Supplementary Materials for "Improving Subnational Opinion Estimation from Cluster-Sampled Polls"

Michael Auslen*

A Cluster-Sampling Procedures of Common Surveys

This appendix details the steps taken by polling firms to produce cluster-sampled surveys used in the main paper.

A.1 Gallup Polls, 1968–1984

Gallup's sampling procedure was relatively consistent during the period from 1968–1984. Codebooks for individual surveys (available from the Roper Center for Public Opinion Research) generally outline the steps taken to produce a sample. The summary here is based on the codebook from Roper Poll No. USAIPO1980-1162G, fielded September 12-15, 1980.

First, each state was assigned to one of eight geographic regions (New England, Middle Atlantic, East Central, West Central, South, Southwest, Rocky Mountain, and Pacific). Then, within each region, geographic areas were categorized into one of seven size-of-community strata based on urban/rural status and the population of the area (urbanized areas of cities over 50,000 people; remaining parts of cities over 1,000,000 people; cities with 250,000-999,999 people; cities with 50,000-249,999 people; cities with 2,500-49,999 people; rural villages; and farm or open country rural areas).

Based on these two categorization schemes, every location in the country is identifiable by a region×stratum pair. Within each pair, areas were assigned to equally-sized primary sampling units (PSUs). From each PSU, Gallup selected two localities, weighting by population so that larger areas were more likely to be sampled than smaller ones. This ensured that the probability of an individual being sampled within a region×stratum combination was roughly equal while still maintaining a balance across regions and strata.

In each selected locality, Gallup then continued to sample progressively smaller areas, weighting by population each time, until a census block or cluster of blocks was identified. Interviewers were assigned a household at random to begin interviewing and then were directed to move to other households within the block or cluster of blocks in a replicable pattern.

*Ph.D. candidate, Department of Political Science, Columbia University, New York, NY. Email: michael.auslen@columbia.edu. URL: https://michaelauslen.com

In practice, this sampling procedure appears to have been followed only when new clusters needed to be identified; memos from former Gallup statisticians made available by the Roper Center discuss needing new clusters after old ones have been "exhausted."

A.2 American National Election Studies (ANES), 2000

In 2000, the ANES used a design that produced 999 respondents interviewed face-to-face via cluster sampling and 800 respondents contacted via telephone using random-digit dialing. Here, I detail the procedure used for the cluster sample, which I use in the validation exercise in the paper. ANES sampling procedures are generally detailed in the introduction to the codebook for the given survey (e.g., Burns et al. 2001) but may also rely on technical reports produced following prior years' studies. For example, the 2000 procedure references Survey Research Center (1991).

The 2000 ANES sample included clusters chosen from a subsample of 108 PSUs. First, clusters from each of the eight largest PSUs were included with certainty. This included six large metropolitan statistical areas (MSAs), the Boston New England County Metropolitan Area (NECMA), and Dallas-Ft. Worth Combined Metropolitan Statistical Area (CMSA). Next, 10 PSUs representing of the next 20 largest MSAs or CMSAs were sampled, each of which was meant to represent itself plus one other. Then, additional MSAs and non-MSA regions were sampled from each of four Census regions (Northeast, Midwest, South, and West).

Within each sampled PSU, blocks were then grouped together to produce "area segments," of which several were chosen proportional to the number of households in 1990. The number of area segments representing each PSU varied, ranging from five to 12. Households and respondents were then chosen randomly from within sampled area segments.

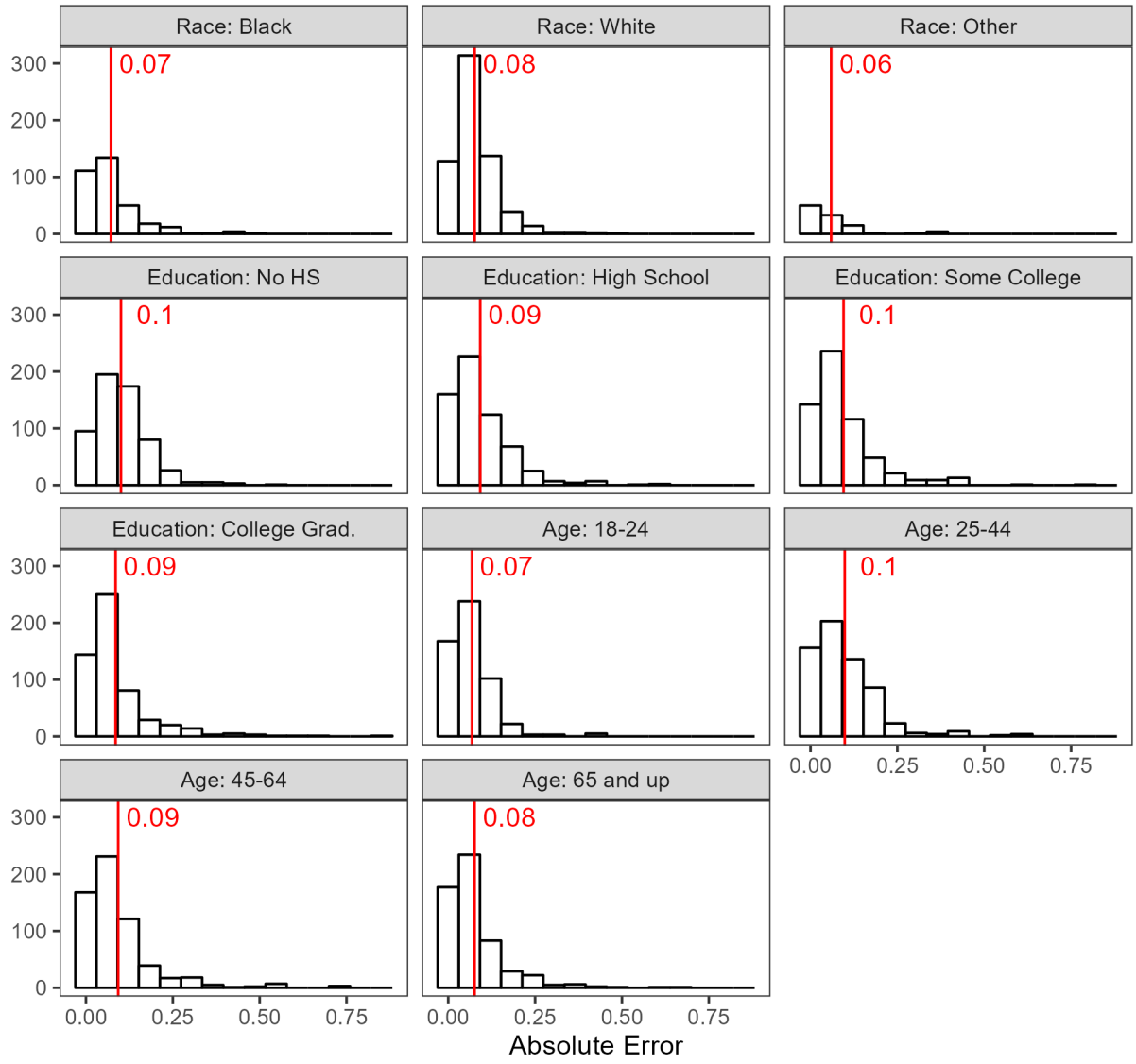
B Representativeness of Cluster Samples

In assessing the "representativeness" of cluster sampling methods, it is useful to distinguish between whether approaches produce representative samples of opinion *in expectation* and whether any individual poll produces a representative state in practice. In general, cluster-sampling approaches are well designed to produce nationally representative samples at a low cost. However, because these approaches rely on sampling a limited number of clusters in each primary sampling unit, the respondents in a given survey are likely not to be representative of their states—even if they would be in expectation over a larger, pooled sample.

To illustrate this point, I used 15 Gallup polls fielded in 1980 and compared state-level demographics estimated using disaggregation on the poll with true values.[1] First and most obviously, many states do not appear in a given cluster-sampled poll: on average, 10 states have zero observations in each Gallup poll I studied.

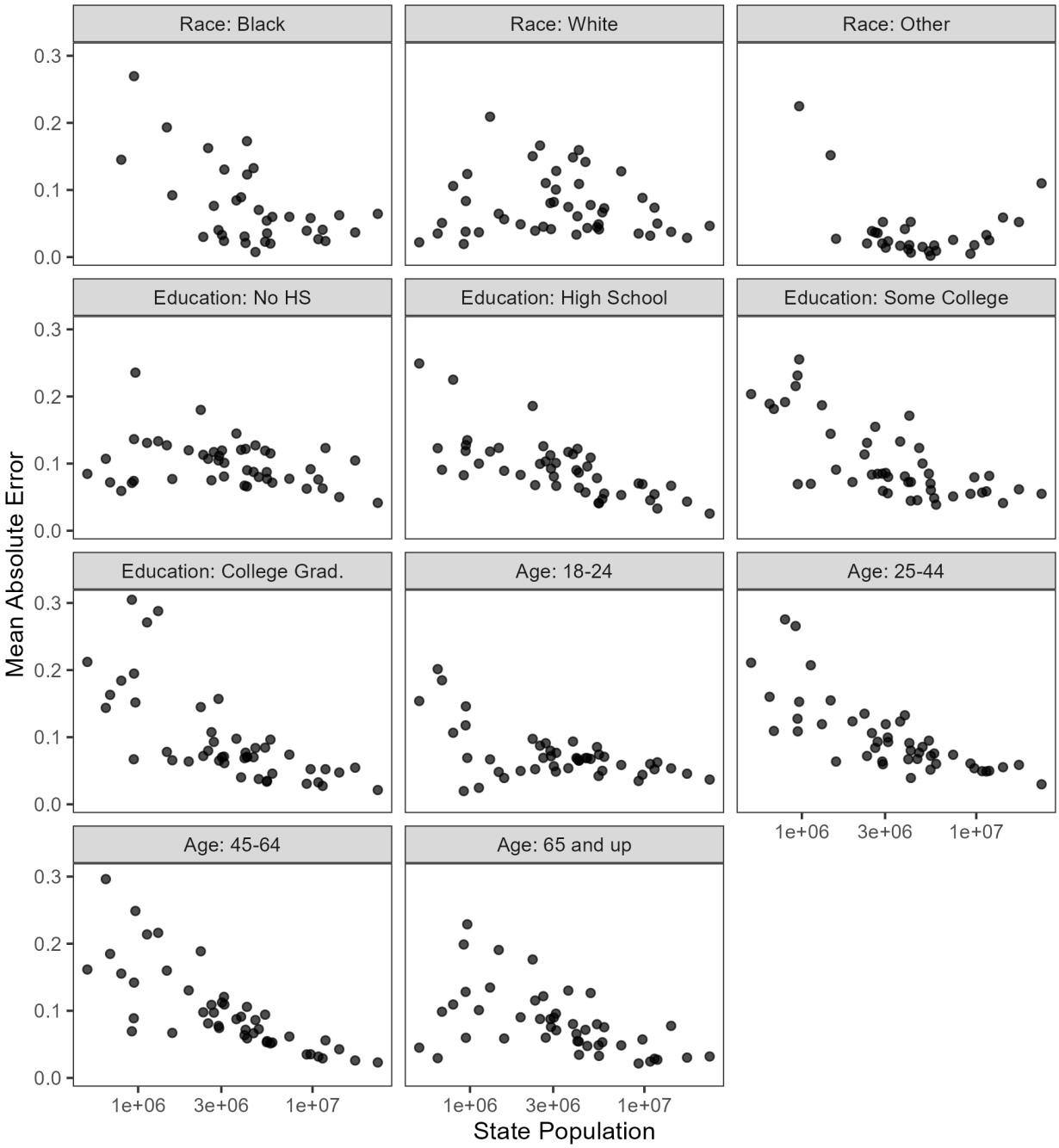
[1]The Gallup surveys studied are those fielded on January 4, February 1, February 29, March 28, April 11, May 2, May 16, May 30, June 13, June 27, July 11, August 1, August 15, September 12, and October 10.

Figure A1: Demographic Unrepresentativeness in Gallup Polls



Note: Panels report the distribution of the absolute difference between state-level estimates of demographic group size from 1980 Gallup surveys and their true value from the Census. Red lines report averages across all states and polls.

Figure A2: Demographic Unrepresentativeness in Gallup Polls by Population



Note: Panels report the correlation between population and the mean absolute error in estimates of demographic group size from 1980 Gallup surveys and their true value from the Census, averaged within each state.

Less obviously, but still problematic, among states that are included, state-level subsamples are often unrepresentative of the population. Figure A1 reports the distribution of this

unrepresentativeness using demographic characteristics of the public for which true values can be observed in the census. Specifically, I compute the share of respondents from each state that identify with various racial, education, and age categories. I do this separately for each of the 15 surveys and find the absolute value of the (disaggregated) survey's error compared to the true value in the Census. Among only those states with clusters included in a poll, the demographic makeup in the survey is generally 7 to 10 percentage points off from the truth. But strikingly, in some states and surveys, there is far greater error. On average, polls over- or under-represent states' Black and white populations by at least 15 percentage points in 12% of states, and they incorrectly represent educational categories by more that 15 percentage points in 17.5% of states, on average.

Polls are especially likely to be unrepresentative of the population in smaller states. Small states have fewer clusters, on average, and are less likely to have clusters sampled at all (due to population weighting in most multi-stage cluster-sampling procedures). Figure A2 plots the mean absolute error between survey-estimated demographics and true demographics (on the y-axis) against state population (on the x-axis). Smaller states consistently have more error, suggesting that they are especially likely to be poorly represented in cluster-sampled surveys.

C Poststratification using Historical Censuses

In the poststratification step of CMRP, state opinion is estimated by first predicting the average survey response for members of each combination of demographics within each state×stratum combination. Then, a weighted average is taken in which the weights are proportional to the frequency with which each demographic and stratum "type" exists in each state. Borrowing from Lax and Phillips (2009), we can represent the MRP estimate for a state s as:

$$Estimate_s^{MRP} = \frac{\sum_{c \in s} N_c \theta_c}{\sum_{c \in s} N_c}$$

in which c is a unique combination of all demographic characteristics and stratum within state s, θ_c is the predicted opinion for c, and N_c is the number of people represented by c.

Computing these estimates requires a "poststratification matrix" with information about the joint distribution of the population over all demographic groups and strata used in the model.[2] That is, the N_c for every possible combination of race, sex, education, age group, and stratum within each state. Notably, the defining characteristics of stratum vary depending on the poll being used. For Gallup polls, stratum reflects the seven size-of-community strata; for the ANES, stratum represents a tier of MSAs or non-MSA counties. See discussion below of the data sources and assumptions made in modeling these polls.

In general, census microdata published by IPUMS can be used to quickly and easily generate poststratification matrices for traditional MRP (see Lopez-Martin, Phillips and Gelman (2021) for instructions to produce a poststratification matrix with IPUMS microdata). However, because CMRP requires a joint distribution of demographics at the sub-state stratum

[2]While obtaining joint distributions is ideal for MRP, Leemann and Wasserfallen (2017) have proposed an alternative method based only on the distribution of individual demographic groups within each subnational unit.

level, microdata are not ideal, as they rely on relatively few observations in less populated state×stratum combinations, and detailed sub-state identifiers are not always available in the microdata to protect respondents' privacy. Instead, scholars should use joint distributions from census tables, which can be downloaded from IPUMS-NHGIS (Manson et al. 2021). Below I detail the steps taken to create poststratification matrices for use with the Gallup polls circa 1970 and 1980, and the 2000 ANES.

C.1 Poststratification for Gallup Polls using 1970 and 1980 Census

Gallup's sampling procedure from the 1960s into the 1980s relied on strata that divided geographic areas based on size of community. As such, a poststratification matrix can be computed by aggregating up from place-level census tables (in U.S. Census terminology, place refers to cities, towns, or other localities) after matching each place to the correct stratum in the Gallup procedure.

To do so, we must first identify a Census table with joint distributions of the demographic characteristics of interest and ensure it exists at both the place and state levels for the relevant year. The 1970 census did not report a joint distribution of race, sex, age, and education. This forced me to drop education from the analyses of 1968 and 1972 presidential election polls. I used "Table NT17: Sex by Age" from the sample-based population dataset, which includes the distribution of sex and age groups, broken down by race. From the 1980 census, I combined information from "Table NTPB46: Sex by Age by Years of School Completed," which contains information about people from 18-24 years old, and "Table NTPB48: Sex by Age by Years of School Completed," which has age breakdowns for those 25 and older.

With data at the place and state level on hand, I then followed the below steps:

1. Eliminate any information duplicated across Census tables. For the 1980 poststratification matrix, Table NTPB46 lumps together the population aged 25 and over, while Table NTB48 includes the breakdown. I dropped the information about those over 25 from NTPB46 and merged the remaining tables together.

2. Within each place and in the statewide dataset, arrange the dataset so that each row corresponds to a single combination of demographics. E.g., Black women between 18-24 with a college degree in Birmingham, AL.

3. Merge demographic groups to reflect the information in the Gallup survey and the desired level of granularity for the model. Census data may contain more detailed age breakdowns than one might optimally fit a model to, or may have more detailed information about racial groups that the pollster did not collect information on. E.g., for Gallup polls of this era, race was collected only as Black, White, or Other; as such, information from the Census on Hispanic origin, as well as other racial groups need to be combined. In order to poststratify based on the multilevel model, all levels must be the same.

4. Each place should be assigned to a stratum based on the total population reported by the Census. For Gallup polls of this era, places should be divided into (1.) 1,000,000 or more people; (2.) 250,000-999,999 people; (3.) 50,000-249,999 people; (4.) 25,000-49,999 people; (5.) places under 2,500 people.

5. Aggregate up from the place level to the stratum level within each state by summing up population of each demographic combination within any place that is part of each stratum.

6. The final stratum in the Gallup cluster-sampling procedure is farms and open-country rural areas. By definition, these are not within places and, therefore, are not part of the places dataframe. However, joint distributions in these areas can be estimated by taking the state-level joint distributions and subtracting out any people included in a different stratum. The remaining population is the open-country rural stratum.

The resulting dataframe should have columns for race, sex, age group, education, state, and stratum that align to the groupings in the Gallup poll, as well as the number or share of the population represented by each combination of variables. State-level variables (region, presidential vote, religiosity, etc.) can easily be merged into this data frame. Adding these variables to both the multilevel model and poststratification phase improve the quality of estimates obtained from MRP (Lax and Phillips 2009; Buttice and Highton 2013). Poststratification matrices for use with Gallup polls in this period can be found in the replication materials for this paper.

C.2 Poststratification for ANES, 2000

The ANES produces its samples based on MSAs (and occasionally CMSAs or NECMAs). While the specific sampling procedure cannot be identified, ANES documentation indicates that there are three tiers of MSAs and one tier of non-MSA counties, which I attempt to replicate as a stratum variable. Because MSAs are almost always made up of whole counties, it is relatively straightforward to map from the joint distribution of demographics within each county to a poststratification matrix. To do so for 2000, I used "Table NPCT065D: Population 18 Years and Over by Sex by Age by Educational Attainment" from the Sample-Based Data in the 2000 Census. This file includes several age and education variables by sex and can be broken down into racial and ethnic groups. With this county-level data in hand, I followed the below steps to produce a poststratification matrix:

1. Arrange the dataset so that each row corresponds to a single combination of demographics within a particular county.

2. Merge in a file with information about which MSA each county belongs to (or if they are not part of an MSA). This can be obtained from the Census Bureau or from the Missouri Census Data Center's Geocorr tool.[3] Note that counties in Connecticut, Rhode Island, Massachusetts, New Hampshire, Vermont, and Maine do not always map perfectly onto MSAs, but they do map onto NECMAs, which the ANES uses in New England. I merged in NECMAs using a crosswalk file from Geocorr 2000 for these states.

3. Match each MSA or NECMA (and as a result, each county) with the correct stratum used in the ANES sampling procedure. I included four tiers: (1.) Eight largest MSAs,

[3]Geocorr 2000 can be found at `https://mcdc.missouri.edu/applications/geocorr2000.html`

which are listed in Burns et al. (2001) and included in the sample with certainty; (2.) next 20 largest MSAs, which I identified using the lists in Burns et al. (2001) and Survey Research Center (1991), as well as Census data ranking MSAs by population; (3.) all remaining MSAs; and (4.) all counties not part of an MSA.

4. Aggregate up from the county level to the stratum level within each state by summing up population of each demographic combination within any county that is part of each stratum.

As in the case of the Gallup poststratification files, The resulting dataframe should have columns for race, sex, age group, education, state, and stratum that align to the groupings used by the ANES, as well as the number or share of the population represented by each combination of variables. This allows for easy merging of state-level variables. A poststratification matrix for the 2000 ANES is in the replication materials for this paper.

C.3 Availability of Poststratification Data

I have published the poststratification data used for this paper on the *State Politics and Policy Quarterly* Dataverse. This includes matrices for 1970 and 1980.

D Simulation Study: Pooling

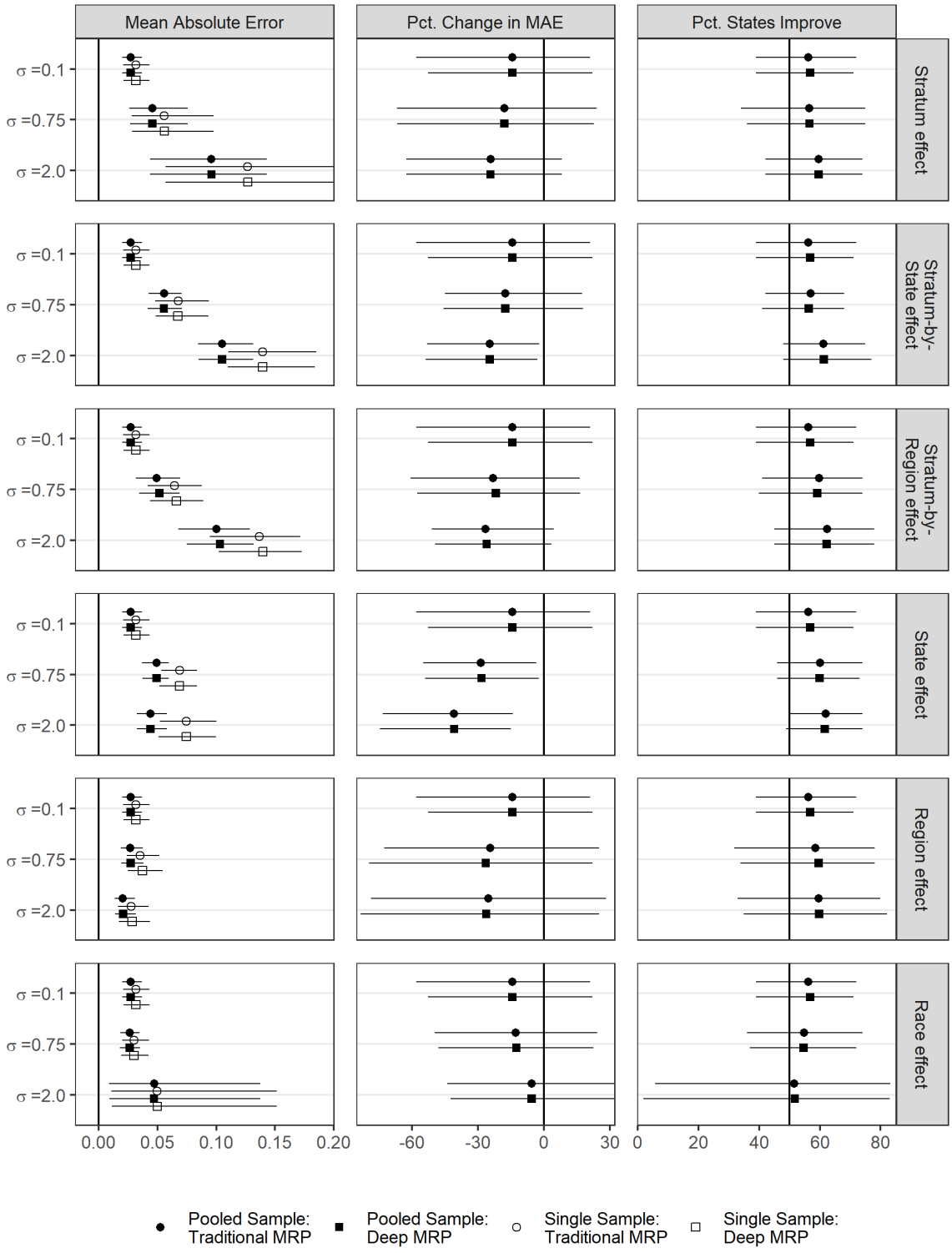
Here, I present results for a simulation study of pooling multiple surveys, discussed in Section 3.1 of the main text.

In the first study, I generated two versions of each sample: a "simple" sample that included 14 respondents from each of two clusters in every stratum×region combination, and a "pooled" sample that takes 14 respondents from each of four clusters in every stratum×region. This has the effect of doubling the overall sample size and the number of clusters, akin to pooling two surveys of equal size that utilize different clusters. For each sample, I estimate opinion using Traditional and Deep MRP.

Figure A3 shows the results of these simulations. As before, I report summary statistics from 100 simulations, varying the effect size σ for one variable at a time and holding all others constant at $\sigma = 0.1$. I report results using pooled surveys (filled circles and squares), as well as the single polls (hollow circles and squares).

The lefthand column of Figure A3 reports the MAE across simulations. As the effect sizes increase for stratum, stratum×region, and especially state and stratum×state, pooled surveys perform better than a single sample. The middle column reports the difference in MAE between pooled and single surveys as a percentage of the error in the single survey. A negative result means that the MAE decreases (improves) when pooling is used. Depending on the conditions shaping opinion, a pooled survey might reduce error by as much as 30% when the magnitude of the state effect is particularly high. The rightmost column reports the average share of states in each simulation whose modeled estimates of opinion get closer to true opinion. While there is a considerable amount of variation, the average state tends to improve when pooled surveys are used.

Figure A3: Results of Simulations: Pooling Surveys



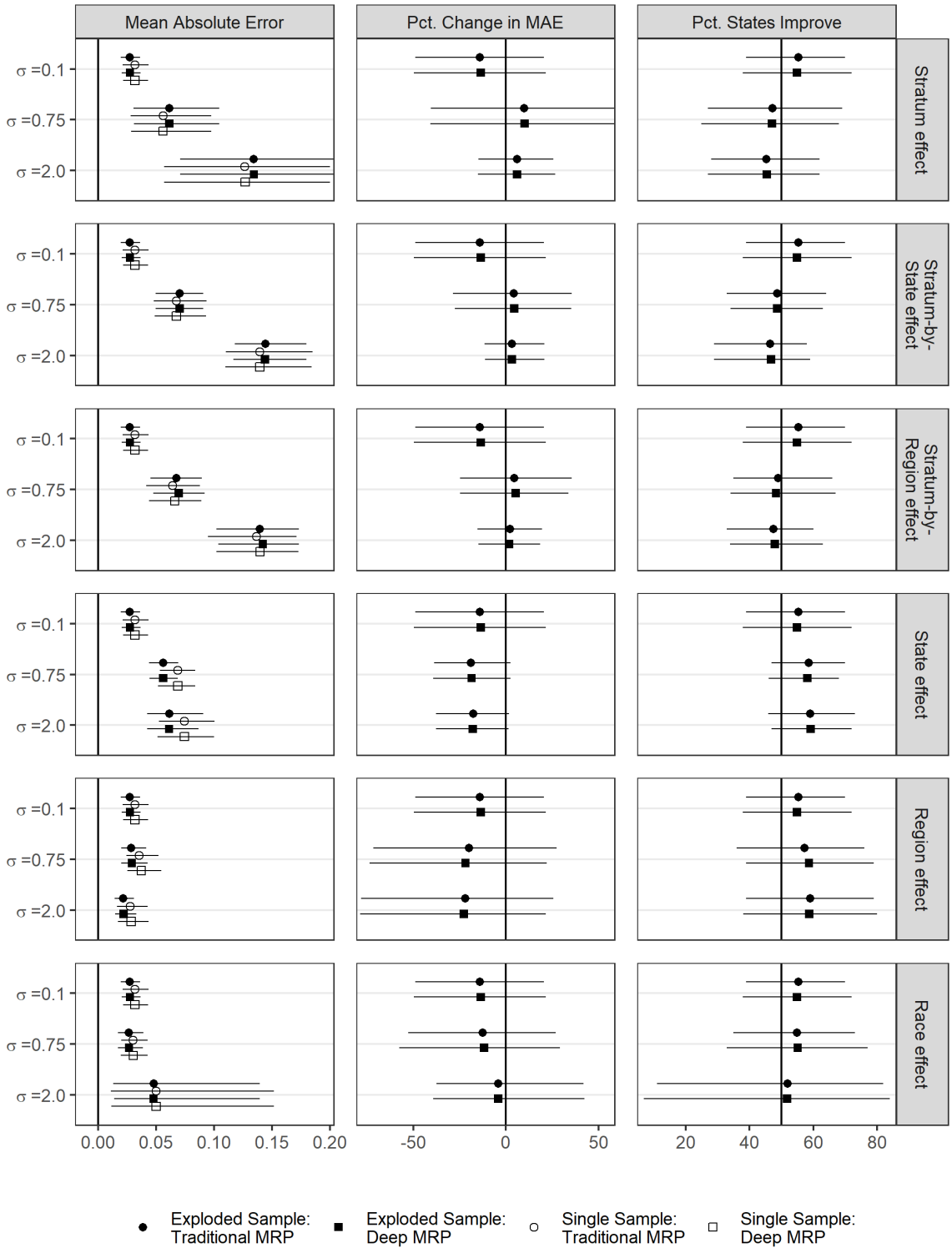
Note: Points report results for 100 simulations under a set of conditions. Simulations vary the standard deviation σ for one variable's effect on opinion at a time. All models were performed on clustered samples. Percent Change in MAE and Percent of States Improving are relative measures, comparing MRP or Deep MRP from a pooled sample with that of a single clustered sample. Error bars cover 95% of simulations. Axis limits are constrained to preserve readability.

These results suggest that, when possible, pooling multiple cluster-sampled surveys may improve estimation. This is particularly the case when opinion varies across states or by stratum within states, consistent with the problems associated with unrepresentative state-level subsamples in a single cluster-sampled poll.

It is important to note that pooling appears to dramatically improve estimation only when the *number* of clusters increases, not when the sample size within each does. In Figure A4, I use simulations to examine an alternative case in which the size of each cluster doubles but the overall number of clusters does not. That is, I compare two different ways to produce the same overall increase in sample size. When the cluster samples double in size, the estimates are not systematically improved over simple MRP—and may even get worse when the magnitude of the stratum effect on opinion is particularly high.

The simulation setup otherwise mirrors the one described above. However, rather than comparing a single sample to a pooled sample drawn from different clusters (i.e., doubling the number of clusters), here I "explode" the size of each cluster by doubling it.

Figure A4: Results of Simulations: Doubling Cluster Size



Note: Points report results for 100 simulations under a set of conditions. Simulations vary the standard deviation σ for one variable's effect on opinion at a time. All models were performed on clustered samples. Percent Change in MAE and Percent of States Improving are relative measures, comparing MRP or Deep MRP from a pooled sample with that of a single clustered sample. Error bars cover 95% of simulations. Axis limits are constrained to preserve readability.

Figure A4 reports results from the simulations. The first column reports mean absolute error (MAE) for each model. The second column reports this as a share of error in the baseline (single survey) model. This suggests that the potential gains from doubling the within-cluster sample size does very little to improve the quality of estimates; in some cases, it may even make estimates slightly worse. The third column shows the share of states whose estimates improve under a pooled sample.

E Presidential Elections Results

E.1 Data Sources

Poll data for the presidential election models come from five Gallup polls fielded in the final month of the 1968–1984 presidential elections. Respondent-level data and codebooks for all polls were downloaded from the Roper Center for Public Opinion Research's iPoll tool.

1968: Gallup Poll #1968–0770, fielded October 17-22, 1968. Respondents were randomly assigned to either express their presidential vote preferences via a secret or non-secret ballot. Secret ballot condition: "Suppose you were voting today for President of the United States. Here is a Gallup Poll ballot listing the candidates for this office. Will you please mark that ballot for the candidate you favor as you would in a real election if it were being held today—and then drop the folded ballot into this box" (N=805). Non-secret condition: "If the presidential election were being held today, which candidate would you vote for—Humphrey, the Democrat; Nixon, the Republican; or Wallace, the candidate of the American Independent Party?" (N=800).

1972: Gallup Poll #859, fielded October 13-16, 1972. Respondents were randomly assigned to either a secret or non-secret ballot condition. Secret ballot condition: "Here is a Gallup Poll secret ballot listing the candidates for this office. Suppose you were voting today for President of the United States. Will you please mark that secret ballot for the candidate you favor as you would in a real election if it were being held today—and then drop the folded ballot into this box" (N=758). Non-secret condition: "If the presidential election were being held today, which candidate would you vote for—Nixon, the Republican, or McGovern the Democrat?" (N=759).

1976: Gallup Poll #961, fielded October 22-25, 1976. All respondents were asked the same version of the question: "Now I'd like to get your honest opinion on this next question. It doesn't make any difference to me how you vote...I only want to get your views accurately: If the presidential election were being held TODAY, which candidate would you vote for—the Democratic candidates Carter and Mondale, or the Republican candidates Ford and Dole?" (N=1,505).

1980: Gallup Poll #1980–1163G, fielded October 10-13, 1980. All respondents were asked the same version of the question: "Now I'd like to get your honest opinion on this next question. It doesn't make any difference to me how you vote...I only want to get YOUR views accurately: If the presidential election were being held TODAY, which wuold you vote for—the Republican candidates Reagan and Bush, the Democratic candidates Carter and Mondale, or the Independent candidates Anderson and Lucey?" (N=1,593).

1984: Gallup Poll #1244G, fielded October 26-29, 1984. Respondents were randomly

assigned to either express their presidential vote preferences via a secret or non-secret ballot. Secret ballot: "Suppose you were voting today for President and Vice President of the United States. Here is a Gallup Poll secret ballot listing the candidates for these offices. Will you please mark that secret ballot for the candidates you favor today and then drop the folded ballot into the box." (N=739). Non-secret ballot: "Now I'd like to get your honest opinion on this next question. It doesn't make any difference to me how you vote...I only want to record your opinion accurately. If the presidential election were being held TODAY, which would you vote for—the Republican candidates, Reagan and Bush, or the Democratic candidates, Mondale and Ferraro?" (N=811).

ANES: For the pooled models in Section 3 of the paper, I incorporate additional responses from the American National Election Study (ANES). These samples incorporate responses from the ANES Time Series Cumulative Data File (American National Election Studies 2022). The file compiles responses to the following question from the 1968, 1972, 1976, 1980, and 1984 ANES: "So far as you know now, do you expect to vote in the national elections this coming November or not? (IF YES:) Who do you think you will vote for in the election for President?"

In all cases, I modeled two-party vote for the Democrat by coding the outcome variable as 1 for respondents who supported the Democrat, 0 for respondents who supported the Republican, and dropping those who supported third-party candidates or were undecided. Demographic and geographic variables (`race`, `female`, `agegrp`, `region`, `state`, and `educ`) also came from the Gallup polls. The share of the population that identify as evangelical Christians or Mormons (`evang` in the models below) came from the Churches and Church Membership in the United States studies (Glenmary Research Center 1974; Grammich et al. 2019).

The `stratum` variable used in the CMRP models maps a city size variable included in the Gallup data onto six of the seven strata used by Gallup to produce clustered random samples. Specifically, these strata were:

1. Cities of 1,000,000 people and over, plus suburbs
2. Cities of 250,000-499,999 and 500,000-999,999 people, plus suburbs
3. Cities of 50,000-99,999 and 100,000-249,999 people, plus suburbs
4. Cities of 2,500-4,999; 5,000-9,999; 10,000-24,999; and 25,000-49,999 people
5. Places under 2,500 people
6. Residents of farms and open-country non-farms

E.2 Models

Once data were cleaned, I produced seven separate models to estimate Democratic support as a share of the two-party vote at the state level. All multilevel models were fit in Stan using `rstanarm`, and poststratification was done using weights generated from IPUMS-NHGIS. Tables A1 and A2 report the formulas used to produce each model. All variables included in the models were used to poststratify. The 1968 and 1972 models differ from the later ones in that they do not include an education variable. This is because of census data limitations that made it impossible to produce a joint distribution of sex, race, education, and age at the stratum×state level in 1970.

Table A1: Model Syntax: Presidential Election Polls (1968 and 1972 Elections)

Model	**R (lme4/rstanarm) Syntax**
MRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|state) + evang`
Deep MRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|state) + (1|race:agegrp) + (1|female:agegrp) + (1|race:female:agegrp) + (1|race:state) + (1|female:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|agegrp:region) + evang`
CMRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|stratum) + (1|stratum:region) + (1|state) + (1|stratum:state) + evang`
Deep CMRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|state) + (1|stratum) + (1|stratum:region) + (1|stratum:state) + (1|race:agegrp) + (1|female:agegrp) + (1|race:female:agegrp) + (1|race:state) + (1|female:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|agegrp:region) + (1|race:stratum) + (1|female:stratum) + (1|agegrp:stratum) + evang`

Table A2: Model Syntax: Presidential Election Polls (1976, 1980, and 1984 Elections)

Model	**R (lme4/rstanarm) Syntax**
MRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|educ) + (1|state) + evang`
Deep MRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|educ) + (1|state) + (1|race:agegrp) + (1|race:educ) + (1|female:agegrp) + (1|agegrp:educ) + (1|race:female:agegrp) + (1|race:female:educ) + (1|female:agegrp:educ) + (1|race:state) + (1|female:state) + (1|educ:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|educ:region) + (1|agegrp:region) + evang`
CMRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|educ) + (1|stratum) + (1|stratum:region) + (1|state) + (1|stratum:state) + evang`
Deep CMRP	`(1|race) + (1|female) + (1|race:female) + (1|agegrp) + (1|region) + (1|educ) + (1|state) + (1|stratum) + (1|stratum:region) + (1|stratum:state) + (1|race:agegrp) + (1|race:educ) + (1|female:agegrp) + (1|agegrp:educ) + (1|race:female:agegrp) + (1|race:female:educ) + (1|female:agegrp:educ) + (1|race:state) + (1|female:state) + (1|educ:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|educ:region) + (1|agegrp:region) + (1|race:stratum) + (1|female:stratum) + (1|educ:stratum) + (1|agegrp:stratum) + evang`

The pooled models include only MRP and Deep MRP. However, they add a random effect for survey design (i.e., `(1|survey)`, to distinguish design effects from Gallup and ANES), which is not included in poststratification.

E.3 Detailed Results

Table A3 reports detailed results from the presidential election analysis with pooled samples. Each cell reports the mean absolute error (MAE) from running a particular MRP model on survey data from each election. Pooled samples are denoted as "pooled" and all other columns report single-sample results. MAE is computed by averaging the absolute difference between the estimated Democratic two-party vote share for each state and the observed "true" results in the election. The top panel reports MAE in which the estimates and ground truth are rescaled by subtracting out the national support for the Democratic candidate in the poll, or the national vote for the Democratic candidate in the election. These results—also reported in the main paper—account for national-level changes in the election over the weeks between polling and voting, as well as any bias inherent to the poll itself. The bottom panel reports MAE as the absolute difference between estimated and observed two-party vote share.

Table A3: Detailed Results: Presidential Elections with Pooled Sample

	Election Year	MRP	Deep MRP	Pooled MRP	Pooled Deep MRP
Scaled Vote	1968	0.0575	0.0526	0.0594	0.0540
	1972	0.0758	0.0500	0.0771	0.0517
	1976	0.0417	0.0401	0.0437	0.0437
	1980	0.0568	0.0478	0.0603	0.0477
	1984	0.0523	0.0482	0.0561	0.0462
	Avg.	**0.0568**	**0.0477**	**0.0593**	**0.0487**
Raw Vote	1968	0.0632	0.0620	0.0633	0.0622
	1972	0.0705	0.0563	0.0708	0.0547
	1976	0.0582	0.0461	0.0573	0.0525
	1980	0.1132	0.0855	0.1141	0.0846
	1984	0.0507	0.0449	0.0545	0.0431
	Avg.	**0.0712**	**0.0590**	**0.0720**	**0.0594**

Note: Cells report the mean absolute error of state-level two-party vote share for Democratic candidates, predicted using the listed model in each election year. The top panel compares state estimates as a deviation off the national vote against similarly scaled observed election results. The bottom panel compares estimated Democratic two-party vote share against observed vote share without transforming the data.

Table A4 reports detailed results from the presidential election analysis using CMRP. Each cell reports the mean absolute error (MAE) from running a particular MRP or CMRP model on survey data from each election.

Table A4: Detailed Results: Presidential Elections with CMRP

	Election Year	MRP	Deep MRP	CMRP	Deep CMRP
Scaled Vote	1968	0.0573	0.0590	0.0527	0.0545
	1972	0.0709	0.0719	0.0691	0.0710
	1976	0.0376	0.0421	0.0377	0.0410
	1980	0.0507	0.0560	0.0508	0.0557
	1984	0.0563	0.0613	0.0559	0.0592
	Avg.	**0.0546**	**0.0581**	**0.0532**	**0.0563**
Raw Vote	1968	0.0632	0.0633	0.0623	0.0622
	1972	0.0705	0.0708	0.0709	0.0692
	1976	0.0582	0.0573	0.0585	0.0577
	1980	0.1132	0.1141	0.1101	0.1098
	1984	0.0507	0.0545	0.0513	0.0559
	Avg.	**0.0712**	**0.0720**	**0.0706**	**0.0709**

Note: Cells report the mean absolute error of state-level two-party vote share for Democratic candidates, predicted using the listed model in each election year. The top panel compares state estimates as a deviation off the national vote against similarly scaled observed election results. The bottom panel compares estimated Democratic two-party vote share against observed vote share without transforming the data.

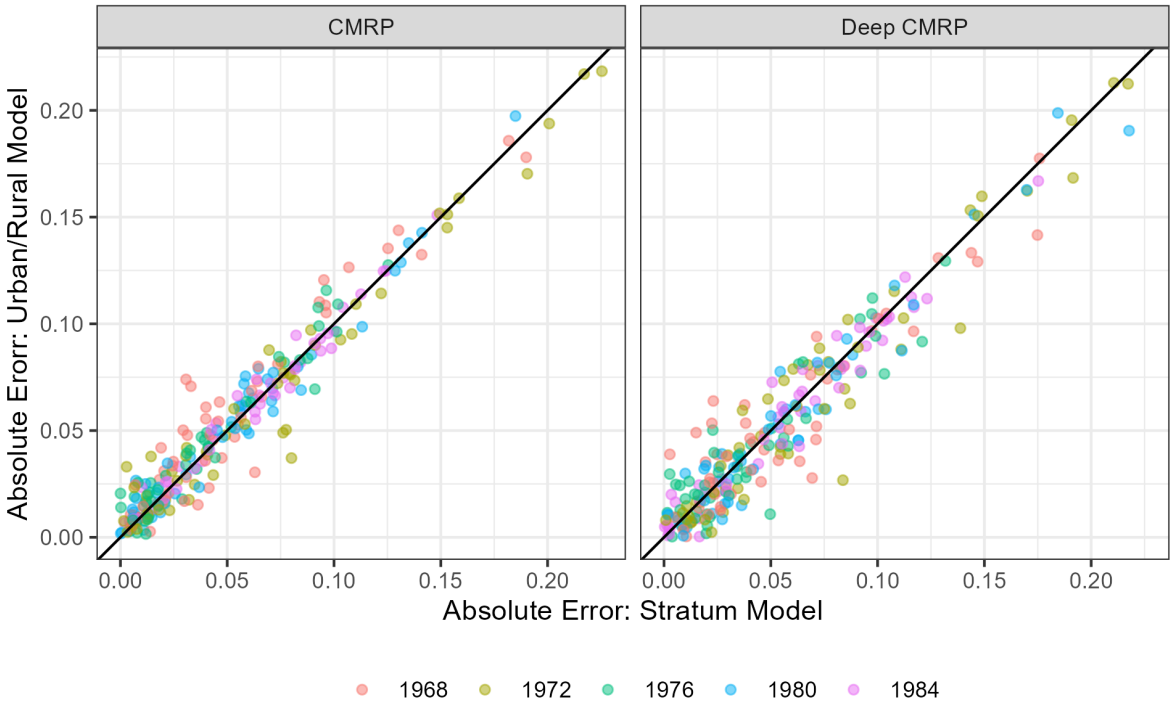
E.4 Alternative Coarsened MRP Model

One reasonable alternative in cases where detailed geographic variables corresponding to how a survey was cluster-sampled are unavailable may be to include some other, higher-level geographic variable in lieu of `stratum`. For example, one could code whether respondents live in urban or rural areas, and then model and poststratify opinion using this information.[4]

In this appendix, I report results for estimating two-party presidential vote replacing `stratum` with `urban`. In this case, I use the data about city size to identify respondents as urban or rural. I designate places with populations of 50,000 or more as urban, and smaller towns and rural and farmland area as rural. The models otherwise follow those in Tables A1 and A2.

[4]Caughey and Warshaw (2018) include a random effect for urban-ness to their MRP estimates, much of which come from a historical period in which cluster-sampling was used. Likewise, common approaches for estimating sub-state (e.g., district-level) opinion include information about the share of target areas that are urban (Tausanovitch and Warshaw 2013).

Figure A5: Comparison with Urban/Rural MRP Setup



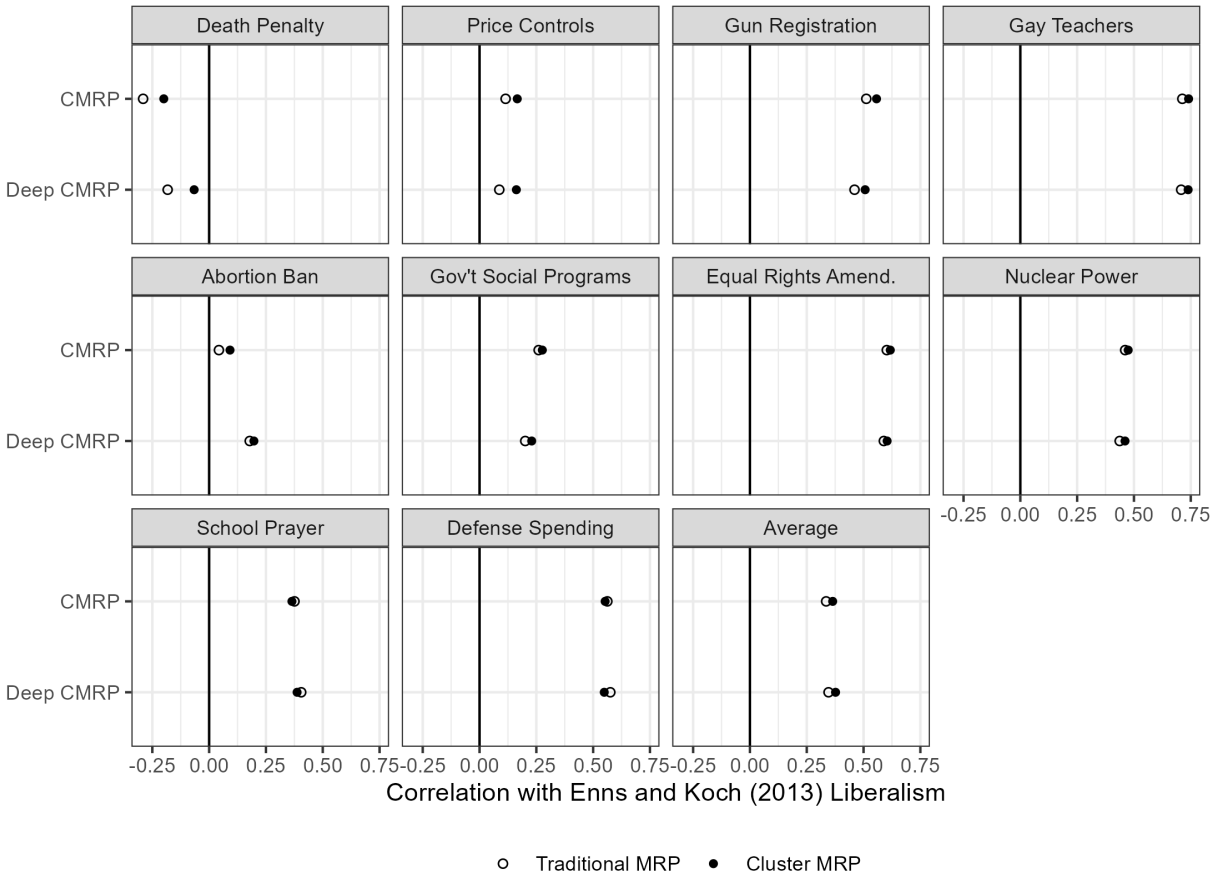
Note: Points compare mean absolute error for state-level estimates of two-party Democratic presidential vote from two models. The x-axis reports error for the primary CMRP approach discussed in the paper using Gallup polling strata; the y-axis substitutes strata with whether respondents live in an urban or rural area.

Figure A5 compares the MAE of individual state-year estimates from the two models, with the error from the standard CMRP model on the x-axis and the MAE from the model using urban/rural on the y-axis. I find little difference in the errors for either CMRP or Deep CMRP, at least in the case of estimating presidential vote.

F Gallup Issues Results

The September 1980 Gallup poll included 10 questions about particular policy issues. Using CMRP methods, I estimated opinion for each issue in the poll. The models include Republican vote share but are otherwise identical to those used in the presidential election polls. I then compared the correlation between CMRP and the Enns and Koch (2013) state liberalism scores with similar traditional MRP models. I recoded all questions so that responses of $y_i = 1$ correspond to the liberal position so that higher correlation is always better. The full text of the questions from the Gallup poll, as well as the specification of all models is in Appendix F.

Figure A6: CMRP on Gallup Issue Questions



Note: Points represent the correlation between estimated issue opinion and state-level liberalism from Enns and Koch (2013). All issues are recoded so that a higher correlation always indicates improvement.

Figure A6 shows the results of this validation effort. Shaded points indicate the correlation between state-level liberalism and issue opinion estimated with CMRP methods, while hollow points correspond to traditional or deep MRP. On average, CMRP methods increase the correlation between opinion and liberalism by approximately 0.03, depending on the model. Notably, on some issues, such as price controls or the death penalty, CMRP produces large improvements, although on others the gains are more modest. Regardless, the largest improvements from using CMRP outweigh the slight reductions in accuracy on issues such as school prayer.

F.1 Data Source and Issue Questions

Polling data come from the Gallup Poll (#1162G) fielded September 12–15, 1980. All questions were asked of all respondents (N=1,602). Respondent-level data and codebooks for all polls were downloaded from the Roper Center for Public Opinion Research's iPoll tool.

In all cases, I modeled state-level support for the question, and then recoded the estimates so that greater support for the issue corresponded to a liberal position (these issues are identified below). This allows me to compare the results against state-level estimates of latent policy liberalism. As in the presidential vote question, I obtained demographic and geographic variables from the Gallup survey data, and used Gallup's city size variable to produce `stratum`. Religion data used in the model come from Grammich et al. (2019). State Republican vote share in 1980 (`repvote`) is from Leip (2021). Below are listed the exact questions asked in the survey:

Price Controls: "Would you favor or oppose having the government bring back wage and price controls?"

Death Penalty: (Recoded.) "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... A mandatory death penalty for anyone convicted of murder."

Gun Registration: "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... Federal registration of all firearms."

Equal Rights Amendment: "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... The ERA—giving women equal rights and equal responsibilities."

Government Social Programs: "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... Government social programs as a way to deal with social problems."

Gay Teachers: "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... Allowing homosexuals to teach in public schools."

Nuclear Power: (Recoded.) "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... More nuclear power plants."

Abortion Ban: (Recoded.) "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... A ban on all abortions."

Defense Spending: (Recoded.) "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... Increased spending for defense."

School Prayer: (Recoded.) "This card lists various proposals being discussed in this country today. Would you please tell me whether you generally favor or generally oppose each of these proposals. ... Requiring prayers in the public schools."

F.2 Models

Once data were cleaned, I produced seven separate models of each question to estimate support for the policy. All multilevel models were fit in Stan using `rstanarm`, and post-stratification was done using weights generated from IPUMS-NHGIS data. Table A5 reports

the formulas used to produce the models. All variables included in the models were used to poststratify.

Table A5: Model Syntax: Gallup Policy Issues (September 1980)

Model	**R (lme4/rstanarm) Syntax**
MRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|state) + evang + repvote`
Deep MRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|state) + (1|race:agegrp) + (1|race:educ) + (1|female:agegrp) + (1|agegrp:educ) + (1|race:female:agegrp) + (1|race:female:educ) + (1|female:agegrp:educ) + (1|race:state) + (1|female:state) + (1|educ:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|educ:region) + (1|agegrp:region) + evang + repvote`
CMRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|stratum) + (1|stratum:region) + (1|state) + (1|stratum:state) + evang + repvote`
Deep CMRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|state) + (1|stratum) + (1|stratum:region) + (1|stratum:state) + (1|race:agegrp) + (1|race:educ) + (1|female:agegrp) + (1|agegrp:educ) + (1|race:female:agegrp) + (1|race:female:educ) + (1|female:agegrp:educ) + (1|race:state) + (1|female:state) + (1|educ:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|educ:region) + (1|agegrp:region) + (1|race:stratum) + (1|female:stratum) + (1|educ:stratum) + (1|agegrp:stratum) + evang + repvote`

F.3 Detailed Results

Table A6 shows the results from analyzing issue questions in the Gallup Poll. Each cell reports the Spearman correlation coefficient for the relationship between state opinion estimates from a given MRP or CMRP model and the liberalness of state electorates as estimated by Enns and Koch (2013). Because all state-level opinion estimates from the polls have been rescaled such that higher support indicates more liberal opinion, a higher correlation can always be interpreted as a more accurate prediction.

Table A6: Detailed Results: Gallup Issues and Liberalism

	MRP	Deep MRP	CMRP	Deep CMRP
Death Penalty	-0.291	-0.183	-0.200	-0.066
Price Controls	0.115	0.087	0.166	0.162
Gun Registration	0.512	0.460	0.557	0.507
Gay Teachers	0.713	0.708	0.740	0.738
Abortion Ban	0.043	0.179	0.092	0.197
Gov't Social Programs	0.260	0.201	0.276	0.230
Equal Rights Amend.	0.602	0.589	0.618	0.604
Nuclear Power	0.461	0.437	0.475	0.461
School Prayer	0.376	0.405	0.364	0.387
Defense Spending	0.562	0.576	0.553	0.549
Average	**0.335**	**0.346**	**0.364**	**0.377**

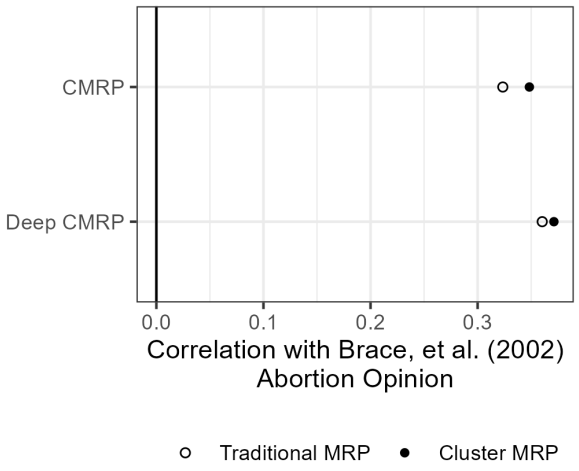
Note: Cells report Spearman correlations between estimated issue opinion and state-level liberalism from Enns and Koch (2013). For all issues, higher correlation indicates a better estimate.

Although I find only modest improvements in most cases, this may in part be a function of noise inherent to using a more general ideological scale—which was itself constructed from public opinion polls—as a ground truth measure of opinion. This underscores a broader challenge with validating MRP in particular cases. MRP is generally most useful in understanding public opinion on specific issues where survey data do not exist; as a result, it is generally impossible to determine whether a particular MRP model works well in a specific case.

F.4 Validating with Abortion Opinion

I now take a closer look at the abortion polls reported in the main paper and provide validations using two measures of "ground truth" opinion. First, in Figure A7, I show Spearman correlations of opinion estimates from traditional MRP and CMRP models with state abortion liberalism scores from Brace et al. (2002). These scores were generated by pooling General Social Survey (GSS) respondents over the period from 1974-1998, and none of the questions in the index asked about an outright abortion ban. As a result, they may be an imperfect proxy for support for banning abortion. Nevertheless, I show slight increases on the correlations.

Figure A7: CMRP Improvement in Abortion Opinion Estimates

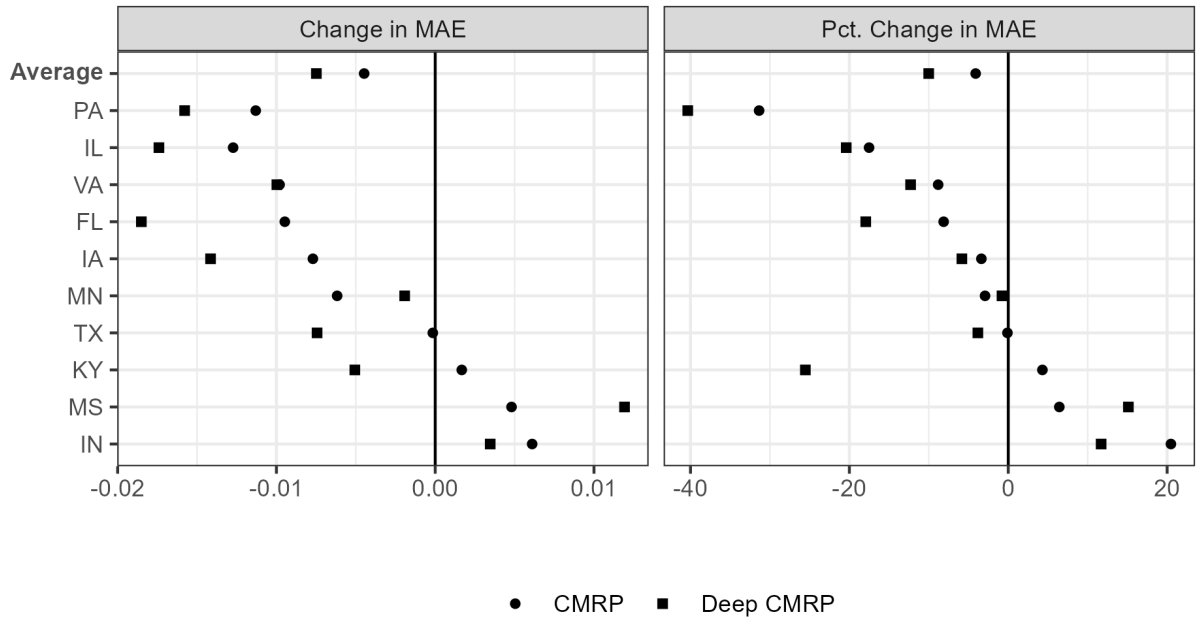


Note: Points represent the Spearman correlation between estimated abortion opinion and abortion liberalism from Brace et al. (2002).

A better approach at estimating ground-truth opinion would be to turn to large-sample state polls asking similar questions to the Gallup survey in 1980. This is made difficult by the limited number of state polls routinely in the field (Parry, Kisida and Langley 2008). I have identified and collected survey data from 10 such cases, where state-level pollsters included a question (or series of questions) about whether abortion should be illegal. I detail the questions and surveys below, though I note that none of the state poll questions are identical to the one posed by Gallup.

Figure A8 reports changes in the MAE comparing CMRP methods against analogous traditional MRP models. Although I find that the reductions in MAE are relatively modest, they represent 4% (CMRP) or 10% (Deep CMRP) reductions in the MAE compared to the baseline. As in the case of the validation using presidential votes, this is consistent with error reductions from using machine learning with MRP (Ornstein 2020; Goplerud 2023).

Figure A8: Abortion Estimates and State Polls



Note: Points report improvement using a CMRP method on abortion polls versus an analogous traditional MRP method. MAE is computed against a baseline from state-level polls taken during the years 1980-1986 asking some variation of a question as to whether abortion should be made illegal.

The specific state-level surveys and questions used are listed below. All data were downloaded from Roper iPoll or the National Network of State Polls Dataverse.[5] Because most polls did not include survey weights, I modeled state opinion from the poll using only available demographic characteristics (race, sex, education, and age) and poststratified using Census data. This approach is similar to MRP as an adjustment for unrepresentative surveys (Lopez-Martin, Phillips and Gelman 2021), though notably excludes any geographic predictors or continuous variables. I note that several other states asked questions about abortion; however, I focus here only on questions that asked whether abortion should be made illegal, rather than those about personal opinions as to the morality of abortion or about constitutional amendments which are sufficiently different from the Gallup question wording as to render unhelpful comparisons.

Florida: Florida Annual Policy Survey (February 1986). Respondents were asked a series of eight abortion questions: "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion... If there is a strong chance of serious defect in the baby? ... If the woman wants it for any reason? ... If she is married and does not want any more children? ... If the woman's own health is seriously endangered by the pregnancy? ... If the family has a very low income and cannot afford any more children? ... If she became pregnant as a result of rape? ... If she is not married and does not want to

[5]The dataverse can be found at `https://dataverse.unc.edu/dataverse/nnsp`.

marry the man? ... If the woman is less than three months pregnant?" I code respondents as supportive of banning all abortions if they answered "yes" to any five of these conditions.

Illinois: CBS News exit poll (November 6, 1984). "Should abortion be legal?"

Indiana: Indiana University Center for Survey Research (November/December 1983). "Do you think that abortion should be ... Legal with no restrictions ... Legal with restrictions ... Illegal"

Iowa: Iowa Poll (*Des Moines Register* and Tribune Co.; September 1884). "Which ONE of the following statements best describes your opinion toward abortion? ... All abortions should be illegal ... Abortions should be legal on a highly limited basis—only to save the mother's life ... Abortions should be illegal on a somewhat limited basis—when any one of a variety of mental or physical problems is anticipated ... All abortions should be legal."

Kentucky: University of Kentucky Survey Research Center (Fall 1981). "Do you personally believe that abortion is wrong?" Respondents who answered "Yes" or "Don't know" or refused to answer were then asked, "Do you think abortion should be illegal?" I take care to code only those who were asked if it should be illegal and responded that it should be illegal as supportive of a ban.

Minnesota: The Minnesota Poll (*Minneapolis Tribune*; April/May 1980). Respondents were asked a series of seven questions about abortion: "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion ... If there is a strong chance of serious defect in the baby? ... If she is married and does not want any more children? ... If the woman's health is seriously endangered by the pregnancy? ... If the family has a very low income and cannot afford any more children? ... If she became pregnant as a result of rape? ... If she is not married and does not want to marry the man? ... If the woman wants it for any reason?" I code respondents as supportive of banning all abortions if they answered "yes" to any five of these conditions.

Mississippi: CBS News exit poll (November 6, 1984). "Should abortion be legal?"

Pennsylvania: CBS News exit poll (November 6, 1984). "Should abortion be legal?"

Texas: Texas Poll (Texas A&M University; Spring 1984). "Do you think abortions should be legal under any circumstances, legal only under certain circumstances, or illegal in all circumstances?"

Virginia: Virginia Commonwealth University Survey Research Laboratory (October/November 1985). "There has been a lot of discussion about abortion in recent years. Do you think it should be legal for any unwanted pregnancy, legal only under certain conditions, or not legal at all?"

G ANES Results

To test CMRP under a different cluster-sampling procedure, I turned to the 2000 ANES. This offers an opportunity to test CMRP under a different cluster-sampling procedure than Gallup's. The ANES identifies PSUs using a population-ranked list of Metropolitan Statistical Areas (MSAs) (Burns et al. 2001).[6] I compare CMRP-estimated opinion with results from

[6]Specifically, the ANES includes each of the eight largest MSAs with certainty; half of the next 20 largest; and the remaining MSAs and non-MSA areas selected with probabilities weighted by their population. More detail on the cluster-sampling procedure can be found below.

the 2000 National Annenberg Election Survey (NAES), which included 58,373 respondents and several similar topics to the ANES.

For each of six similar questions asked in both the ANES and NAES, I used CMRP to estimate public opinion in each state and compared it against average opinion from the NAES. However, this approach has two limitations. First, question wordings in the ANES and NAES are not identical, nor are the dates that questions were in the field. Second, the 2000 ANES includes a small sample, with just 999 chosen by cluster sampling and the rest via random-digit dialing.

G.1 Data Sources and Issue Questions

Polling data for the models come from the 2000 American National Election Study (ANES), which included a set of cluster-sampled respondents interviewed in-person (N=999), as well as a smaller telephone sample produced via random-digit dialing. For the purposes of testing CMRP, I only use the face-to-face sample. Data were obtained from the ANES archives.

I compare the modeled estimates of state-level opinion from the ANES against the average responses in each state to similar questions asked by the 2000 National Annenberg Election Survey (NAES), which used a random sample of 58,373 respondents (sample sizes for particular questions vary, as the windows in which they were fielded are not identical).

Demographic and geographic data come from the ANES response file. In order to produce the `stratum` variable, I used information from Burns et al. (2001) and Survey Research Center (1991) to identify the procedure used by the ANES to assign MSAs to PSUs. Specifically, I identified four tiers of places and matched them to the MSA variable included in the 2000 ANES data release:

1. Largest eight metropolitan areas
2. Next 10 largest metropolitan areas
3. Remaining MSAs
4. Remaining non-MSAs

Evangelical and mormon share of the population comes from the Religious Congregations and Membership Studies (Grammich et al. 2019), and state Republican vote share in the 2000 election is from Leip (2021).

G.1.1 Restrict Abortion

ANES: "There has been some discussion about abortion during recent years. Which one of the opinions on this page best agrees with your view? You can just tell me the number of the opinion you choose."

1. By law, abortion should never be permitted.

2. The law should permit abortion only in the case of rape, incest, or when the woman's life is in danger.

3. The law should permit abortion for reasons other than rape, incest, or danger to the woman's life, but only after the need for the abortion has been clearly established.

4. By law, a woman should always be able to obtain an abortion as a matter of personal choice.

I recoded responses 1 and 2 as "restrict abortion", while responses 3 and 4 were recoded as "do not restrict abortion."

NAES: "Do you personally favor or oppose making it harder for a woman to get an abortion?" (Fielded April 4-September 7, 2000; N=21,558)

1. Favor
2. Oppose

G.1.2 School Vouchers

ANES: Respondents were randomly assigned to one of two versions of the question: "Do you favor or oppose a school voucher program that would allow parents to use tax funds to send their children to the school of their choice, even if it were a private school?" OR "Do you favor or oppose a school voucher program that would allow parents to use tax funds to send their children to the school of their choice, even if it were a private school, or haven't you thought much about this?"

1. Favor school voucher program
2. Oppose school voucher program

NAES: "Do you personally favor or oppose using government money to help some parents send their children to private schools?" (Fielded April 4-September 7, 2000; N=22,554)

1. Favor
2. Oppose

G.1.3 Death Penalty

ANES: "Do you favor or oppose the death penalty for persons convicted of murder?"

1. Favor
2. Oppose

NAES: "Do you personally favor or oppose the death penalty for some crimes?" (Fielded April 4-November 27, 2000; N=32,877)

1. Favor
2. Oppose

G.1.4 Gays in the Military

ANES: "Do you think homosexuals should be allowed to serve in the United States Armed Forces or don't you think so?"

1. Homosexuals should be allowed to serve
2. Homosexuals should not be allowed to serve

NAES: "Do you personally favor or oppose allowing homosexuals to serve openly in the United States military?" (Fielded April 4-November 27, 2000; N=31,048)

1. Favor
2. Oppose

G.1.5 Increase Social Security

ANES: "Next I am going to read you a list of federal programs. For each one, I would like you to tell me whether you would like to see spending increased or decreased. ... What about social security? (Should federal spending on social security be increased, decreased, or kept the same?)

1. Increased
2. Decreased
3. Kept about the same

I coded responses either as "increased" or as "not increased" for respondents who answered that social security spending should be kept the same or decreased.

NAES: "Social security benefits—should the federal government spend more money on this, the same as now, less, or no money at all?" (Fielded December 14, 1999-December 12, 2000; N=54,089)

1. More
2. Same
3. Less
4. None

I coded responses either as "increased" or as "not increased" for respondents who answered that spending should be kept the same, reduced, or cut altogether.

G.1.6 Gun Restrictions

ANES: "Do you think the federal government should make it more difficult for people to buy a gun than it is now, make it easier for people to buy a gun, or keep these rules about the same as they are now?"

1. More difficult
2. Make it easier
3. Keep these rules about the same

I coded responses either as "restrict" for those who want to make it more difficult to buy a gun, or "not restrict" for those who want to make it easier or keep rules the same.

NAES: "Restricting the kinds of guns that people can buy—should the federal government do more about this, the same as now, less, or nothing at all?" (Fielded December 14, 1999-December 12, 2000; N=54,599)

1. More
2. Same
3. Less
4. None

I coded responses either as "restrict" for those who want the federal government to be more restrictive, and "not restrict" for those who want the government to do the same as now, less, or nothing.

G.2 Models

Once data were cleaned, I produced seven models of each question to estimate support for the policy. Multilevel models were fit in Stan using `rstanarm`, and poststratification was done using weights generated from IPUMS-NHGIS census data. Table A7 reports the formulas used to produce the models. All variables included in the models were used to poststratify.

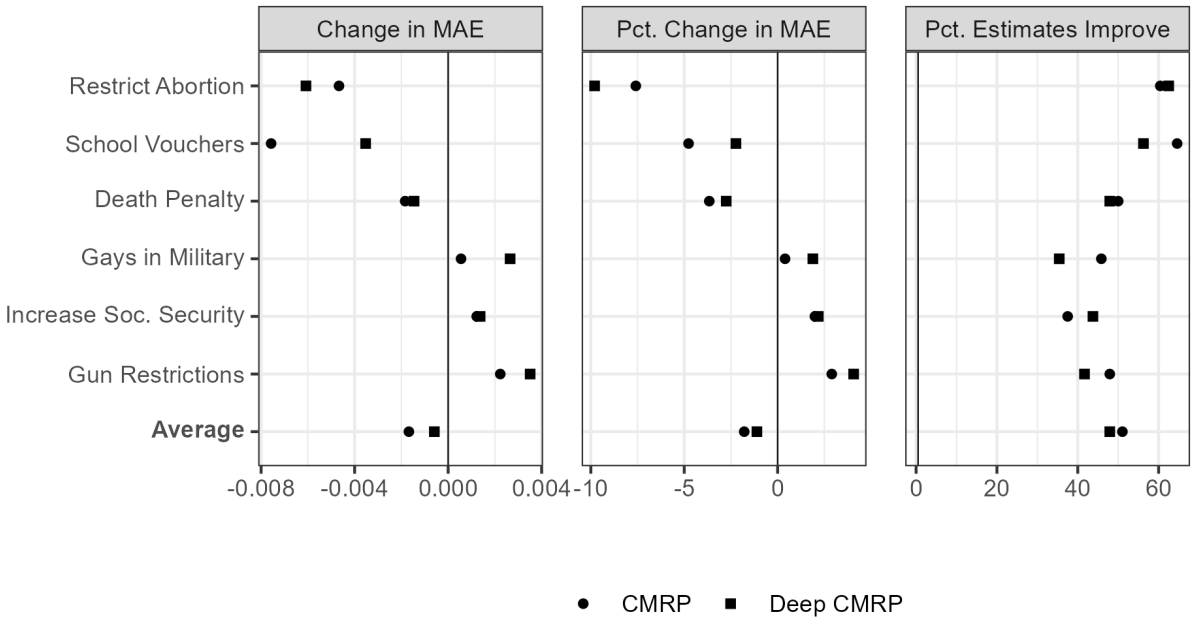
Table A7: Model Syntax: ANES 2000

Model	**R (lme4/rstanarm) Syntax**
MRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|state) + evang + repvote`
Deep MRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|state) + (1|race:agegrp) + (1|race:educ) + (1|female:agegrp) + (1|agegrp:educ) + (1|race:female:agegrp) + (1|race:female:educ) + (1|female:agegrp:educ) + (1|race:state) + (1|female:state) + (1|educ:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|educ:region) + (1|agegrp:region) + evang + repvote`
CMRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|stratum) + (1|stratum:region) + (1|state) + (1|stratum:state) + evang + repvote`
Deep CMRP	`(1|race) + (1|female) + (1|race:female) + (1|educ) + (1|agegrp) + (1|region) + (1|state) + (1|stratum) + (1|stratum:region) + (1|stratum:state) + (1|race:agegrp) + (1|race:educ) + (1|female:agegrp) + (1|agegrp:educ) + (1|race:female:agegrp) + (1|race:female:educ) + (1|female:agegrp:educ) + (1|race:state) + (1|female:state) + (1|educ:state) + (1|agegrp:state) + (1|race:region) + (1|female:region) + (1|educ:region) + (1|agegrp:region) + (1|race:stratum) + (1|female:stratum) + (1|educ:stratum) + (1|agegrp:stratum) + evang + repvote`

G.3 Detailed Results

Figure A9 reports results from the ANES analysis. The leftmost column reports the difference in MAE between the CMRP model and a corresponding MRP model, while the second column reports this opinion difference as a percent of the MAE for traditional MRP. On average, using CMRP methods produces a 1.2% decrease in MAE across the six questions. The rightmost column shows the share of states whose estimates improve using CMRP compared to traditional MRP. On average, about half of states' estimates improve when CMRP is used.

Figure A9: Using CMRP with the ANES



Note: Points represent improvement from using a CMRP method on issue questions in the 2000 ANES, compared to an analogous traditional MRP method. MAE is computed by taking the absolute difference of estimated support for the issue in the ANES and support from the NAES. Negative values represent improvement (reduction) in MAE. Pct. Change in MAE reports the same MAE as a percentage of MAE under the baseline (traditional MRP) models. Pct. States Improve reports the share of states whose estimates under CMRP were more accurate than under traditional MRP.

Table A8 reports detailed results from analyzing the ANES. Each cell shows the MAE from running a particular MRP or CMRP model on survey data. MAE is computed relative to the matching NAES question.

Table A8: Detailed Results: ANES

	MRP	Deep MRP	CMRP	Deep CMRP
Restrict Abortion	0.0615	0.0621	0.0568	0.0560
Death Penalty	0.0502	0.0530	0.0484	0.0515
Gays in the Military	0.1387	0.1406	0.1393	0.1432
Gun Restrictions	0.0769	0.0862	0.0792	0.0897
Increase Social Security	0.0612	0.0628	0.0624	0.0642
School Vouchers	0.1588	0.1581	0.1512	0.1545
Average	**0.0912**	**0.0938**	**0.0895**	**0.0932**

Note: Cells report the mean absolute error of estimated opinion from the ANES, for which the NAES is used as a ground truth.

References

American National Election Studies. 2022. "ANES Time Series Cumulative Data File.".

Brace, Paul, Kellie Sims-Butler, Kevin Arceneaux and Martin Johnson. 2002. "Public Opinion in the American States: New Perspectives Using National Survey Data." *American Journal of Political Science* 46(1):173–189.

Burns, Nancy, Donald R. Kinder, Steven J. Rosenstone and Virginia Sapiro. 2001. *National Election Studies, 2000: Pre-/Post-Election Study Codebook*. Ann Arbor, MI: University of Michigan, Center for Political Studies.

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

Caughey, Devin and Christopher Warshaw. 2018. "Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014." *American Political Science Review* 112(2):249–266.

Enns, Peter K. and Julianna Koch. 2013. "Public Opinion in the U.S. States: 1956 to 2010." *State Politics & Policy Quarterly* 13(3):349–372.

Glenmary Research Center. 1974. "Churches and Church Membership in the United States, 1971 (States)." https://www.thearda.com/Archive/Files/Descriptions/RCMSMGST.asp.

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

Grammich, Clifford, Kirk Hadaway, Richard Houseal, Dale E. Jones, Alexei Krindatch, Richie Stanley and Richard H. Taylor. 2019. "Longitudinal

Religious Congregations and Membership File, 1980-2010 (State Level)." https://www.thearda.com/Archive/Files/Descriptions/RCMSMGST.asp.

Lax, Jeffrey R. and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53(1):107–121.

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.

Leip, David. 2021. "Dave Leip's Atlas of U.S. Presidential Elections.".

Lopez-Martin, Juan, Justin H. Phillips and Andrew Gelman. 2021. "Multilevel Regression and Poststratification Case Studies." https://bookdown.org/jl5522/MRP-case-studies/.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler and Steven Ruggles. 2021. "IPUMS National Historical Geographic Information System: Version 16.0.".

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

Parry, Janine A., Brian Kisida and Ronald E. Langley. 2008. "The State of State Polls: Old Challenges, New Opportunities." *State Politics & Policy Quarterly* 8(2):198–216.

Survey Research Center. 1991. *Technical Description: 1990 National Election Study Sample Design*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

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