Estimating Subnational Opinion with Cluster-Sampled Polls: Challenges and Suggestions

Alissa Stollwerk*

July 20, 2013

Abstract

Multilevel regression with poststratification (MRP) has become widely used in political science to estimate subnational opinion from national polls. This method takes into account both demographic and state-level effects and has greatly expanded knowledge of public opinion at the subnational level. In this paper, I assess the use of MRP on two cluster-sampled polls, the American National Election Studies (ANES) and the General Social Survey (GSS). I discuss concerns that cluster sampling raises and demonstrate that MRP can produce problematic results when used on a typical cluster-sampled poll. I evaluate several ways to improve upon MRP with cluster sampling. I find that adding a state-level predictor to the model, pooling surveys across years and sampling frames, and incorporating additional geographic information can all improve MRP’s performance on cluster-sampled polls, but that caution should still be used when applying MRP to cluster-sampled data.

*Alissa Stollwerk is a doctoral candidate in the Political Science Department at Columbia University. I would like to thank Jeffrey Lax and Justin Phillips for their insightful comments at all stages of this work, and I thank Andrew Gelman for helpful comments and his assistance with the General Social Survey data as well. I am also grateful to Jim Gibson for his advice on the 1980s joint SRC-NORC sampling frame.
Introduction

Over the last two decades, interest in state public opinion has increased as political scientists have consistently shown that state public opinion is meaningful and that it informs public policy and elections at the state level. Unfortunately, polls that measure public opinion at the state level are few and far between, especially when one is looking for consistency across the fifty states. In response to this paucity of polling data, researchers have used a variety of strategies to derive subnational public opinion estimates from national polls, primarily applying statistical techniques such as disaggregation and simulation. As simulation techniques become more popular, researchers are using simulation methods to generate state-level opinion estimates from a wide range of public opinion polls, many with different sampling methods. This paper examines how one simulation method, multilevel regression with poststratification, performs on cluster-sampled polls, focusing on how a geographically-clustered sample may impact the accuracy and efficiency of sub-national opinion estimates.

As early as the 1960s, political scientists attempted to use simulation, using the demographic characteristics of the survey respondents and of the populations of each state to generate state-level estimates of public opinion from national polls (Pool et al 1965, Weber and Shaffer 1972-3). These early simulations, however, were criticized for ignoring potential geographic variation. In these studies, demographics were the only variables, and they were estimated using simple linear or fixed effects. Erikson, Wright, and McIver’s work on state politics directly criticized this assumption, arguing that geography mattered. In their book Statehouse Democracy, they showed that state public opinion was not simply an aggregation of the opinion of demographic categories, but contained distinct information connected to the geography itself. In fact, they argued that “most of the variance in state partisanship and state ideology is due to state-to-state differences that cannot be accounted for by the demographic variables measured here” (61). Rather than using simulation methods, which relied only on demographics, they used a technique known as disaggregation, combining several polls that one could then break down by state, to demonstrate this geographic variation. By emphasizing the importance of geography in
forming state-level opinion, they showed that geographic effects could not be ignored.

Recently, however, political scientists have returned to simulation methods and developed more sophisticated techniques that utilize meaningful geographic information as well as the traditional demographic variables. Gelman and Little (1997) developed a multilevel modeling technique, known as multilevel regression with poststratification, or MRP, which creates reliable and valid measures of public opinion at the state level from a single national poll of roughly 1500 respondents. Unlike earlier models of simulation, MRP relies on both demographic and geographic effects in simulating public opinion. MRP also employs partial pooling across subgroups, and Park, Gelman, and Bafumi (2006) showed that partial pooling across states outperformed both running individual models for each state (no pooling) and a model with complete pooling in which only demographic information was used in the modeling phase. In their work, Park et al test MRP by using it to reproduce the same state-level ideology measures that disaggregation produced in *Statehouse Democracy*. They find that their measures are nearly identical, and show that in areas where one has seen change over time, such as in a rightward shift in Republican partisan identification during the time series in question, MRP outperforms disaggregation. Lax and Phillips (2009b, 2013) further test how different forms of MRP perform against disaggregation and conclude even a simple MRP model often outperforms disaggregation in producing valid state-level estimates.

MRP is quickly becoming an approachable and accepted technique that researchers can use to estimate both state and local public opinion. This growing acceptance, however, warrants some caution. As researchers learn how to conduct MRP and use it on a wide range of polls, it is important that we understand how MRP functions on polls that rely on different sampling methods. Specifically, both Park, Gelman, and Bafumi (2006) as well Lax and Phillips (2009b) evaluate MRP using polls that rely on a full probability sample. As MRP gains in popularity, however, political scientists will be motivated to use it on two of the discipline’s largest surveys, the American National Election Study (ANES) and the General Social Survey (GSS).1 These polls, however,

1 Some political scientists have already used multilevel modeling to develop subnational estimates from these surveys, including Berkman and Plutzer (2006). Anticipating that cluster sampled polls may be problematic, Berkman and Plutzer do drop a handful of states with unusual clusters from their analysis. Others, such as Brace et al (2002),
are conducted in person and thus rely on area-based cluster sampling rather than random digit
dialing or some other completely random sampling procedure. By design, cluster sampling is
a type of non-random sampling, where different individuals in the population have different
probabilities of being chosen. If all respondents within a state do not have an equal chance of
being selected for the survey, as is the case in cluster-sampled polls, it is unclear how accurately
MRP will be able to estimate state-level effects and state-level opinion more broadly. Especially
given the importance of geographic indicators in predicting state-level opinion discussed above,
it is necessary to understand how a nonrandom sample based on geographic location will impact
simulation methods such as MRP.

This paper examines how MRP performs on cluster-sampled polls, specifically focusing
on the GSS and the ANES. After discussing why cluster sampling may impact state-level esti-
mates produced with MRP, the paper will test MRP on the GSS and ANES and propose some
corrections to perform when using MRP on cluster-sampled polls.

**Cluster Sampling**

Two of the longest running and most respected polls in political science, the American National
Election Study and the General Social Survey, both rely on cluster sampling rather than random
digit dialing or any other full probability sampling method. Given the importance of these two
polls, it is important to understand how cluster sampling works and why it is used for some
surveys.

Cluster sampling has been used by major survey houses for several decades. When based
on geographic location, this practice may also be known as area sampling. The GSS has used
cluster sampling since it fielded its first large poll in 1972, and the ANES has done so since its
early polls in the 1950s. At the first level of sampling, large, pre-defined regions are sampled from
the country as a whole. Then, in successive rounds, increasingly smaller geographic areas are
chosen from within those clusters, eventually narrowing down to the household level and then
have used disaggregation on the GSS to generate state-level estimates.
Cluster sampling can include certain aspects of random sampling within the chosen clusters, but it is different from random sampling because not every respondent in the population of interest has an equal chance of being chosen. Rather, those in the selected clusters have a higher probability of being selected and those outside of those clusters have zero probability of being included. While the GSS and ANES employ slightly different sampling techniques and have also both changed their techniques as new technologies and population listings have become available, the general principle remains the same. A national frame of clusters is designed, typically guaranteeing extra or automatic weight to the country’s largest population centers. Increasingly smaller clusters nested within these larger clusters are chosen until only certain neighborhoods are left, and it is from these neighborhoods, often as small as 300 households, that respondents will be chosen.

Cluster sampling has many advantages, the most important of which are that it decreases both the cost and the time of conducting in-person interviews by selecting respondents clustered within specific, smaller geographic areas. Despite the lack of purely random sampling, cluster sampling has long been considered a valid and respected polling technique, and cluster-sampled polls have even been used to develop public opinion estimates for sub-national units. Brace, Sims-Butler, Arceneaux, and Johnson (2002) pool and then disaggregate the GSS over a 25-year period from 1974 to 1998, showing that their estimates hold up to a variety of stability and reliability tests. While their work is promising, they do not directly test their estimates against similar measures that are already viewed as valid, such as state-level polls or national polls designed to be representative at the state level, leaving the accuracy of measures derived through disaggregation of cluster-sampled polls an open question. Furthermore, Brace and his co-authors pool 25 years of the GSS to come to this conclusion, creating such a large sample that it is easy to see why they could defend disaggregation of the GSS, despite problems one might expect because of cluster sampling. As they note, the GSS has changed its sampling frame periodically, and pooling over 25 years would have increased the number of clusters dramatically compared to using a single GSS poll. Given that this pooling over 25 years is not a practical research design for scholars interested in changes over time or questions that were not asked consistently over
long periods of time, it is important to investigate the accuracy of subnational estimates derived from a single cluster-sampled poll.

**MRP and Geography**

To understand how MRP may behave on a cluster-sampled poll, it is important to understand exactly how MRP simulates subnational public opinion and how a nonrandom geographic sample may impact this simulation. Following the model tested by Lax and Phillips (2009b) to estimate public opinion at the state level, I begin by modeling individual responses to public opinion questions as a function of each individual’s demographic information and state of residence. Specifically, the model is written out below for each individual $i$, with index markers $j$ for race/gender combination, $k$ for age, $l$ for education level, and $s$ for state.

$$
Pr(y_i = 1) = \logit^{-1}(\beta^0 + \alpha_{race,gender}^{j[i]} + \alpha_{age}^{k[i]} + \alpha_{edu}^{l[i]} + \alpha_{state}^{s[i]})
$$

Each of the terms after the intercept is modeled itself based on a normal distribution with a mean of zero and an estimated variance. Race/gender is either a four or six category variable, depending on the model specification and the individual data. Age and education are both four-category variables. State is modeled based on the region the state is in and some state-level variance. In some models, as will be discussed further in the results section, I model the state effect based on both region and an aggregate state level measure, such as state presidential vote for the Democratic candidate, and show that this improves the performance of the model. Finally, region is modeled as a random effect as well.

$$
\alpha_{race,gender}^j \sim N(0, \sigma_{race,gender}^2), \text{for } j = 1, \ldots, 6
$$

$$
\alpha_{age}^k \sim N(0, \sigma_{age}^2), \text{for } k = 1, \ldots, 4
$$

$$
\alpha_{edu}^l \sim N(0, \sigma_{edu}^2), \text{for } l = 1, \ldots, 4
$$
\[ a_{s}^{state} \sim N(a_{m[s]}^{region} + \beta^{presvote} \cdot presvote, \sigma_{state}^{2}), \text{for } s = 1, \ldots, 50 \] (5)

\[ a_{m}^{region} \sim N(0, \sigma_{region}^{2}), \text{for } m = 1, \ldots, 4 \] (6)

These demographic factors do correlate strongly with ideology, partisanship, and presidential vote, as well as with other attitudes that I estimate in this paper. While this model is quite simple, such a standard model has been shown to perform quite well (Park, Gelman, and Bafumi 2006, Lax and Phillips 2009b, 2012). Thus, it is appropriate to use this model to assess how well MRP performs on cluster-sampled polls.

Once opinion is estimated, the next step is to post-stratify the estimated opinion by state-level population. Specifically, the above model will have generated 4,800 combinations of different demographic and state values. I use the “1-Percent Public Use Microdata Sample” from the US Census to learn the number of people in each state, \( N_{c} \), in each demographic cell type \( c \). From here, I can generate state-level public opinion estimates by weighting the opinion prediction in each cell, \( \theta_{c} \), according to the state’s population:

\[ y_{MRP}^{state[s]} = \frac{\sum_{c \in s} N_{c} \theta_{c}}{\sum_{c \in s} N_{c}} \] (7)

This estimation technique thus relies on geographic information both in the simulation phase and in the poststratification phase. The question I consider, then, is how this geographic information, particularly as used in the simulation phase, may impact the validity of state-level estimates of public opinion produced through MRP on cluster-sampled polls.

**Potential Problems with Cluster Sampling and MRP**

For MRP to produce valid estimates, it presumes that survey respondents are fairly representative of their geographic-demographic type, an assumption that is reasonable when respondents are selected randomly and a poll has an appropriately large sample size. Cluster sampling, however,

---

2This presumes the model with 6 race/gender categories, 4 age and education categories each, and 50 state categories. The model that excludes Hispanic as a racial category has 3,200 demographic-geographic types.
alters how sampling occurs within states in a non-random way. If geography matters not only because of interstate variation but also because of intrastate variation, then the accuracy and efficiency of subnational opinion estimates, whether generated through disaggregation or MRP, could be impacted by cluster sampling. Intuitively one would expect that clusters could be unrepresentative of the states they are within, especially in states with diverse populations that are often geographically sorted by ideology and partisanship.

Erikson, Wright, and McIver (1993) note this potential problem in *Statehouse Democracy*, citing it as the reason that they chose to use the CBS/NYT poll as opposed to the ANES and GSS. Unlike disaggregation, however, one might expect MRP’s use of poststratification to correct at least partially for any demographic unrepresentativeness of a given cluster (as Lax and Phillips 2009b suggest). On the other hand, MRP might not produce accurate measures of geographic effects in the simulation phase if the sampled respondents are not representative of their state’s population, and it may even exaggerate that unrepresentativeness. The validity of cluster sampling might also vary among different cluster-sampled polls based on the number of clusters, the size of clusters, and other more idiosyncratic factors of a cluster-sampled poll.

To understand how cluster-sampled polls may be problematic for MRP, I first look at the clusters themselves, showing how clusters often seem unrepresentative of the state within which they are nested. I then measure design effects in some representative cluster-sampled polls, showing that intracluster homogeneity should be cause for further concern.

**Cluster Sampling in the Real World**

It is illustrative to look at the geographic distribution of clusters in an area-sampled poll. While specific geographic information for respondents is typically withheld from the public release of a dataset to ensure the privacy of respondents, some polls do release more information than others. The ANES, for example, releases the state and congressional district for each respondent along with an identification code for the first two levels of sampling for each respondent – the strata and cluster. While not nearly as precise as having the zip code or census tract, this allows us to understand broadly the location of ANES respondents by state and congressional district.
Looking at the 2004 sample, with its almost 1200 respondents, 139 of the country’s 435 congressional districts are represented, nested within only 29 of the 50 states, as well as an additional respondent from the District of Columbia. While MRP can handle imputation of missing states, this alone represents a tall order. Further, an inspection of the congressional districts that are represented might cause additional reason for concern. For example, in New Jersey this poll samples 3 people from the state’s first congressional district and 33 from the state’s second congressional district, with zero sampled from the state’s other 11 districts. In 2004, the first congressional district was incredibly Democratic, encompassing the city of Camden. However, the second congressional district, from which almost all of the New Jersey respondents are sampled, was a more Republican area, unusual in a typically Democratic-leaning state. Indeed if one looks at the party identification for the 36 New Jersey residents in this sample, one finds almost a perfectly even split of 18 Democrats, 2 independents, and 16 Republicans – more even then one might expect given the sampling but less Democratic than one would expect for New Jersey.

On the opposite side, one can look at Utah, a small state from which almost as many people – 32 – are sampled, with the majority coming from the more “blue” second congressional district, which includes Salt Lake City, and the rest coming from outside of it. Utah is considered one of the most Republican states in the country, with Bush receiving 71.5% of the vote in 2004. Nonetheless, this sample from Utah oversampled Democrats, including 18 Democrats, 4 independents, and 10 Republicans. An even more extreme example, perhaps, is that of Louisiana. This poll included 46 respondents from Louisiana, but they were all sampled from the conservative 4th congressional district in the northwestern part of the state.

One can show several more examples like these, and they illustrate a problem with any national poll – that one cannot disaggregate at the state level from a single poll to produce valid results. Nevertheless, these results are more worrisome for the use of MRP given that respondents are clustered within certain areas of a state. If these non-random areas are unrepresentative of the state, as they appear to be in these cases, we would expect MRP to produce inaccurate state random effects for these states. We would also expect state random effects to be more problematic in states with fewer clusters.
Design Effects and MRP

The above discussion explains why cluster-sampled polls may present a challenge for statistical techniques such as MRP. We can more precisely gauge the challenges posed by cluster-sampled polls by measuring the design effect for our variables of interest.

Specifically, one reason area-sampled polls are potentially problematic for MRP is because one would expect there to be intra-cluster homogeneity within these samples, distinguishing the people in the cluster from the other people in the state. In the case of area-based clusters specifically, people who live in the same neighborhood are more likely to be similar to each other than to those who live in different neighborhoods on a host of variables, including race, economic class, and even political beliefs. This intracluster homogeneity makes the sample less precise than a simple random sample of the same size would be, increasing the variance of the estimates and decreasing the effective sample size of the poll. The design effect measures this lack of precision and loss of sample size. The design effect can be expressed as a function of the average number of cases selected in each cluster, \( b \), and the intracluster correlation coefficient, \( \rho \), which measures how similar respondents are within clusters for a given variable of interest.

\[
deff \approx 1 + (b - 1) \times \rho
\]

We can also define the design effect as the ratio between the variance of a statistic in a clustered sample and the variance of that statistic under the assumption of a simple random sample.

\[
deff = \frac{\text{Var}_{\text{clustered}}(y)}{\text{Var}_{\text{SRS}}(y)}
\]

A simple random sample, therefore, has a design effect of 1 by definition. A design effect greater than 1 reflects that there is intracluster homogeneity within the sample. For example, if a variable has a design effect of 2, that means that the sample variance will be twice as big as it would be under a simple random sample. It can also be interpreted to show that half as many respondents would have been needed to produce the same results under a simple random sample. Design effects are not a constant in a survey but rather are measured separately for each variable. For
example, in a geography-based cluster sample, we might expect a great deal of intra-cluster homogeneity for a variable such as income or race, given current housing patterns in the United States. However, we might not expect a great deal of intracluster homogeneity in terms of gender or disability.

In their work on large-scale multistage area probability designs, Harter et al (2010) examine the design effects of several key variables in the 2006 GSS. As one would expect, they found design effects to be of more concern for some variables than for others. The only variable they record as having a design effect of close to 1 is whether or not one has a happy marriage, with a design effect of 0.88. The other variables they report have design effects from 1.51 to 2.85. Given that the concern of this paper focuses on the use of MRP to measure political attitudes measured in cluster-sampled polls, I will focus solely on design effects for these variables. Harter et al report several design effects of interest:

<table>
<thead>
<tr>
<th>GSS Variable</th>
<th>deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>1.7</td>
</tr>
<tr>
<td>Lower class or working class</td>
<td>1.9</td>
</tr>
<tr>
<td>Abortion should be permitted for any reason</td>
<td>1.5</td>
</tr>
<tr>
<td>Support death penalty for murderers</td>
<td>2.3</td>
</tr>
<tr>
<td>High confidence in the executive branch of the federal gov’t</td>
<td>1.7</td>
</tr>
<tr>
<td>Marijuana should be legal</td>
<td>1.7</td>
</tr>
<tr>
<td>Men are more suited to politics then women</td>
<td>1.7</td>
</tr>
<tr>
<td>Spending too little on social security</td>
<td>1.7</td>
</tr>
</tbody>
</table>

This table lists the design effects for several key variables in the 2006 GSS.

These design effects show that intracluster homogeneity matters across several different political attitudes, as well as demographics. While these questions are primarily about social attitudes, the design effect for spending on social security (a relatively popular program) shows that one must be aware of design effects for economic and spending issues as well. As Harter and her co-authors conclude, this lack in precision and resulting underestimation of the variance is severe enough that researchers should use alternate procedures such as either a linearization method or a replication method to estimate accurate variances and avoid creating a false sense of confidence.
in the data. For researchers to be able to do these calculations and better understand the design effects themselves, Harter notes that survey houses must release replication weights as well as the cluster and strata indicators. 3

Likewise, the ANES must also be examined for design effects. In his technical report on how to analyze ANES survey data, Matthew DeBell (2010) advises using a Taylor series approximation, what Harter et al (2010) call linearization, to take the design effect of the cluster sample into account and produce accurate standard errors. While the design effect is different for each variable, DeBell provides the average design effect for recent ANES surveys, which range from 1.21 in the 2004 pre-election study to 1.82 in the 2006 pilot study. Unlike the GSS, the ANES includes the three necessary variables for this procedure in its public release of the data: the weight, the stratum, and the cluster. Below, I use the Taylor series approximation method to calculate design effects for variables of interest in the 2004 pre-election study.

<table>
<thead>
<tr>
<th>NES Variable</th>
<th>deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1.5</td>
</tr>
<tr>
<td>Social Class (self-identified)</td>
<td>1.6</td>
</tr>
<tr>
<td>Race (6 category)</td>
<td>2.1</td>
</tr>
<tr>
<td>Party ID (7 point scale)</td>
<td>2.6</td>
</tr>
<tr>
<td>Ideology (7 point scale)</td>
<td>2.1</td>
</tr>
<tr>
<td>Abortion self-placement</td>
<td>1.3</td>
</tr>
<tr>
<td>Same-sex Couples Allowed to Marry</td>
<td>1.6</td>
</tr>
<tr>
<td>Government Spending on Welfare</td>
<td>1.5</td>
</tr>
<tr>
<td>Economy Better Since GW Bush Took Office</td>
<td>1.4</td>
</tr>
</tbody>
</table>

This table lists the design effects for several key variables in the 2004 ANES.

These results add to our concerns about intracluster homogeneity and how it may impact the ability of MRP to produce accurate estimates of public opinion data gathered through multistage area designs. If the presence of a design effect greater than one reflects a decrease in the effective sample size of the model, it might be the case that larger sample sizes are needed

---

3While the GSS historically released survey weights and cluster indicators, it only issued the strata indicators recently, with the release of the 1972-2012 cumulative datafile. However, challenges with the data structure currently prohibit the calculation of design effects for several variables of interest (Pedlow 2013).
to conduct MRP on a cluster-sampled poll than one would need to conduct it on a randomly- 
sampled poll. In addition, while knowledge of the design effect is typically used to reflect the 
lack of precision in a survey and increase the variance of the estimates, the presence of a design 
effect much greater than 1 might impact the accuracy of the coefficient estimates in addition to 
the variances when using MRP. If the respondents from a given state are clustered in one loca-
tion and there is a high amount of intracluster homogeneity relative to heterogeneity at the state 
level, an MRP model may well produce inaccurate state random effects; the fewer clusters a poll 
contains in a given state, the more problematic this might be. One might even expect inaccurate 
demographic coefficients in the model, if the respondents are substantially unrepresentative of 
their demographic groups as well as their state population. Given these concerns, it is important 
to investigate how MRP functions when applied to cluster-sampled polls, and how one might 
 improve MRP estimates based on these polls.

Given the presence of unusual clusters and large design effects, it is important to examine 
whether a cluster-sampled poll may introduce error or inaccuracy into MRP estimates. If cluster 
sampling results in a population that is not representative of the state, this could impact the 
value of the state random effects, introducing error into the simulation phase of the multilevel 
model. The state random effects can be thought of as the state-level errors that dictate how much 
the model’s intercept should be shifted up or down for each state (Gelman and Hill, 260). In 
addition to impacting the state random effects, the use of area-based cluster sampling may also 
impact the coefficients of the variables in our model. For example, if a demographic subgroup 
in the sample is not representative of that subgroup in the overall population, it might lead 
to an inaccurate coefficient on a variable included in the model, such as race, gender, age, or 
education. While theoretically this seems less likely than issues with the geographic effects, it 
is still something to be aware of as we consider sources of error from an area sample. In the 
next section, I compare several MRP estimates produced from cluster-sampled polls with “true 
values” produced through disaggregation of polls that do not rely on cluster sampling, and 
I examine potential solutions for researchers who would like to use MRP on cluster-sampled 
polls.
Results and Discussion

There are several ways to test the validity of estimates produced by MRP when using cluster-sampled polls. I begin by looking at several sets of state-level estimates generated by MRP from individual GSS polls and pooled sets of GSS polls from several years, considering the face validity of each. I then compare GSS and ANES state-level estimates with state-level results from other polls that are representative at the state level, either by design or after several polls are pooled together. Following this, I use a feature of the 1980 sampling frame shared by the GSS and ANES to examine how the number of clusters impacts performance of MRP. Last, for the ANES opinion estimates, I then investigate how knowledge of a sub-state geographic indicator, in this case congressional district, affects state-level estimates.

Face Validity: Who’s Liberal and Who’s Conservative?

When evaluating the results from multilevel regression with poststratification, it is helpful to look at the face validity of these results. Are liberal states liberal and conservative states conservative in the measures created by MRP? In other words, do the state-level opinion estimates make sense? The short answer, based on an initial analysis of MRP on GSS data, is sometimes.

I use the GSS from recent years to generate state-level estimates for party identification and political ideology. I first create four variables from these: Democratic, Republican, liberal, and conservative, and then I use logistic regression to create estimates for each of these at the state level.\(^4\) I use MRP to model public opinion responses as a function of both geographic factors (state and region) and demographic factors (age, education, race, and gender), allowing for partial pooling across states. Then, I use Census data for poststratification, weighting the estimates for each type of respondent by the percentage of that type in the actual state population.\(^5\)

\(^4\)Both party identification and political ideology are measured along a seven point scale in the GSS. In recoding, I count those who are “extremely liberal,” “liberal,” and “slightly liberal,” as liberal, and likewise for conservative; similarly, I count those who are a “strong Democrat,” a “not very strong Democrat,” and “Independent, close to Democrat” as Democratic and likewise for Republican. Furthermore, since I focus on estimating percent Democratic and Republican of total respondents, I do include those who responded “other party” in the base category, but I drop the (few) respondents who said “don’t know,” and “no answer.” Likewise, for the ideology variables, I drop those who said “don’t know” or “no answer,” as well as those who were not asked the question.

\(^5\)Note that in this and subsequent models in this section, race is a dichotomous white/black variable, given con-
The figures below graph percent Democratic and percent liberal as generated by MRP in four different models. First, I use a simple MRP model on the 2004 GSS (N = 1305 for ideology, N = 2791 for partisanship). Second, I add an aggregate measure of Democratic presidential vote share in 2004 as a state-level predictor. This statistic highly correlates with ideology and especially with partisan identification at the state level, and, as shown by Lax and Phillips (2009b), adding such an aggregate measure to a logistic estimation should increase its accuracy. Third and fourth, I repeat these graphs on a larger dataset, pooling the 2002, 2004, and 2006 GSS (N = 6938 for ideology, N = 9966 for partisanship). While Lax and Phillips posit that a single national poll of approximately 1400 respondents should produce valid estimates, I hypothesize that a larger N will be required for a cluster-sampled poll. Given that, as shown earlier, cluster sampling results in large design effects for key political variables, I would expect a corresponding reduction in the poll’s effective sample size. Furthermore, the sampling frame used by the GSS changed between the 2002 and 2004 waves of the survey. Thus, pooling GSS surveys from 2002, 2004, and 2006 creates a survey that, in effect, samples from twice the number of clusters as a single GSS; this should produce more valid estimates, since sampling from additional clusters improves the efficiency and accuracy of a cluster-sampled poll (Harter et al 2010).

These figures help assess the face validity of the estimates in two different manners. First, by looking at the values on the x- and y-axes, one can see how accurate each state-level estimate appears to be. For example, a conservative state like Texas should have a high conservative estimate and a low liberal estimate. Second, by looking at both partisanship and ideology simultaneously, one sees how strong the overall relationship is between the estimated measures. While this would not have been a useful example several decades ago, increased polarization had led to a strong correlation between state-level ideology and partisanship, especially after 2000. In an updated analysis of their work in Statehouse Democracy, Erikson, Wright, and McIver (2006) note that by Bush’s first term, this correlation had reached $r = 0.66$.

Indeed, looking at the pattern of these four graphs in Figure 2, MRP seems to perform

---

6These models and figures also exclude DC, which is an outlier on these measures.
as hypothesized, producing problems that would not occur if MRP were performed on a poll that did not use area-based clusters. In the first regression, when MRP is done on a single cluster-sampled poll without an aggregate state-level measure, the results do not fully meet face validity. Some results make sense: New York is one of the most liberal and Texas is one of the most conservative, for example. However, the correlation between party and ideology is basically zero, and several individual states present surprising results. One would not expect Indiana to be the most liberal state or Idaho and Kansas to be near the top as well. Likewise Massachusetts is in the middle for both partisanship and ideology, and Washington state is more Republican and conservative than one would expect. Adding state-level vote for Kerry in 2004 to the 2004 estimation changes a few results, but keeps the pattern roughly similar, showing that using a more complicated model on a single cluster-sampled poll is not a sufficient solution. Pooling 2004 with the surveys from 2002 and 2006, more than tripling the N and introducing additional clusters, has a more substantive effect. Massachusetts becomes one of the most liberal states, and Washington improves as well. States that were in a more accurate place, such as Texas and New York, stay where they are relative to the other states. Overall, the correlation between ideology and partisanship strengthens, and this relationship increases further when I add Kerry vote to the model. The pooled model with an aggregate predictor produces a correlation of 0.67 between ideology and partisanship, similar to that found by Erikson, Wright, and McIver.

The Republican and conservative graphs in Figure 3 show a similar pattern. In the 2004 data, however, the South has a distinct pattern, with low Republican identification given its conservatism. While this relationship is common in historical data, it is odd it is so strong in 2004. The pooled data, however, corrects for this, and, especially in the model that includes the aggregate measure, the states are where one would expect them to be and the correlation between partisanship and ideology reaches 0.70.

These initial face validity tests seem to demonstrate that a simple MRP analysis on a small dataset will not produce sufficiently reliable results. Rather, they imply that one must simultaneously add a strong aggregate state-level measure and use a larger dataset to increase
This figure shows several sets of MRP estimates from the GSS for liberal and Democratic. The first graph uses a simple MRP model on the 2004 GSS. The second (top right) adds presidential vote as an aggregate predictor. The bottom two graphs repeat these two models on a pooled 2002, 2004, and 2006 dataset. To reach the expected correlation between ideology and partisanship and have face validity for individual states, one must both pool clustered data and use a more sophisticated MRP model.
Figure 2: MRP GSS Estimates for Republican and Conservative – 2004 v. 2002-2006 Estimates.

This figure shows several sets of MRP estimates from the GSS for conservative and Republican. As in Figure 1, the first graph uses a simple MRP model on the 2004 GSS. The second (top right) adds presidential vote as an aggregate predictor. The bottom two graphs repeat these two models on a pooled 2002, 2004, and 2006 dataset. To reach the expected correlation between ideology and partisanship and have face validity for individual states, one must both pool clustered data and use a more sophisticated MRP model.
the validity of MRP estimates from cluster-sampled polls. These results also seem to indicate that pooling surveys across sampling frames, thus increasing the number of clusters within the sample, further improves the results. In the next section, I evaluate how MRP can be used on cluster-sampled polls by comparing these estimates produced under different specifications to results from a simple random sample, the National Annenberg Election Study.

GSS, ANES, and the National Annenberg Election Study

The National Annenberg Election Study (NAES) in a national survey that uses random digit dialing to sample respondents in 48 states and the District of Columbia. The Annenberg survey has a sample size of over 80,000 people and is designed to be representative at the state level as well as the national level. Thus, since the NAES is a full probability sample that can be disaggregated into its state components, I can consider its results to be a valid standard against which I can evaluate MRP estimates on cluster-sampled polls. Disaggregation of significantly large datasets is considered to be an ideal method of measuring state-level opinion, and thus the Annenberg dataset provides a useful standard against which to compare MRP estimates. Here, I compare the state-level NAES estimates on partisanship and political ideology to the state estimates produced by MRP of similar questions in the GSS and ANES. While there are minor differences in question wording, these questions are similar enough to assess MRP’s performance on the GSS and ANES.

First, I run a series of correlations between the MRP estimates from the GSS data and the disaggregated values from the NAES data, looking at four state-level estimates (percent Democratic, percent Republican, percent liberal, percent conservative) and testing all four specifications for MRP estimation used earlier in this paper against the NAES value. This way, one can see if and how correlation changes when one moves from using the single dataset to the pooled dataset and from the simple MRP model to the MRP model using a predictive state-level measure, in this case statewide presidential vote for John Kerry in 2004. This yields the following results:
As the table shows, the simple MRP model using a single poll produces far from desirable results. While the correlations are all correctly signed, they are quite small except for the conservative measure, and they do not inspire much trust in the MRP estimates. Lax and Phillips (2009b) test how a MRP model without an aggregate predictor performs on a sample size of 2800 and find correlations with true values of 0.77. The GSS for 2004 is just under this \( N \) for the two partisan measures (2787), and half of this for the two ideology measures (1303), and none of these correlations come close to the correlations Lax and Phillips demonstrate from non-clustered polls. Since I follow Lax and Phillips in their model specifications, this suggests an issue with the underlying data. Specifically, these results suggest that cluster sampling introduces error into MRP estimates because the respondents are non-randomly sampled at the state level.

Including a state-level predictor highly correlated with what one is trying to predict, such as Democratic presidential vote share, somewhat improves the accuracy of the model on the 2004 data, producing an average correlation of 0.55.\(^8\) This improvement reinforces Lax and Phillips’ general recommendation to include an aggregate predictor when using MRP, especially when using a single national poll. A state level variable is especially important for states without any data, and there are more such states in a cluster-sampled poll than in a randomly-sampled poll. However, while including the aggregate predictor improves the correlation with true values, it does not improve it as much as one would expect given non-clustered data. Lax and Phillips

---

\(^7\)Therefore, for this part of the analysis, I will drop Alaska and Hawaii from the analysis. I also drop DC, as I did above, given that it is an outlier on these measures.

\(^8\)The liberal measure, however, presents an unusual outlier, perhaps because of the relatively small number of people who identify as liberal in the survey. Not including the liberal outlier, the average correlation would be 0.68.
(2009b) find that a standard MRP model applied to a random poll of 1400 typically has a correlation of approximately 0.73 with true values and a correlation of 0.82 for a poll of 2800.

In addition to adding an aggregate predictor, another potential solution to improve MRP’s performance is pooling multiple surveys to have a larger $N$. In this example, I pool the 2002, 2004, and 2006 GSS surveys ($N = 9942$ for partisanship identification and $N = 6917$ for ideological identification), and the results show an increase in correlations both in the simple model and the model with the aggregate predictor, the latter of which reaches correlations similar to those expected by Lax and Phillips mentioned above. Pooling in this case might increase correlations for two reasons. First, an increased $N$ might be necessary to improve MRP estimates under cluster sampling, given that cluster sampling may reduce the effective sample size of a survey. However, this answer is not completely satisfying. If cluster sampling reduces the efficiency of the sample, including more respondents from within the same cluster alone should not improve it substantially (Harter et al 2010). Erikson, Wright, and McIver (1993) touched on this concern, as quoted above, when explaining why they use CBS/NYT surveys, rather than GSS and ANES polls, to construct their state-level ideology and partisanship measures. However, the second reason that pooling might improve estimates is that, depending on the sampling design, pooling surveys from several years might increase the number of clusters from which respondents were drawn if the sampling frame has changed over the years pooled. That is the case when pooling GSS surveys from 2002, 2004, and 2006, as the frame changed in 2004. Harter et al (2010) advise that sampling from additional clusters will improve the quality of one’s sample, and that indeed seems to be the case here.

These results indicate that simultaneously pooling data and using an aggregate state-level predictor can produce the most accurate measure and make it much less problematic to use MRP on cluster-sampled data. To confirm, I replicate the above section using the American National Election Studies. To analyze the ANES, I use three different forms of the data. First, as with the GSS, I look at only 2004, the same year as the Annenberg survey. The ANES has smaller $N$’s than the GSS ($N = 1175$ for partisanship identification and $N = 912$ for ideological identification), and I therefore do not expect a strong performance for MRP. Next, I pool the 2004 data with
surveys from 2002 and 2000, which I refer to as Pooled Model A, increasing the sample size to an $N$ of 4313 for partisanship data and 2755 for ideology.\(^9\) Finally, given that these $N$'s are still smaller than what might be needed with a cluster-sampled poll, I also add 2008 data to the pooled model (labeled Pooled Model B), yielding an $N$ of 6516 for the partisanship variables and an $N$ of 4329 for the ideology variables. I generate MRP estimates both using the basic model and adding state-level Kerry vote as an aggregate predictor.\(^{10}\) Comparing the MRP-generated state-level estimates with Anneberg’s state-level estimates, I find the following correlations:

<table>
<thead>
<tr>
<th>Proportion</th>
<th>2004</th>
<th>2004 w/ KV</th>
<th>Pooled A</th>
<th>Pooled A w/KV</th>
<th>Pooled B</th>
<th>Pooled B w/KV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>0.24</td>
<td>0.66</td>
<td>0.38</td>
<td>0.71</td>
<td>0.42</td>
<td>0.68</td>
</tr>
<tr>
<td>Republican</td>
<td>0.19</td>
<td>0.61</td>
<td>0.36</td>
<td>0.65</td>
<td>0.41</td>
<td>0.68</td>
</tr>
<tr>
<td>Liberal</td>
<td>0.58</td>
<td>0.71</td>
<td>0.54</td>
<td>0.70</td>
<td>0.57</td>
<td>0.71</td>
</tr>
<tr>
<td>Conservative</td>
<td>0.53</td>
<td>0.67</td>
<td>0.40</td>
<td>0.65</td>
<td>0.54</td>
<td>0.74</td>
</tr>
<tr>
<td>Average</td>
<td>0.39</td>
<td>0.66</td>
<td>0.42</td>
<td>0.68</td>
<td>0.49</td>
<td>0.70</td>
</tr>
</tbody>
</table>

$N$ (2004) = 912 (ideology), 1175 (partisanship)

$N$ (Pooled A) = 2755 (ideology), 4313 (partisanship)

$N$ (Pooled B) = 4329 (ideology), 6516 (partisanship)

Bold font indicates an improvement over the single-year, basic model.

As with the GSS data, I find that using a strong aggregate predictor at the state-level improves the model considerably, producing stronger correlations with true values than the model that does not include the predictor. Pooling data, however, does not lead to as strong of an improvement as it did with the GSS data. Since both the GSS and the ANES changed sampling frames over this time period, it is unlikely that the number of clusters explains the difference. The pooled datasets here are smaller than those used in the GSS example, so it may be the case that a larger sample size continues to be important for MRP conducted on cluster-sampled polls. It may also be the case that pooling introduces a different bias if the variables that are being estimated change over time. While the GSS example pools 2002, 2004, and 2006 surveys, the ANES data requires pooling from 2000 to 2008 to achieve a similar $N$. This larger time frame may create additional tradeoffs, introducing a pooling bias which could harm accuracy over time, rather than help it.

---

\(^9\)I cannot pool with 2006 data, as I did with the GSS, since the ANES did not conduct a survey in 2006.

\(^{10}\)Unlike the GSS model, the ANES model uses three categories for race: black, white, and Hispanic.
(Gelman 2013). Despite this, the large pooled model which includes the aggregate predictor does still return strong correlations for all four variables which approach the standard set by Lax and Phillips (2009b).

Another statistic that is helpful to examine is the mean absolute error. This measure takes the average absolute difference between the generated estimates, in this case those generated by MRP from the GSS and ANES, and the baseline measure, which are those measures produced by disaggregating the 2004 Annenberg data. The table below shows mean absolute errors calculated by comparing the GSS and the NAES. These mean absolute errors are, on average, slightly larger than one would expect, given that Lax and Phillips (2009b) report mean absolute errors for MRP estimates produced using a full model specification ranging from 4.9 to 4.1 for sample sizes of 1400 and 7000, respectively. Similar to the correlation results, these numbers do improve, on average, when using a state-level predictor, but they actually worsen when pooling the data. This may be another reflection of a potential bias caused by pooling data. This might be especially reflected in the increased error of the Republican measure in the pooled models, given greater instability in partisanship overtime.

<table>
<thead>
<tr>
<th>Mean Absolute Error – GSS and NAES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Abs. Err.</td>
</tr>
<tr>
<td>Democratic</td>
</tr>
<tr>
<td>Republican</td>
</tr>
<tr>
<td>Liberal</td>
</tr>
<tr>
<td>Conservative</td>
</tr>
<tr>
<td>Average</td>
</tr>
</tbody>
</table>

N (2004) = 1303 (ideology), 2787 (partisanship)
N (Pooled) = 6917 (ideology), 9942 (partisanship)
Bold font indicates an improvement over the single-year, basic model.

I then examine the mean absolute errors for the ANES data as compared to the Annenberg data. The below table shows similar results to the GSS data, and the mean absolute errors range from 6.5 to 3.7. The ANES estimates do not have the same higher errors for partisanship measures that the GSS estimates showed, but they do further enforce the limited power of pooling alone. By pooling and using a state-level predictor, however, these results do approach the same mean
absolute errors reported by Lax and Phillips (2009b).

Mean Absolute Errors – ANES and NAES

<table>
<thead>
<tr>
<th>Mean Abs. Err.</th>
<th>2004</th>
<th>2004 w/ KV</th>
<th>Pooled A</th>
<th>Pooled A w/KV</th>
<th>Pooled B</th>
<th>Pooled B w/KV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>5.9</td>
<td>4.8</td>
<td>5.1</td>
<td>3.7</td>
<td>5.1</td>
<td>4.1</td>
</tr>
<tr>
<td>Republican</td>
<td>6.2</td>
<td>5.1</td>
<td>6.5</td>
<td>5.1</td>
<td>6.5</td>
<td>5.1</td>
</tr>
<tr>
<td>Liberal</td>
<td>4.5</td>
<td>3.8</td>
<td>5.5</td>
<td>4.4</td>
<td>5.3</td>
<td>4.4</td>
</tr>
<tr>
<td>Conservative</td>
<td>5.4</td>
<td>6.4</td>
<td>5.5</td>
<td>5.4</td>
<td>5.2</td>
<td>4.8</td>
</tr>
<tr>
<td>Average</td>
<td>5.5</td>
<td>5.0</td>
<td>5.7</td>
<td>4.7</td>
<td>5.5</td>
<td>4.6</td>
</tr>
</tbody>
</table>

N (2004) = 912 (ideology), 1175 (partisanship)
N (Pooled A) = 2755 (ideology), 4313 (partisanship)
N (Pooled B) = 4329 (ideology), 6516 (partisanship)
Bold font indicates an improvement over the single-year, basic model.

GSS, ANES, and Ideology Measures

I next replicate these results using a different measure of “true” values: the state-level measures of ideology and partisanship that Erikson, Wright, and McIver created using disaggregation of several national level polls. These measures are widely used in political science, and thus it seems appropriate to measure MRP estimates against them. Given that the time frame of the GSS and ANES data used here spans from 2000 to 2008, I use an updated version of the disaggregated data that was produced using polls from 1996 to 2003 (Erikson, Wright, and McIver 2006). As before, I drop DC as an outlier from my MRP analysis. Beginning with the GSS, this produces the following correlations:

Pearson’s Product-Moment Correlations – GSS and EWM

<table>
<thead>
<tr>
<th>Proportion</th>
<th>2004 Model</th>
<th>2004 w/ Kerry Vote</th>
<th>Pooled Model</th>
<th>Pooled w/Kerry Vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>0.55</td>
<td>0.61</td>
<td>0.47</td>
<td>0.58</td>
</tr>
<tr>
<td>Republican</td>
<td>0.26</td>
<td>0.69</td>
<td>0.54</td>
<td>0.78</td>
</tr>
<tr>
<td>Liberal</td>
<td>0.22</td>
<td>0.06</td>
<td>0.57</td>
<td>0.77</td>
</tr>
<tr>
<td>Conservative</td>
<td>0.46</td>
<td>0.82</td>
<td>0.58</td>
<td>0.79</td>
</tr>
<tr>
<td>Average</td>
<td>0.37</td>
<td>0.55</td>
<td>0.54</td>
<td>0.73</td>
</tr>
</tbody>
</table>

N (2004) = 1303 (ideology), 2787 (partisanship)
N (Pooled) = 6917 (ideology), 9942 (partisanship)
Bold font indicates an improvement over the single-year, basic model.
In general, these correlation results are similar to those produced with the Annenberg data. The single-year, simple model produces poor to moderate correlations, and while the pooled model improves on them slightly, the correlations are still smaller than one would expect compared to a non-clustered poll. Adding a state-level predictor to the single year model produces strong correlations for all variables but percent liberal, which is the measure I attempt to estimate with the fewest respondents. As in the earlier tables that compared MRP estimates to the Annenberg data, the model using pooled data with the state-level predictor performs the best, reaching an average correlation of 0.73 among the four variables tested. This pattern is somewhat maintained in the mean absolute errors reported below.

<table>
<thead>
<tr>
<th></th>
<th>Mean Abs. Err.</th>
<th>2004 Model</th>
<th>2004 w/ Kerry Vote</th>
<th>Pooled Model</th>
<th>Pooled w/ Kerry Vote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>8.4</td>
<td>8.0</td>
<td>8.0</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>Republican</td>
<td>7.2</td>
<td>8.0</td>
<td>4.9</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>Liberal</td>
<td>4.5</td>
<td>5.0</td>
<td>5.5</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>Conservative</td>
<td>4.3</td>
<td>4.2</td>
<td>3.6</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>6.1</td>
<td>6.3</td>
<td>5.5</td>
<td>5.2</td>
<td></td>
</tr>
</tbody>
</table>

*Bold font indicates an improvement over the single-year, basic model.*

I repeat this analysis using data from the American National Election Studies. Again, I consider six models rather than four, given the lower $N$. Initially, there is a similar pattern as in the GSS results above, with correlation becoming stronger when an aggregate predictor, vote for Kerry, is added to the model. Even more so than with the GSS, the ANES shows that pooling alone adds little additional value; as discussed earlier, this may be because the longer time frame of the ANES data included here. These results are further reflected in the mean absolute error analysis.

However, this ANES analysis, both in the correlations and especially the mean absolute errors, amplifies a trend also seen in the GSS results above: In these validity tests using the disaggregated Erikson, Wright, and McIver data, the results for ideology are more comforting to those who wish to use MRP on cluster-sampled data than those for party identification, reflected

$N (2004) = 1303$ (ideology), 2787 (partisanship)

$N (Pooled) = 6917$ (ideology), 9942 (partisanship)
### Pearson’s Product-Moment Correlations – ANES and EWM

<table>
<thead>
<tr>
<th>Proportion</th>
<th>2004</th>
<th>2004 w/KV</th>
<th>Pooled A</th>
<th>Pooled A w/KV</th>
<th>Pooled B</th>
<th>Pooled B w/KV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>0.38</td>
<td>0.50</td>
<td>0.48</td>
<td>0.58</td>
<td>0.36</td>
<td>0.50</td>
</tr>
<tr>
<td>Republican</td>
<td>0.20</td>
<td>0.61</td>
<td>0.26</td>
<td>0.57</td>
<td>0.34</td>
<td>0.62</td>
</tr>
<tr>
<td>Liberal</td>
<td>0.61</td>
<td>0.75</td>
<td>0.57</td>
<td>0.74</td>
<td>0.59</td>
<td>0.74</td>
</tr>
<tr>
<td>Conservative</td>
<td>0.53</td>
<td>0.74</td>
<td>0.41</td>
<td>0.71</td>
<td>0.55</td>
<td>0.78</td>
</tr>
<tr>
<td>Average</td>
<td>0.43</td>
<td>0.65</td>
<td>0.43</td>
<td>0.65</td>
<td>0.46</td>
<td>0.66</td>
</tr>
</tbody>
</table>

*N (2004) = 912 (ideology), 1175 (partisanship)*

*N (Pooled A) = 2755 (ideology), 4313 (partisanship)*

*N (Pooled B) = 4329 (ideology), 6516 (partisanship)*

Bold font indicates an improvement over the single-year, basic model.

In both the correlations and the mean absolute errors, this at first seems surprising, given there was no such difference in the validity testing using the Annenberg data. One explanation may be that, as Erikson, Wright, and McIver argue, partisanship is less stable than ideology; given that the time periods of the various measures do not completely overlap, especially with the ANES data (which spans up to 2008), this might explain the lower correlations and the higher mean absolute errors in this series of validity tests. These results, therefore, may serve as another caution against pooling when the variable of interest changes over time. It is also worth remembering that in both the individual and pooled models, the MRP estimates generated from the GSS rely on a larger N than those generated from the ANES, which could further explain discrepancies in the results.

### Mean Absolute Error – ANES and EWM

<table>
<thead>
<tr>
<th>Proportion</th>
<th>2004</th>
<th>2004 w/KV</th>
<th>Pooled A</th>
<th>Pooled A w/KV</th>
<th>Pooled B</th>
<th>Pooled B w/KV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>10.7</td>
<td>9.8</td>
<td>12.2</td>
<td>11.6</td>
<td>13.2</td>
<td>12.8</td>
</tr>
<tr>
<td>Republican</td>
<td>12.5</td>
<td>13.7</td>
<td>11.0</td>
<td>11.0</td>
<td>9.0</td>
<td>9.3</td>
</tr>
<tr>
<td>Liberal</td>
<td>5.3</td>
<td>4.4</td>
<td>6.4</td>
<td>5.6</td>
<td>6.4</td>
<td>5.7</td>
</tr>
<tr>
<td>Conservative</td>
<td>8.8</td>
<td>10.3</td>
<td>8.0</td>
<td>9.2</td>
<td>7.8</td>
<td>8.6</td>
</tr>
<tr>
<td>Average</td>
<td>9.3</td>
<td>9.6</td>
<td>9.4</td>
<td>9.4</td>
<td>9.1</td>
<td>9.1</td>
</tr>
</tbody>
</table>

*N (2004) = 912 (ideology), 1175 (partisanship)*

*N (Pooled A) = 2755 (ideology), 4313 (partisanship)*

*N (Pooled B) = 4329 (ideology), 6516 (partisanship)*

Bold font indicates an improvement over the single-year, basic model.
Increasing the Number of Clusters: GSS, ANES, and the 1980 Sampling Frame

The results in the above sections show that pooling data from multiple surveys often leads to more reliable state-level estimates, especially when coupled with a state-level predictor. This increase in accuracy is the result of one of two potential factors. Pooling surveys increases the $N$, which also increases the available information on the population, thus improving the ability of MRP to estimate opinion for each different geographic-demographic subtype. However, pooling across surveys may also introduce additional clusters if the sampling frame of the survey has changed during the period of pooling, as was the case in many of the examples above. While it can be difficult to tease out the role of an increased $N$ versus the role of an increased number of clusters, data from the 1980s provides an opportunity to examine the separate role that each factor plays. National sampling frames can be quite expensive to create, and after the 1980 Census, the GSS and ANES decided to collaborate and use the same national frame (Heeringa 1986, Gibson 1995). The organizations thus agreed on a common sample of primary sampling units (PSUs) but then each did their own individual sampling within those PSUs. The advantage for this study, however, is that while the GSS employed 84 PSUs, the ANES chose only to use 61 of those 84 PSUs. This is a substantive decrease in the number of clusters used in the ANES surveys during this decade. Thus, with a constant sampling frame, we can compare the estimates that MRP produces on each survey with a “true value” to see if increasing the number of clusters sampled increases the validity of the survey. Given that the 1980 sampling frame overlaps nicely with the original data collected by Erikson, Wright, and McIver, I will use EWM ideology measures of liberal and conservative as a true value for comparison. Note that as above, I drop Hawaii and Alaska from the analysis as the EWM scores do not contain measures for these states. I also drop the District of Columbia, which has consistently performed as an outlier, and Nevada, which Erikson, Wright, and McIver drop from their analyses as well given its unusually high liberal score.

In the following two figures, I summarize the results from a series of models that use MRP to estimate both liberal and conservative ideology from both the ANES and the GSS. I include models that use and do not use Democratic presidential vote share, and for both polls, I
apply the models to 5 datasets - 1984, 1986, 1988, a pooled 1984 and 1986 dataset, and a pooled 1984, 1986, and 1988 dataset. Figure 4 compares the correlation of each set of MRP estimates to the “true value” as defined by Erikson, Wright, and McIver’s ideology measures. While these results are mixed, on average one can see that the GSS makes up a higher share of the models with high correlations, whereas the ANES dominates the lower correlations. Models that are conducted on pooled years typically outperform those that are conducted on a single year; even more consistently, presidential vote share almost always increases the correlation with the true value. Figure 5 uses the same MRP estimates and EWM ideology measures to look at mean absolute error. Here one sees again that the GSS does, on average, better than the ANES, and similar patterns as above hold, though presidential vote share is not as consistently a strong performer as it was in predicting higher correlations.

These results confirm the theory that increased clusters do improve the performance of MRP on cluster-sampled polls. However, they also show that the number of clusters is not the only factor, as increasing the $N$ of the poll and using a state-level predictor also shape the accuracy of the model.

---

11While I do not constrict the datasets to have identical $N$’s, the size of the polls are quite similar. For the GSS, 1984 has an $N = 1405$, 1986 has an $N = 1394$, and 1988 has $N = 1411$. For the ANES 1984 has an $N = 1477$, 1986 has an $N = 1563$, and 1988 has an $N = 1351$. 

28
This figure shows the correlation between the original Erikson et al ideology measures and several MRP estimates for these measures. The correlation is plotted along the x-axis, and each combination of data and model choice along the y-axis. The labels on the y-axis indicate the poll (GSS or NES), what is being estimated (“Lib” for liberal and “Con” for conservative), the year(s) of the poll(s) included, and whether a state-level predictor is used (D for Dukakis vote-share, M for Mondale vote-share). While the GSS does not always outperform the ANES, the 8 worst performing models are from the NES and 7 of the top 10 are from the GSS, indicating that the increased number of clusters in the GSS gives it an advantage over the ANES when performing MRP. This comparison also shows that using an aggregate predictor almost always improves upon a basic model, and pooling can be quite helpful as well.
This figure shows the mean absolute error between the original Erikson et al ideology measures and several MRP estimates for these measures. The mean absolute error is plotted along the x-axis, and each combination of data and model choice along the y-axis. As in Figure 3, the labels on the y-axis indicate the poll (GSS or NES), what is being estimated (“Lib” for liberal and “Con” for conservative), the year(s) of the poll(s) included, and whether a state-level predictor is used (D for Dukakis vote-share, M for Mondale vote-share). Again, the GSS does better overall than the ANES: 8 of the 10 worst performing models use the NES and 9 of the top 10 use the GSS, indicating that the increased number of clusters in the GSS gives it an advantage. Including an aggregate predictor and pooling data both seem to be helpful, though neither as consistently as they were in the correlation illustration.
Modeling Additional Geographic Information

While increasing the number of clusters, increasing the $N$, and utilizing an aggregate predictor are all potential ways to improve the accuracy of MRP on cluster-sampled models, another potential tool that researchers could use is to include geographic information about the cluster in the multilevel model itself. If one could account for the distinctiveness of the cluster, it might help to mitigate the impact of an idiosyncratic cluster on state-level estimates. Unfortunately (and unsurprisingly), survey firms do not release specific information about the location of their clusters without intensive IRB scrutiny, and thus the average researcher cannot model clusters directly as a level in the multilevel model. While the public data does detail which respondents are in the same PSU, for the purposes of calculating design effects or including sampling weights, it does not say where these PSUs are. Thus, it is not information that could be mapped onto the Census data necessary for MRP.

However, the ANES does provide an opportunity to test this question in its public release data by using another variable as a proxy for cluster. Specifically, the ANES releases the congressional district as well as the state for every respondent in its public data files. Previous work has shown that MRP can be used to produce accurate and efficient estimates of public opinion at the congressional district level (Krimmel, Lax, and Phillips 2013; Warshaw and Rodden 2012). For this example, however, I will include congressional district in the model but continue to post-stratify at the state level. Specifically, the only changes in our earlier model can be expressed as follows:

$$\alpha_{sd} \sim N(\alpha_{state}^{\text{cd}}, \sigma^2_{cd}), \text{ for } s = 1, \ldots, 435$$

$$\alpha_{state} \sim N(0,\sigma^2_{state}), \text{ for } m = 1, \ldots, 50$$

Here, rather than modeling state as a function of region, I instead model congressional district as a function of state and presidential vote at the congressional district level. Note that as I did earlier, I drop DC from this analysis.

I rerun some of my earlier models performed on the ANES, examining how my earlier
state-level models compare with models that include congressional district. Below, I look at 2004 and 2008, focusing on models that include presidential vote share as an aggregate predictor. For the partisanship estimates, 2004 has an \(N = 1175\), and 2008 has an \(N = 2182\); for the ideology estimates, 2004 has an \(N = 912\), and 2008 has an \(N = 1558\). Since the 2008 survey included more respondents than the 2004 survey, I would also hypothesize that the 2008 results should be slightly stronger on average. Note that because of changes in redistricting, I do not pool data in these instances. As earlier, I treat disaggregated data from the Annenberg survey as "true" values and look at both the correlation and the mean absolute error.\(^{12}\)

### Comparing State and Congressional Models – Correlation

<table>
<thead>
<tr>
<th>Proportion</th>
<th>2004 State Model</th>
<th>2004 CD Model</th>
<th>2008 State Model</th>
<th>2008 CD Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>0.66</td>
<td>0.71</td>
<td>0.58</td>
<td>0.69</td>
</tr>
<tr>
<td>Republican</td>
<td>0.61</td>
<td>0.69</td>
<td>0.64</td>
<td>0.71</td>
</tr>
<tr>
<td>Liberal</td>
<td>0.71</td>
<td>0.64</td>
<td>0.77</td>
<td>0.83</td>
</tr>
<tr>
<td>Conservative</td>
<td>0.67</td>
<td>0.69</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>Average</td>
<td>0.66</td>
<td>0.68</td>
<td>0.69</td>
<td>0.78</td>
</tr>
</tbody>
</table>

\(N (2004) = 912\) (ideology), 1175 (partisanship)
\(N (2008) = 1558\) (ideology), 2182 (partisanship)
Bold font indicates an improvement over the state model.

### Comparing State and Congressional Models – Mean Absolute Error

<table>
<thead>
<tr>
<th>Mean Abs. Err.</th>
<th>2004 State Model</th>
<th>2004 CD Model</th>
<th>2008 State Model</th>
<th>2008 CD Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democratic</td>
<td>4.8</td>
<td>4.4</td>
<td>4.7</td>
<td>4.8</td>
</tr>
<tr>
<td>Republican</td>
<td>5.1</td>
<td>4.5</td>
<td>6.2</td>
<td>6.3</td>
</tr>
<tr>
<td>Liberal</td>
<td>3.8</td>
<td>3.5</td>
<td>3.5</td>
<td>3.2</td>
</tr>
<tr>
<td>Conservative</td>
<td>6.4</td>
<td>6.0</td>
<td>3.8</td>
<td>2.7</td>
</tr>
<tr>
<td>Average</td>
<td>5.0</td>
<td>4.6</td>
<td>4.6</td>
<td>4.3</td>
</tr>
</tbody>
</table>

\(N (2004) = 912\) (ideology), 1175 (partisanship)
\(N (2008) = 1558\) (ideology), 2182 (partisanship)
Bold font indicates an improvement over the state model.

In general, this data shows a broad pattern of the congressional district model performing slightly

\(^{12}\)Since I am using the NAES data, Alaska and Hawaii are dropped from the analysis of correlations and mean absolute errors as before. While the 2008 NAES survey was smaller than the 2004 NAES survey, with an \(N\) just under 58,000, this is still large enough to be disaggregated to the state level. Like the 2004 Annenberg data, the 2008 data also relied on random digit dialing rather than cluster sampling.
better than the model which includes state and region only. This is especially true when looking at correlations, where the congressional district model performs better than the state model in all but one instance. The mean absolute error analysis, reveals less strong results, though the overall pattern indicates modeling the congressional district decreases the mean absolute error. These results also for the most part confirm the hypothesis that the 2008 estimates should be more accurate than the 2004 estimates, given that they are generated from a survey with more respondents. It may be that unusual circumstances surrounding Obama’s candidacy decreased the ability of aggregate presidential vote share to predict partisan self-identification.

This data, however, only looks at a small slice of the ANES data available, and a more thorough examination will be necessary to understand to what extent including congressional district in the multilevel model corrects for cluster sampling. While this does seem to produce a potential alternative fix, constant changes in redistricting limit the ability to use pooling if one is modeling the congressional district. Thus, researchers may face a trade-off in choosing which solution, pooling or modeling the congressional district, to use.

Conclusion

Cluster-sampled polls present specific challenges for analytical tools such as MRP. While cluster-sampled polls are considered representative at the national level, their use of clusters can introduce problems at the state and local level that can complicate simulation-based modeling such as MRP. Given the mixed evidence found in this paper, authors should use caution when applying subnational opinion estimation techniques to clustered data.

This is not to say, however, that MRP should never be used with cluster-sampled polls. The evidence above shows that the combination of pooling several polls across years and sampling frames, as well as using a state-level aggregate predictor in the MRP model, yields improved state-level MRP estimates on cluster-sampled data. However, pooling may be an unhelpful solution to some researchers if their questions of interest were not asked on several polls or if one expects responses to change over time. In some ways, even researchers who might
consider pooling may find these solutions unpalatable, as one of the main advantages of MRP is that it made pooling across several polls unnecessary. For these researchers, modeling the congressional district as an additional level in the multilevel model may be another solution that performs comparably to pooling multiple surveys. Many researchers rely on the rich datasets of the GSS and ANES, and even modified use of MRP could help political scientists to consider several questions concerning state level public opinion.

While using an aggregate predictor and pooling data over a few polls seem to increase the validity of the data, further research will be required to determine to what extent it is appropriate to use MRP estimation techniques on data collected through cluster sampling. Given data restrictions to protect the privacy of respondents, we still know very little about what clusters are chosen, and thus cannot incorporate this information into a model. Further analysis might show that the number, size, and other characteristics of the geographic clusters chosen impact the validity and efficiency of MRP estimates. It may be the case that such estimates are only fruitful in certain instances, or that certain adjustments can be made to increase the accuracy of MRP estimates in this situation, such as by using survey or sampling weights, though the benefits of using these weights in the context of multilevel models is still an open question. As MRP becomes a more popular tool in political science, it is important that we understand how it operates on polls with different sampling techniques so that we can apply it appropriately in future research.
References


Gelman, Andrew. 2013. “Everyone’s trading bias for variance at some point, it’s just done at different places in the analyses.”. URL: http://andrewgelman.com/2013/03/14/everyones-trading-bias-for-variance-at-some-point-its-just-done-at-different-places-in-the-analyses/


