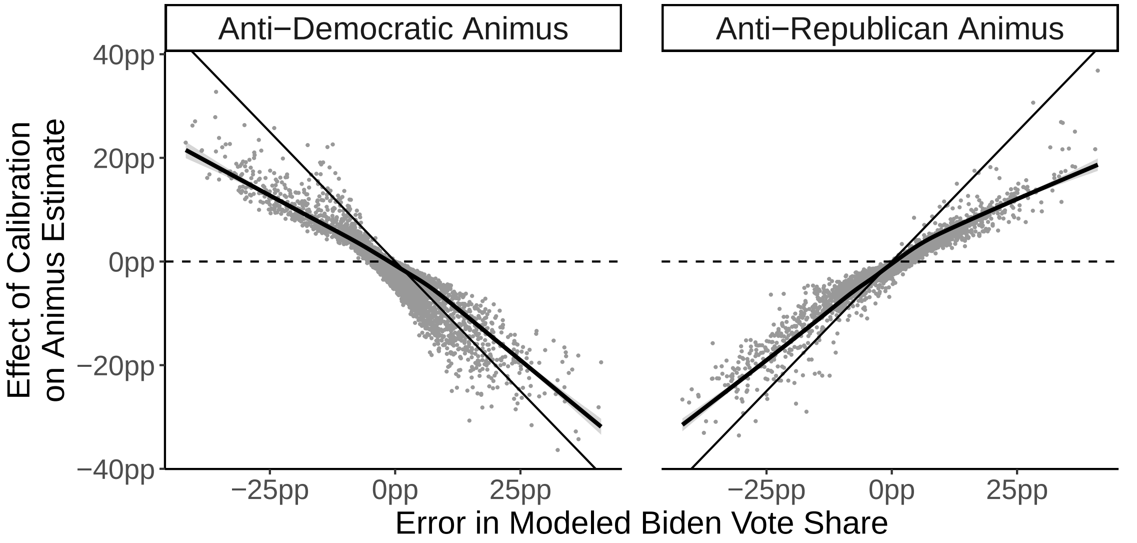
Notes: The y-axis shows the average partisan animus within a county and the x-axis shows Biden's two-party vote share in the county in 2020. Dashed lines show the uncalibrated estimates and the solid line shows the estimates calibrated to 2020 presidential election results.

Figure 11: Effect of Calibration on Animus Estimates by 2020 Election Results



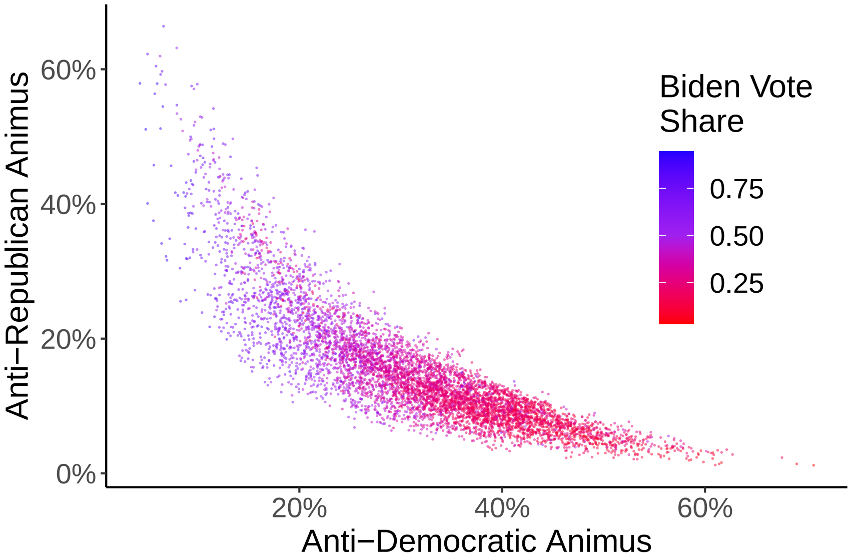
Notes: The y-axis shows the difference between the calibrated and uncalibrated animus estimates at the county level. The x-axis shows the county-level Biden vote share in 2020. The solid curved line is the smoothed county-level regression line, weighted by county population. The solid diagonal lines are 45-degree lines.

Figure 12: Effect of Calibration on Animus Estimates by Error in Modeled Vote Share



Notes: The y-axis shows the difference between the calibrated and uncalibrated animus estimates at the county level. The x-axis shows the error in county-level modeled vote share (higher values mean an underestimate of Biden vote share). The solid curved line is the smoothed county-level regression line, weighted by county population. The solid diagonal lines are 45-degree lines.

Figure 13: County-Level Relationship Between Anti-Democratic and Anti-Republican Animus


Biden Vote
Share
Anti−Republican Animus
Anti−Democratic Animus