Advance Access publication September 24, 2013

Political Analysis (2013) 21:449–467
doi:10.1093/pan/mpt017

How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?

Matthew K. Buttice
California Research Bureau, California State Library, Sacramento, CA 94237-0001
e-mail: matthew.buttice@library.ca.gov

Benjamin Highton
Department of Political Science, University of California, Davis, CA 95616-8682
e-mail: bhighton@ucdavis.edu (corresponding author)

Edited by Jonathan Katz

Multilevel regression and poststratification (MRP) is a method to estimate public opinion across geographic units from individual-level survey data. If it works with samples the size of typical national surveys, then MRP offers the possibility of analyzing many political phenomena previously believed to be outside the bounds of systematic empirical inquiry. Initial investigations of its performance with conventional national samples produce generally optimistic assessments. This article examines a larger number of cases and a greater range of opinions than in previous studies and finds substantial variation in MRP performance. Through empirical and Monte Carlo analyses, we develop an explanation for this variation. The findings suggest that the conditions necessary for MRP to perform well will not always be met. Thus, we draw a less optimistic conclusion than previous studies do regarding the use of MRP with samples of the size found in typical national surveys.

1 Introduction

In a representative democracy, the preferences and opinions of those in the mass public play a critical role in normative accounts and theoretical models of electoral and legislative behavior. Of course, empirical assessments require valid and reliable measures. Although national opinion polls are readily available, they have been of limited use for generating estimates of constituency preferences for subnational units.[1] The fundamental problem is that "typical" national surveys include too few respondents within subnational units for reliable opinion estimates. Thus, studies of representation have generally been limited to the use of indirect and diffuse preference measures like the two-party division of the presidential vote.

In recent years, a new method for producing estimates of preferences in subnational constituencies from national surveys has gained prominence. Multilevel regression and poststratification (MRP) is based on Gelman and Little (1997) and extended in Park, Gelman, and Bafumi (2004, 2006). It is described as the "latest advanced technique that has been used to estimate state-level public opinion as well as public opinion at other levels of aggregation" (Shapiro 2011, 999), which is "emerging as a widely used gold standard" (Selb and Munzert 2011, 456). Research employing MRP has been published in highly regarded political science journals, including three explorations of the method's utility (Lax and Phillips 2009b; Pacheco 2011; Warshaw and Rodden 2012), analyses of the translation of mass opinion into public policy

Authors' note: We appreciate helpful advice from Kyle Joyce, Eric McGhee, Matt Pietryka, and Walt Stone on this article. Buttice began work on this project while at UC Davis and finished while at the California Research Bureau. The research results and conclusions expressed in this article do not necessarily reflect the views of the California Research Bureau or California State Library. The replication archive for this article is available at the *Political Analysis* Dataverse as Buttice and Highton (2013). Supplementary materials for this article are available on the *Political Analysis* Web site.

[1] In the United States, subnational units include states, congressional districts, state legislative districts, cities, school districts, etc.

across a host of issue areas (Lax and Phillips 2009a, 2012), and the influence of public opinion on Supreme Court confirmation votes in the US Senate (Kastellec, Lax, and Phillips 2010b).

With national surveys of typical size ($N \approx 1500$), MRP has the potential to succeed where the conventional method of estimating constituency preferences—survey disaggregation—fails. Given the enormous number of available national surveys on which MRP could be used to derive opinion estimates, there is the possibility for its widespread use to investigate many political phenomena previously believed to be outside the bounds of systematic empirical inquiry. The critical question regards the quality of the MRP estimates produced with such surveys. The two previous examinations of MRP performance with samples the size of typical national surveys (Lax and Phillips 2009b; Warshaw and Rodden 2012) suggest that MRP can perform well. But, those studies are based on a limited number of cases, leaving unanswered the question of whether MRP consistently and generally produces high-quality estimates.

To investigate this aspect of MRP and explain its variation in performance, we analyze MRP opinion estimates for eighty-nine survey items and conduct a host of Monte Carlo simulations. We find that when MRP is used with national surveys of typical size, its performance is highly variable. The sources of the variation in MRP performance derive from properties of the MRP model and the nature of the distribution of the opinion being estimated. The findings of this article therefore imply important qualifications to the initial views.

2 Background

Since Miller and Stokes (1963) wrote their seminal analysis of "Constituency Influence in Congress," scholars have devoted considerable attention to understanding the nature of "dyadic" representation—the relationship between individual representatives and their constituencies. In addition, researchers also focus on "collective" representation—the relationship between the aggregate behavior of elected officials (e.g., policies enacted) and overall public opinion. Empirically analyzing dyadic representation and comparative collective representation[2] poses a variety of difficulties, chief among them the need for reliable measures of mass preferences. Ideally, across whatever geographic units one is interested in, identical surveys would be fielded with large numbers of respondents from each unit to minimize sampling error and produce reliable estimates. In practice, scholars often use large national surveys or pool multiple national surveys and then disaggregate them to the desired geographic level.

Some common limitations keep survey disaggregation from more widespread use. First, large national surveys are rare and expensive. Second, in instances where large national surveys that include the desired measures and geographic identifiers are available, scholars are often limited to cross-sectional analyses due to the lack of over-time data. Third, pooling surveys requires that identical (or very similar) questions be asked in repeated national samples, and this is an uncommon occurrence. When it does happen, it is usually for general political attitudes like partisanship and ideology, which can be used to good effect (e.g., Erikson, Wright, and McIver 1993), but leaves many interesting and important questions unanswerable. It is against this backdrop that Lax and Phillips (2009b) ask "How Should We Estimate Public Opinion in the States?" and Warshaw and Rodden (2012) ask "How Should We Measure District-Level Public Opinion on Individual Issues?" Both provide the same answer: MRP.

3 MRP

MRP is used to estimate aggregate opinion across geographic units from a survey of individual-level opinion.[3] It is based on Gelman and Little (1997) and extended in Park, Gelman, and Bafumi (2004, 2006). At a bare minimum, to employ MRP a researcher needs survey data that includes a

[2] By "comparative collective representation," we mean examining the relationship between public policy and public opinion across governing units. For example, Lax and Phillips (2009a) analyze the relationship between the adoption of gay rights laws and public opinion across the American states.

[3] Previous studies provide in-depth descriptions of MRP, especially Park, Gelman, and Bafumi (2004), Berkman and Plutzer (2005, appendix A), Gelman and Hill (2007), Lax and Phillips (2009b), and Kastellec, Lax, and Phillips (2010a).

question measuring the preference or opinion of interest (e.g., support for gay marriage) and a state identifier.[4] Typically, applications also use individual-level characteristics like age and race along with state-level covariates thought to be correlated with the aggregate opinion of interest, like state ideology or presidential vote. With the data in hand, the first step is to model preferences with a multilevel model:

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \alpha_j[i]^{x_1} + \alpha_k[i]^{x_2} + \alpha_l[i]^{x_3} + \alpha_s[i]^{\text{state}}\right),^5 \qquad \textbf{(1)}$$

where

$$\begin{aligned}
\alpha_j^{x_1} &\sim N(0,\, \sigma_{x_1}^2);\\
\alpha_k^{x_2} &\sim N(0,\, \sigma_{x_2}^2);\\
\alpha_l^{x_3} &\sim N(0,\, \sigma_{x_3}^2);\\
\alpha_s^{\text{state}} &\sim N(\alpha_n^{\text{region}} + \beta^z \times z_s,\, \sigma_{\text{state}}^2);\\
\alpha_n^{\text{region}} &\sim N(0,\, \sigma_{\text{region}}^2).
\end{aligned}$$

Individual responses are modeled as a function of individual-level characteristics x_1, x_2, and x_3 (with j, k, and l indicating the individual's category of x_1, x_2, and x_3, respectively) and the geographic unit for which one desires estimates, in this case state (with s indicating the state of residence). β^0 is an intercept term and each α is an offset. Under this setup, the effects associated with the demographic indicators (x_1, x_2, and x_3) are assumed to be normally distributed with a mean of zero and a variance unique for each item. The effects associated with individual states (α_s^{state}) are themselves modeled as a function of a state-level variable (z) as well as region of the country. And finally, the effects associated with region are assumed to be normally distributed around zero.[6]

The model's parameters may be estimated through standard statistical software packages like Stata and R. Then, predicted values for each "type" of person (r) may be computed (θ_r). If there are a categories of x_1, b categories of x_2, and c categories of x_3, then there are $a \times b \times c$ types of people in each of the fifty states for a total of $50 \times a \times b \times c$ predicted preferences.

The next stage of MRP involves poststratification to produce estimates of the state opinion means. The population frequency of each person type is typically obtained from census data. Those frequencies (N_r) serve as weights to produce the MRP preference estimate:

$$Y_s^{\text{MRP}} = \frac{\sum\limits_{r \in s} N_r \theta_r}{\sum\limits_{r \in s}}. \qquad \textbf{(2)}$$

In sum, the multilevel model produces a preference estimate for each type of person and those preferences are weighted in proportion to the frequencies of those types in the population to produce the state estimates.

4 The Performance and Potential of MRP

How well does MRP perform at producing opinion estimates? The most extensive analyses are Lax and Phillips (2009b) and Warshaw and Rodden (2012). These studies focus on comparing MRP to the more common method of disaggregating national survey data to the geographic level of interest and computing the means within those units (disaggregated means [DM]).[7] Under "optimal"

[4] Our exposition assumes the goal is to estimate state-level preferences. If the goal is to estimate, for example, congressional district preferences, then a geographic identifier indicating the congressional district would be required.

[5] The logit framework is used because the preference indicator is assumed to be dichotomous. All the previous studies except Berkman and Plutzer (2005) employ dichotomous preference indicators.

[6] For simplicity and ease of exposition, our example includes three individual-level covariates (x_1, x_2, and x_3) and a single state-level covariate (z). In practice, more individual-level and state-level covariates can be, and sometimes have been, employed.

[7] We refer to this method as "disaggregated means" (DM).

conditions, namely when overall sample sizes are large enough to produce representative samples of sufficient size in even the least populous geographic units, there is little performance difference between MRP and DM. Because the DM estimates are subject to little sampling error and therefore produce highly reliable estimates, there is little room for improvement from MRP.

As sample sizes get smaller and approach those of typical national surveys, the sampling errors associated with the DM estimates grow, thereby driving down their performance. In contrast, the multilevel model in MRP places less weight on group-level variation as sample sizes decline, thereby limiting the effect of sampling error and avoiding the cause of performance falloff for DM. Thus, while DM necessarily suffers as sample sizes decline, MRP may not. In fact, the existing evidence on this question suggests that MRP performance is barely diminished, if at all, with small samples. Lax and Phillips (2009b) analyze two items—support for same-sex marriage and presidential voting—and Warshaw and Rodden (2012) analyze six—one of which is also same-sex marriage. The findings are generally consistent across cases. Whereas the performance of DM declines precipitously with smaller samples, the performance of MRP does not.[8] If these results generalize, then the substantive implications can hardly be overstated:

> [A] sample of approximately 1,400 respondents or more can produce respectable estimates of opinion, such that the correlation to actual state opinion should be sufficiently high. This can save researchers time, money, and effort . . . Our finding that MRP performs equally as well with small and large samples of survey respondents suggests that MRP can greatly expand the number of issues for which scholars can estimate state opinion and the nuance with which they can do so . . . Using the MRP approach, scholars should now be able to measure opinion across a large set of specific policy concerns. This will greatly enhance research into the responsiveness of state governments. Additionally, since MRP can effectively be used with relatively little data and simple demographic typologies, it can also be applied to studies of public opinion over smaller time periods or in smaller geographic units, such as congressional districts or school districts, for which detailed demographic data are limited, or for other subsets of the population (Lax and Phillips 2009b, 120–1).[9]

Relying on DM *necessarily* limits the scope and type of analyses that scholars can conduct. In contrast, MRP offers the possibility of greatly expanding the range for two reasons. First, as discussed above, the data requirements for MRP are minimal—a survey with geographic identifiers and a relevant opinion question of interest. Second, while large-scale surveys are relatively rare, there are literally thousands of available national surveys that meet the MRP data requirements and are easily accessible through data repositories like the Roper Center and the Inter-university Consortium for Political and Social Research.

Two recent applications give a sense for the research possibilities opened up by the use of MRP with conventional national survey samples. Kastellec, Lax, and Phillips (2010b) analyze the relationship between Senatorial voting on ten Supreme Court nominees and state public opinion on those nominees. To do so, Kastellec, Lax, and Phillips (2010b) use MRP to produce ten sets of estimates (one set of fifty state estimates for each nominee) from national survey samples conducted near the confirmation votes that include questions asking about support for the nominees. Across the ten nominees the median sample size was just 2858 respondents, with less than 1600 respondents in surveys for three nominees. A second application is Lax and Phillips (2012), which analyzes state policy adoption with respect to thirty-nine public policies across eight policy areas. For each of the thirty-nine policies, Lax and Phillips (2012) generate MRP estimates of state public opinion on the policy from national surveys. The median sample size for the thirty-nine policies is 3010, and for eight, the sample sizes are less than 1700. For both of these studies, without MRP the empirical analyses would not have been possible.

Beyond research that has already been conducted, it is easy to identify other important questions that could be addressed by employing MRP on conventional national samples. Consider Arceneaux (2001), which analyzes the relationship between state-level gender role attitudes and the level of

[8] Although never approaching the drop-off observed for MRP, Warshaw and Rodden (2012) do find some variation in the decline of MRP performance.

[9] Park, Gelman, and Bafumi (2006) are more circumspect, even when conducting MRP with large samples. Warshaw and Rodden (2012), although generally optimistic, especially with regard to comparisons between MRP and DM, do take note of the variability in MRP performance.

women's representation in state legislatures. Arceneaux (2001) uses DM for General Social Survey (GSS) data collected over a 22-year period and analyzes the relationship between those estimates and the average level of female representation during that time period for the thirty-eight states for which DM estimates could be produced. With MRP one could produce state-level estimates for all fifty states for every year that the GSS included the gender role questions, enabling one to model within-state change in representation over time and its relationship to within-state change in public opinion, rather than strictly relying on a model where all the variation is cross-sectional and based on pooled values over a period during which attitudes and representation changed dramatically.

In light of the uses for which MRP has already been employed and to which it could be employed in the future, a critical question is whether MRP consistently performs well with samples the size of typical national surveys. Although the results in Lax and Phillips (2009b) and Warshaw and Rodden (2012) are encouraging, they are not well suited to address this issue because of the limited number of items they analyze.[10] Further, regardless of whether MRP routinely performs well, there is an important question about the causes of MRP performance. Lax and Phillips (2009b) and Warshaw and Rodden (2012) focus on the complexity of the multilevel model and show that the performance of MRP is not the mere result of partial pooling toward the grand mean. The inclusion of individual-level and especially geographic-level predictors is what appears to make MRP perform better than DM, at least for the items examined in Lax and Phillips (2009b) and Warshaw and Rodden (2012). Given our interest in the absolute performance of MRP and variation in performance across items, we do not analyze how the presence or absence of predictors in the multilevel model influences MRP performance. Instead, as explained in the next section, we focus on how well the available covariates predict opinion.

5 Hypotheses about the Performance of MRP

Given what researchers have learned about measurement error and reliability, few contemporary public opinion scholars would rely on national samples of 1500 to estimate state or congressional district opinion by simply computing the DM estimates (Erikson 2006). As a consequence, our focus is on how well MRP performs in an absolute sense. In the case of national surveys of typical size, because DM performs so poorly, the fact that MRP outperforms DM does not necessarily imply that MRP performs *well* and that its estimates should be employed in substantive analyses.[11]

In addition to focusing on the absolute performance of MRP, we also examine conditions that influence MRP performance. We expect the properties of the multilevel opinion model and the population from which survey samples are drawn to influence the performance of MRP. Two factors related to the multilevel model are the degree to which (1) the individual-level covariates and (2) the state-level covariates account for opinion. To see the rationale, consider a national survey sample with 1500 respondents randomly drawn from the fifty states and the District of Columbia. Based on the 2010 Census, there are only two states (California, Texas) where one would expect more than a hundred respondents and only five more (New York, Florida, Illinois, Pennsylvania, Ohio) where one would expect more than fifty respondents. For the other forty-four, the expected numbers of respondents range from about three to forty-eight with twenty-five states expected to have less than twenty respondents. With these small state samples, the disaggregated state means will be subject to substantial sampling error, and state-level variation in opinion that is unexplained by the state-level predictors will not be given much weight in the estimation of equation (1) and the state offsets will be pulled toward the grand sample mean, especially for the states with the very smallest samples. With minimal state offsets, differences in MRP estimates across states will be driven by the individual-level and state-level covariates

[10] However, even with just six items Warshaw and Rodden (2012) do find variation in MRP performance with small samples. The reported correlations between MRP opinion estimates and "baseline" opinion are 0.51, 0.56, 0.57, 0.70, 0.78, and 0.81.

[11] To be sure, we do confirm that MRP typically outperforms DM, as shown in the Appendix.

included in the multilevel model. As a consequence, we expect MRP performance will be improved when the opinion under consideration is better accounted for by the included covariates.[12]

A third factor we expect to influence MRP performance is a property of the population opinion distribution from which a survey sample is drawn: the degree to which true opinion varies across states relative to within states. For an intuition, consider two instances where in the national population the proportion preferring policy option A, to policy option B is 0.50. In the first case, when states are sorted from lowest to highest level of support for A they are evenly spaced between 0.49 and 0.51. In the second case, the national mean is also 0.50, but when the states are sorted, they range from 0.20 to 0.80. Given the greater variation, one would expect that MRP—and even DM—will perform better in the second case as it would be easier to differentiate one state from another. There is more variation in an absolute sense and also relative to the amount of variation within states.

6 Analyzing MRP Performance

Assessing MRP performance and testing the hypotheses requires population opinion data in addition to sample data drawn from the population. We address this issue in two ways. First, we identify eighty-nine opinion items asked of at least 25,000 respondents across five relatively large national surveys, and for each item we treat the respondents as the population for that opinion.[13] Then we repeatedly draw samples, use MRP to produce state estimates, and then analyze those estimates. Our second approach relies on simulation. We create a population and generate opinion based on parameters we explicitly set. By varying the parameters, we can isolate causes of MRP performance.

To measure the overall explanatory power of the individual-level covariates, we use McFadden's pseudo r-squared from an opinion model including only the individual-level covariates. This quantity is typically reported with logit/probit estimates from standard statistical programs and is intended to mirror the r-squared from an Ordinary Least Squares (OLS) regression. For a given opinion for which MRP will be used to produce state estimates, we estimate a logit model of opinion for the entire population with the individual-level variables included as independent variables and record McFadden's pseudo r-squared. Across items, then, we will be able to compare how well the individual-level covariates account for true opinion. Our expectation is that MRP will perform better on samples drawn from populations where the predictive power of the individual-level covariates is higher.

To measure the strength of the state-level covariates, we focus on how well the state-level covariates account for variation across states in true opinion. Specifically, we use OLS and regress true state opinion on the state-level covariates and record the r-squared. We refer to this quantity as the state-level r-squared, and given the logic described earlier, we expect that MRP performance will be better when used on samples drawn from populations where these values are higher.

To quantify how opinion varies across states relative to how it varies within states, we focus on the intraclass correlation (ICC) for the population, which is the proportion of the total variation in the population that is interstate variation. When the ICC equals its theoretical maximum (1.0), all opinion variation is across states and opinion within states is perfectly homogeneous. When the ICC equals its theoretical minimum (0), there is no opinion variation across states and all variation is within states. We expect that the performance of MRP estimates will improve when samples are drawn from populations with higher ICCs.

[12] This is a generalization of what Lax and Phillips (2009b) and Warshaw and Rodden (2012) find from their within-item analyses showing that the inclusion of individual-level and geographic-level predictor covariates improves the performance of MRP.

[13] These are described in detail below.

6.1 *MRP with Real Survey Data*

Lax and Phillips (2009b) analyze MRP performance with samples the size of conventional national surveys for two opinions and Warshaw and Rodden (2012) do it for six. To provide greater perspective and test our hypotheses, we identified eighty-nine policy preference questions asked in the two largest ongoing national academic surveys, the National Annenberg Election Studies (2000, 2004, and 2008) and the Cooperative Congressional Election Studies (2006 and 2008). The criterion for inclusion was that an item be asked of at least 25,000 respondents. The Appendix provides the complete list. Across the eighty-nine items, the median sample size is nearly 33,000.

Each item was coded as a dichotomy and we compute "true" state opinion as the disaggregated state means, thereby treating the sample as the population. We also compute the ICC, individual-level pseudo r-squared, and state-level r-squared based on all respondents. Then, sampling with replacement, we select 1500 respondents and use MRP to produce state estimates and repeat the process two hundred times for each opinion item.[14]

The first step of MRP is estimating a multilevel model of opinion. To identify covariates for inclusion, we turn to previous applications of MRP. Table 1 provides a list. The most common individual-level covariates are age, education, gender, and race/ethnicity. We include them all. There is less consistency in the state-level covariates included in previous studies, but the two most common are presidential vote and religious conservatism. Presidential vote is measured as the Republican share of the two-party vote in the most recent or concurrent presidential election. Following Lax and Phillips (2009a, 2012) and Kastellec, Lax, and Phillips (2010b), percentage religious conservative is operationalized as the percentage of state respondents that are Evangelical Protestant or Mormon.[15] The second step of MRP uses the estimates from the multilevel model to compute predicted probabilities for each type of respondent. The last step is poststratification. Because we are treating the overall samples as the populations from which we draw samples, we compute the population frequencies and poststratification weights based on them.

The most common use of MRP estimates is as independent variables in models of responsiveness. Thus, a key performance measure is how well MRP estimates correlate with true values (Lax and Phillips 2009b; Warshaw and Rodden 2012). A second measure of MRP performance is bias. For any opinion in any state, we can measure the difference between the expected value of the MRP estimate and true state opinion. The former is estimated by the mean MRP estimate across the two hundred samples for a given opinion in a state and the latter is true opinion for the state. Bias closer to zero is clearly preferable to increasingly positive or negative bias. Further, the substantive significance of, for example, a bias of five percentage points depends on the level of true interstate variation in opinion. When interstate opinion is highly dispersed, a bias of five percentage points would be less likely to alter significantly the ordering of estimated state preferences. But, the same five percentage points of bias when interstate opinion is less dispersed would be more problematic. To take this into account, we divide the bias by the standard deviation in true state opinion and refer to it as "standardized bias." Finally, although bias refers to how the estimates are *expected* to perform across many samples, in practice, applied researchers will have a single sample from which

[14] Initially, it may seem problematic to treat a sample as the population, but several factors suggest otherwise. First, consider an item on abortion opinion for which there are ninety-five Vermonters in a national sample of 30,000 respondents. We treat those ninety-five as the Vermont population. If they happen to be wildly unrepresentative, then some state-level covariates included in the multilevel opinion model (like state presidential vote) will not perform well, but this will be reflected in the measure of state-level covariate strength. To the extent that state-level covariate strength matters (and below we show that it is very important), the problem will be addressed when we analyze the relationship between the strength of the state-level covariates and MRP performance. Second, in supplemental analyses one of the state-level covariates we included in the multilevel opinion model was state ideology, which was measured as the disaggregated state mean ideology for the "population." If the ninety-five Vermonters are not typical Vermonters and exhibit abortion opinion that is not truly representative of Vermonters, then the same will likely be true for their ideological preferences. So, by including "true" ideological preferences as a state-level covariate in the MRP models, we account for this. As it turns out, for the eighty-nine items, overall MRP performance and the correlates of MRP performance were nearly identical when we included state ideology and when we did not.

[15] These data were downloaded from Kastellec's web site (http://www.princeton.edu/~jkastell/mrp_primer.html) (accessed July 11, 2011).

Table 1 Covariates used to produce MRP estimates of state opinion

Study	*Preference*	*Individual covariates*	*State covariates*
Park, Gelman, and Bafumi (2004)	Presidential vote	Sex, race, age, education	Previous presidential vote
Park, Gelman, and Bafumi (2006)	Presidential vote, partisanship, ideology	Sex, race, age, education	Previous presidential vote
Lax and Phillips (2009a)	Support for eight gay rights policies	Sex, race, age, education	Percent religious conservatives, presidential vote
Lax and Phillips (2009b)	Support for same-sex marriage	Sex, race, age, education	Percent religious conservatives
Kastellec, Lax, and Phillips (2010b)	Support for ten Supreme Court nominees	Sex, race, age, education	Percent religious conservatives, state ideology
Pacheco (2011)	Partisanship, ideology	Sex, race, age, education	None
Lax and Phillips (2012)	Support for thirty-nine public policies	Sex, race, age, education	Percent religious conservatives, presidential vote

to derive MRP estimates. Our third measure of MRP performance calculates the error that arises in such a situation. For each sample drawn for each of the eighty-nine policy items, we compute the mean absolute error (MAE), which is the average absolute difference between the MRP estimate and true value across all the states for the sample. As with bias, we divide this quantity by the standard deviation of true state opinion and refer to it as "standardized MAE."

Figure 1 shows the correlations between MRP estimates and true state values for each of the two hundred samples for each of the eighty-nine policy items. The items are ordered from the lowest average correlation to the highest. Ample variation is immediately evident. Across samples for individual items, the average standard deviation of the two hundred correlations is 0.12, with notably more intra-item variation among those items with the lowest average correlations. The bottom third of the policy items (in terms of average correlations) have an average standard deviation of 0.16 and the top third have an average less than half that of 0.07. Across items, the average correlation between MRP estimates and the true values ranges from 0.09 to 0.91, with a median value of 0.56. The 10th- and 90th-percentile values are 0.30 and 0.81, and the standard deviation is 0.19. Thus, whether looking across samples or across items, there is substantial variation in MRP performance, with poor performance not uncommon. Thirty-five percent of the samples produced MRP estimates with correlations below 0.50. For thirty-three of the eighty-nine items, the average correlation between MRP estimates and true values was less than 0.50.

Turning to the standardized bias in the MRP estimates, we measure it for each policy opinion for each state. Figure 2 shows the set of state estimates for each item, ranking them from those with the least to the most variability. Ideally, of course, bias would be small (near zero) with minimal variability across states. But this does not appear to be the case. Almost two-thirds of the standardized bias estimates exceed ±0.25, and 14% exceed ±1.0.[16] Figure 3 focuses on the magnitude of standardized bias by showing the absolute value of bias for the states across policy items, with items ranked by average bias. Consistent with the correlation and bias results, there is substantial variability in absolute standardized bias. Across the eighty-nine items, average absolute standardized bias ranges from 0.22 to 0.74; the mean is 0.53; and the 10th- and 90th-percentile values are 0.35 and 0.66.

The bias and absolute bias measures are based on the average MRP estimates across the two hundred samples for each item. As discussed above, a user of MRP will typically have a single sample and therefore measuring error on a sample-by-sample basis is also important. Thus, for each sample for each policy item, we compute the absolute error for each state and average the values to produce the mean absolute error (MAE) for each sample, which we standardize by

[16] The average standardized bias across states for each policy item is not a useful measure because equal numbers of exceedingly positively and negatively biased state estimates would produce an average bias near zero even though the set of estimates would not be informative about the true state values. The average absolute bias across states for each policy opinion is more informative, and we report those findings next.

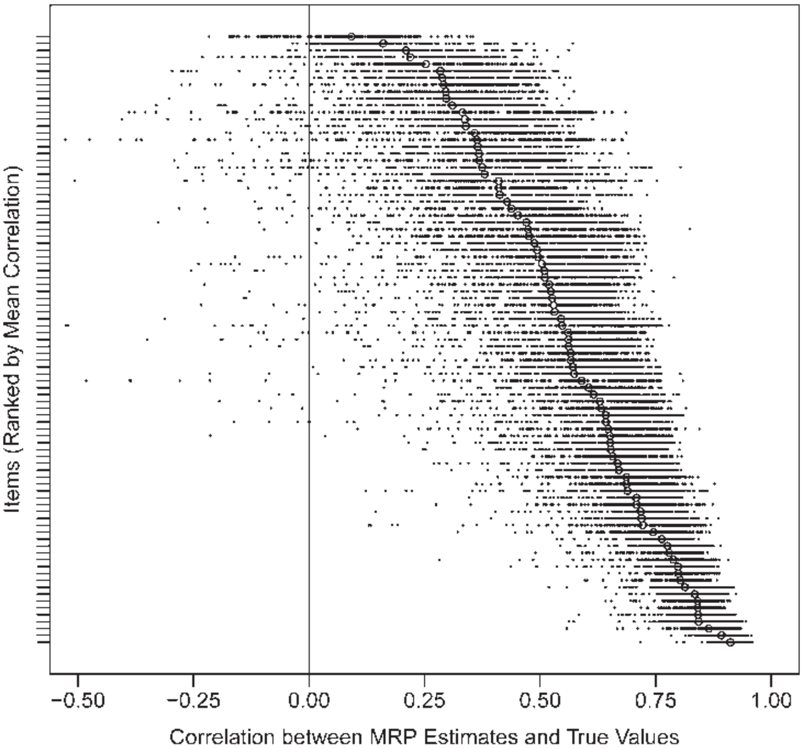

Fig. 1 MRP performance across items and samples (correlation with true values). Each row of the figure is for one of the eighty-nine policy items. For each item, the average correlation with the true state values is represented with a hollow circle. Each dot represents the correlation for one of the two hundred samples drawn for each item.

dividing by the standard deviation in true state preferences for the policy item. The two hundred standardized MAEs for each of the eighty-nine policy items are displayed in Fig. 4, which orders them by average MAE. Visual inspection again shows substantial variation across samples for single items and across items. The average standardized MAE ranges from a low of 0.35 to a high of 1.18, with 10th- and 90th-percentile values of 0.48 and 0.89, respectively. Nearly half of the policy items have an average standardized MAE > 0.75, which means that the average state error is typically at least three-quarters as large as the standard deviation of true state opinion.

Across the eighty-nine items, much of the variability in MRP performance can be accounted for. Figure 5 shows the relationship between the three hypothesized causes of MRP performance and each of the measures of MRP performance. The first row of figures shows the relationships between the three performance measures and the strength of the individual-level covariates. Across all three measures, the relationship is weak, with little observed improvement in performance associated with stronger individual-level models. This is not the case for the strength of the state-level covariates, shown in the second row of Fig. 5. As the strength of the state-level covariates increases from its lowest observed values (near 0.0) to its highest (nearly .8), the average correlation between MRP estimates and true values increases considerably. Likewise, the average absolute standardized bias decreases, as does the average standardized MAE. Thus, for all three measures, MRP performance is considerably better at higher observed levels of state-level covariate strength compared with lower levels. Equally strong relationships are evident when the ratio of interstate to intrastate variation in preferences (the ICC) is examined. The bottom row of Fig. 5 shows how the (natural log of) the ICC is related to MRP performance. In each case, MRP performance is substantially better (larger average correlations between MRP estimates and true values, smaller average absolute standardized bias, and smaller average standardized MAE) at higher values of the ICC to lower ones. Together the factors account for 92% (average correlation),

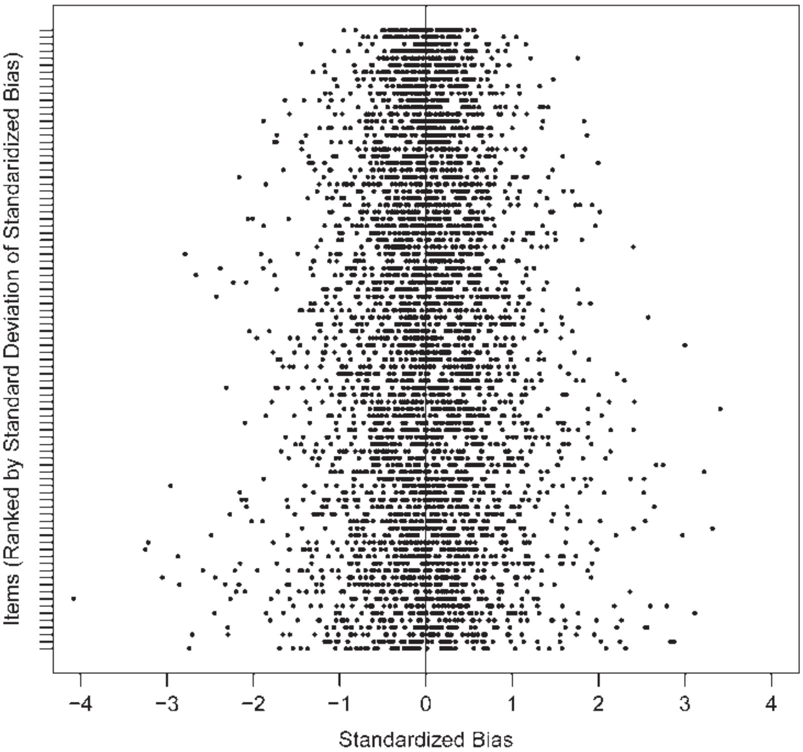

Fig. 2 MRP performance across items and states (standardized bias). Each row of the figure is for one of the eighty-nine policy items. Each dot represents the estimated standardized bias for a single state.

79% (average standardized bias), and 89% (average MAE) of the observed variation in the MRP performance measures.[17]

As discussed earlier, Lax and Phillips (2009b) investigate the small sample properties of MRP with opinion on same-sex marriage and report impressive performance. Warshaw and Rodden (2012) analyze six items and find that MRP performance is notably better with three (same-sex marriage, abortion, and stem cell research) compared with the others (social security, minimum wage, and the environment). Our results are consistent with these findings. We identified thirteen items among the eighty-nine that are similar to the ones that performed well in Lax and Phillips (2009b) and Warshaw and Rodden (2012). Six address abortion, five focus on gay rights, and two deal with stem cell research. For these items, the average correlation with true values is notably higher (0.81 versus 0.51 for the other items); the average absolute bias is substantially lower (0.35 compared with 0.54); and the average MAE is also considerably smaller (0.49 versus 0.75). Clearly on "cultural" items like these, the performance of MRP is unusually strong.[18]

In sum, the analysis of the eighty-nine policy items finds substantial variability in MRP performance, with poor performance not uncommon. An exception is apparent with the set of cultural items relating to abortion, gay rights, and stem cell research for which the MRP estimates are noticeably better. The variation in performance across items appears due to variation in one aspect of the multilevel model on which the MRP estimates are based (the strength of the state-level

[17] These figures are based on regressing each of the performance measures on the three hypothesized causes of MRP performance. Consistent with Fig. 5, the estimated effects of the ICC and the strength of the state-level covariates are substantial, whereas the apparent effects of the strength of the individual-level covariates are small and in two cases (absolute standardized bias and average standardized MAE) cannot confidently be distinguished from zero ($p = 0.89$ and $p = 0.20$, respectively).

[18] The explanation appears to be that for the cultural items the strength of the state-level predictors is high (averaging 0.63 versus 0.31 for the other items), as is the population ICC (averaging 0.024 versus 0.009 for the other items). Whether these conditions would hold during other time periods or when other covariates are used in the multilevel model remains an open question.

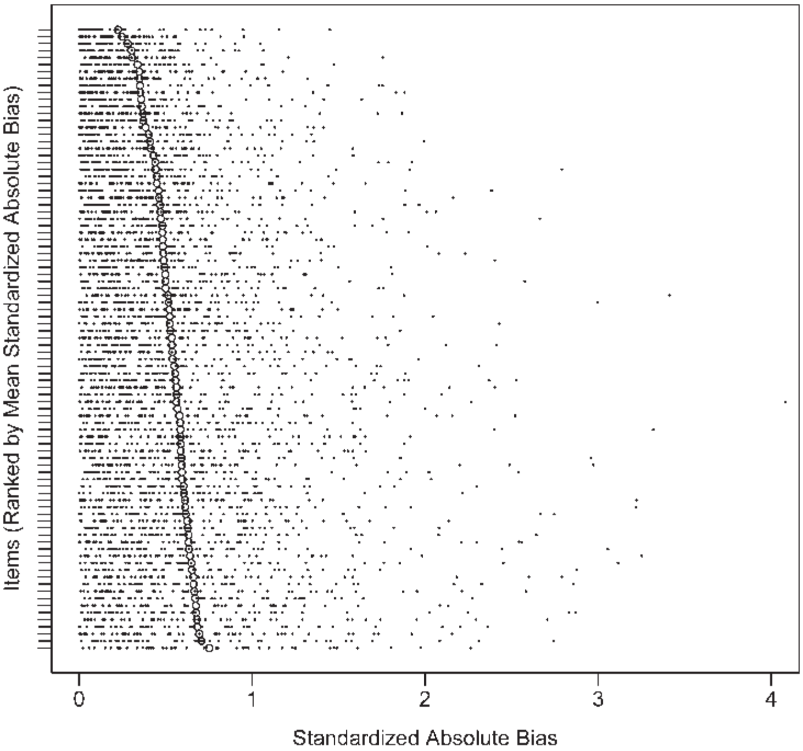

Fig. 3 MRP performance across items and states (absolute standardized bias). Each row of the figure is for one of the eighty-nine policy items. For each item, the average absolute standardized bias is represented with a hollow circle. Each dot represents the absolute standardized bias for a single state.

covariates) and one aspect of the population from which samples are drawn (the ICC). On the basis of these results, one would be hard pressed to infer that MRP would consistently produce highly reliable opinion estimates from typical national surveys.

6.2 *MRP with Simulated Survey Data*

The results based on the eighty-nine survey items reveal highly variable performance when MRP is used with samples the size of typical national surveys. They also suggest that a large population ICC and strong state-level predictors may facilitate significantly enhanced performance. Disentangling the effects is hampered by the empirical relationship between the two. Across the eighty-nine policy items analyzed above, the correlation between the strength of the state-level covariates and the (natural log of) the population ICC is 0.49. There are only a small number of items where the state-level covariates are strong but the ICC is low. Likewise, there are only a few instances where the ICC is high but the state-level covariates are weak.

To gain a deeper understanding of the conditions that influence MRP performance, we turn to a series of Monte Carlo simulations. These simulations give us control over the data generation process, which enables us to manipulate the population ICC and strength of the state-level covariates. Drawing on the range of observed values for the eighty-nine items, we conduct four sets of simulations, varying the population ICC and the strength of state-level covariates. For each variable, we use the values associated with the 5th and 95th percentiles of their observed distributions from the eighty-nine items analyzed above.[19] For the population ICC, the values are 0.002 and 0.04; for the state-level r-squared, they are 0.05 and 0.82.

[19] We focus on the observed rather than theoretical ranges for the variables because our interest is in empirical applications.

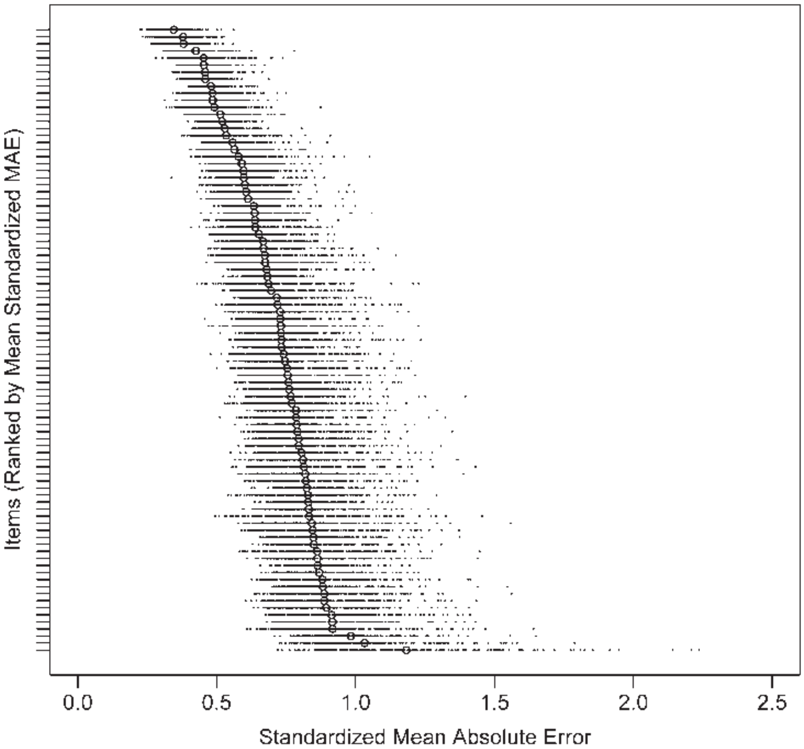

Fig. 4 MRP performance across items and samples (standardized MAE). Each row of the figure is for one of the eighty-nine policy items. For each item, the average standardized MAE is represented with a hollow circle. Each dot represents the standardized MAE for one of the two hundred samples drawn.

The simulations entail several steps. For each condition, we begin by generating an opinion indicator for 100,000 "people." Specifically, y is distributed B(100,000,0.50) so that 50% favor a policy and 50% oppose it. We then create three individual-level covariates:

$$\begin{aligned} x^*_{1i} &= y_i + \alpha \times e_{1i} \\ x^*_{2i} &= y_i + \alpha \times e_{2i} \\ x^*_{3i} &= y_i + \alpha \times e_{3i}. \end{aligned} \tag{3}$$

For each individual i, x^* is determined by the true preference (y) and random error (α^*e). With each $e \sim N(0,1)$, α is a parameter that determines the strength of the relationship between the individual-level covariates and y. Thus, we can control the predictive power of the individual-level covariates by varying α. As α increases, for each x^* the ratio of true opinion to random error declines, lowering the overall predictive power of the covariates. Because the poststratification step of MRP requires frequency distributions across categories of the individual-level covariates, we divide x^*_1, x^*_2, and x^*_3 at their respective quartiles to turn each into four-point variables (x_1, x_2, and x_3), which are used in the multilevel preference model.

The next step is to assign individuals to states. We do this by sorting the population based on a variable ($state^*_i$) and then assigning cutpoints to allocate individuals to fifty "states" in proportions that match the actual state proportions based on the 2010 US Census. For example, after sorting by $state^*_i$, individuals 1 through 1551 are assigned to the first state, giving it 1.55% of the population (1551/100,000 = 0.0155) which corresponds to the population proportion for Alabama. If $state^*$ is a purely random variable, then there will only be trivial differences in the mean preference across states and the ICC will approach 0. To the extent that there is "preference sorting" into states, then interstate preference differences will grow, as will the ICC. To control the degree of preference sorting, we determine $state^*_i$ with the following:

$$state^*_i = y_i + \delta \times u_i. \tag{4}$$

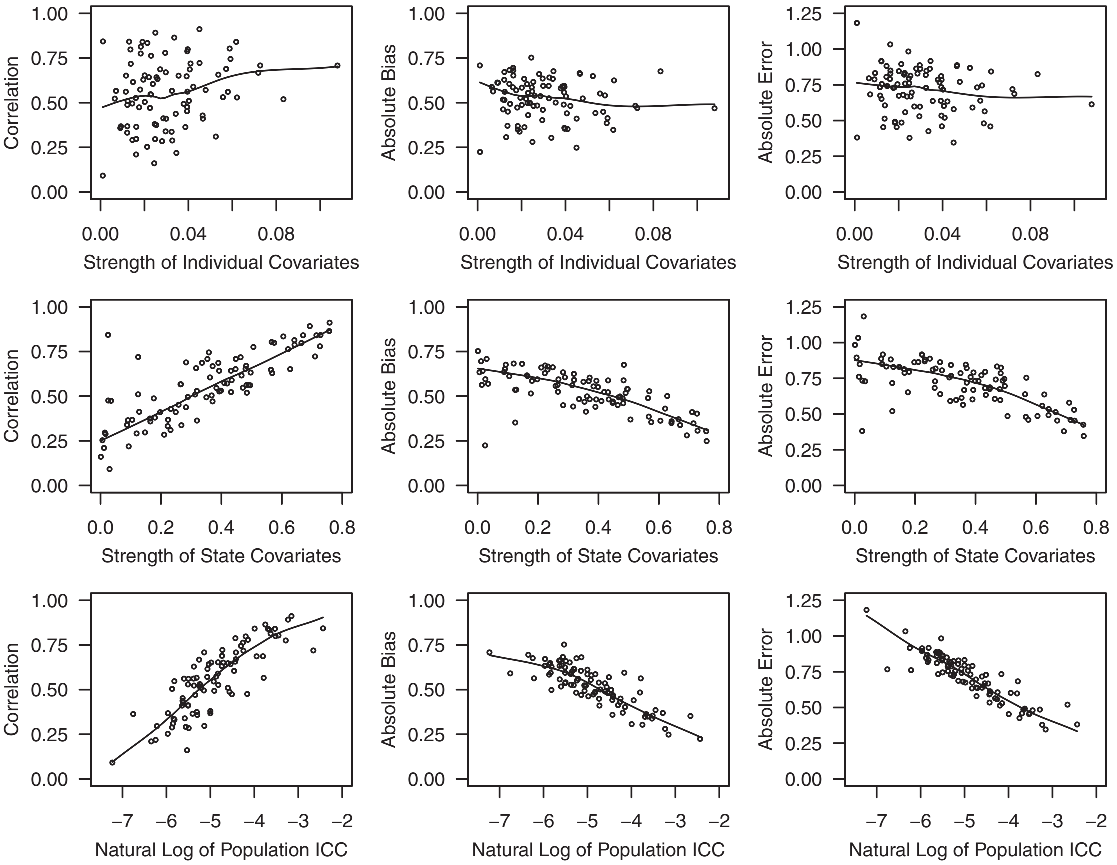

Fig. 5 The correlates of MRP performance. Each figure shows the relationship between one of the hypothesized causes of MRP performance (x-axis) and one of the measures of MRP performance (y-axis).

In equation (5), $u_i \sim N(0,1)$ and δ are the parameters that influence the degree of preference sorting. When δ is large, *state** will be mostly random and there will not be much interstate variation in policy preferences and the value of the ICC will therefore be low. When δ is small, *state** will be more strongly influenced by policy preferences producing greater interstate preference variation and a larger ICC.

The last step of the data generation process is to create a state-level covariate. True state-level opinion (Y_s) is simply the mean preference within a given state. Given our interest in the overlap between true state opinion and state-level covariates, we create a state-level proxy for Y_s that reflects true opinion and random noise:

$$z_s = Y_s + \gamma \times v_s. \tag{5}$$

With $v_s \sim N(0,1)$, the parameter γ determines the strength of the relationship between the state-level covariate (z_s) and Y_s. As γ increases, the shared variance between z_s and Y_s decreases, thereby lowering the value of the state-level r-squared.

In sum, there are three parameters for the data generation process (α, δ, and γ). Because we are interested in the independent effects of the ICC and the strength of the state-level covariates, we assign a value of α that produces a pseudo r-squared equal to the median value from the eighty-nine items analyzed earlier. We then manipulate the values of δ and γ to create the four conditions defined by the population ICC and state-level r-squared taking on their 5th and 95th observed percentile values. For the four conditions, we create a data set of 100,000 observations, repeatedly (five hundred times) draw a sample of $N = 1500$, and produce MRP opinion estimates of state-level opinion (Y_s^{MRP}). We assess the quality of the MRP estimates with the three performance measures used earlier. In addition, for each of the five hundred samples within each of the four conditions we compute bootstrapped standard errors and confidence intervals for the estimates by resampling with replacement—1000 times—from the samples and producing MRP estimates from them. This enables us to assess the coverage of the MRP estimates, as well as the other three measures of MRP performance.[20]

The Monte Carlo results are summarized in Table 2. Most generally, they substantiate the results from the analysis of the eighty-nine policy items by demonstrating substantial variability in MRP performance. First, consider the correlation between MRP estimates and true values. Table 2 reports the average of the five hundred correlations for each condition. When the strength of the state-level covariates and the population ICC are both at their low values, the average correlation is just 0.17. Increasing the ICC to its high value (while keeping the strength of the state-level covariates at its low value) produces an average correlation of 0.61 while increasing the strength of the state-level covariates to its high value (while keeping the ICC at its low value) raises the average correlation to 0.66. When both at their high values, the average correlation is 0.87, and it is consistently high, with each of the five hundred sets of MRP estimates in this condition having a correlation greater than 0.75 with the true values.

Parallel to the increase in the average correlations, the standardized bias is reduced when either—and especially both—the strength of the state-level covariates or the population ICC are set to their high rather than their low values. Compared with its level in the condition where both parameters are at their low values (0.71), the average standardized bias is more than cut in half when both parameters are at their high values (0.34). Nearly three in four samples (70%) have an average standardized bias of less than 0.50 when both parameters are high, compared with 46% when both are low. The same patterns are evident for the average standardized MAE. Across the three measures, then, the Monte Carlo results show that MRP performance is neither uniformly strong nor uniformly weak. The quality of the estimates depends crucially on how well the state-level covariates account for opinion variation across the states along with how the interstate differences in opinion compare to the intrastate differences.

Next consider the coverage rates for the bootstrapped confidence intervals of the MRP estimates. We focus on the 90% confidence interval and produce an estimated interval for each of the fifty states for each of the five hundred samples in each of the four conditions. Each of these 100,000 (50*500*4) intervals either includes the true state value or it does not. Ideally, when using a 90% confidence interval, close to 90% of the estimated confidence intervals would include the true value. We find that this is generally not the case for the MRP estimates. The overall coverage rates across the four conditions are 85%, 80%, 91%, and 79%. Thus, sometimes the rates are too high, indicating that the confidence intervals are too conservative, and sometimes the rates are too low, indicating that they are too narrow.

Further, within each condition we can examine the coverage rates by state, and we find additional variability. The proportion of states with coverage rates close to 90% (between 85% and 95%) peaks at just 48% in the condition where the strength of the state covariates is high and the population ICC is low. In this condition, twenty-four states have coverage rates between 85% and 95%, whereas ten states have coverage rates below 85% and sixteen states have a rate above 95%. Across all four conditions, more than 50% of the states have coverage rates that are too high or too low. In the condition where MRP performs best with regard to the correlation with true values, standardized mean absolute bias, and standardized MAE, only 28% of the states have coverage rates close to the nominal rate of 90%.

7 Summary and Implications

The analysis of eighty-nine policy items and Monte Carlo simulations substantiates the key proposition that with samples the size found in typical national surveys, the performance of MRP is highly variable. Sometimes MRP produces opinion estimates that correlate strongly with true values, have little bias, modest error, and reasonable coverage; sometimes they do not. The

[20] We limit our coverage analysis to the Monte Carlo simulations because analyzing the coverage for the eighty-nine policy items would have been prohibitively time consuming.

Table 2 Monte Carlo results

	Condition			
Strength of state covariates	*Low*	*Low*	*High*	*High*
Population ICC	*Low*	*High*	*Low*	*High*
Correlation with true values				
Average	0.17	0.61	0.66	0.87
Percent of samples < 0.50	99	8	17	0
Percent of samples $0.50 - 0.75$	3	90	36	0
Percent of samples > 0.75	0	2	47	100
Absolute standardized bias				
Average	0.71	0.45	0.38	0.34
Percent of samples < 0.50	46	64	70	70
Percent of samples $0.50 - 0.75$	12	18	16	24
Percent of samples > 0.75	42	18	14	6
Absolute standardized error				
Average	0.92	0.62	0.67	0.42
Percent of samples < 0.50	0	1	10	97
Percent of samples $0.50 - 0.75$	3	98	63	3
Percent of samples > 0.75	97	1	27	0
90% confidence interval coverage				
Average	0.85	0.80	0.91	0.79
Percent of states < 0.85	40	32	20	42
Percent of states $0.85 - 0.95$	38	16	48	28
Percent of states > 0.95	22	52	32	30

Note. For each of the four combinations of strength of state covariates and population ICC, five hundred trials were conducted based on sample sizes of 1500.

general explanation for MRP performance offered in this article can account for the findings reported in Lax and Phillips (2009b) and Warshaw and Rodden (2012) with regard to the relatively high quality of MRP estimates for cultural policy preferences. In contemporary American politics, there is greater geographic heterogeneity on cultural issues than on other issues, and this variation is more readily accounted for by the sort of geographic covariates typically used in MRP analyses.

In light of the findings reported in this article, what should one make of existing studies that rely on MRP with conventional national survey samples? Consider Lax and Phillips (2012), who analyze state policy adoption for thirty-nine different policies across a host of issue domains. Lax and Phillips (2012) estimate an identical multilevel model for each of the thirty-nine surveys. Based on the results in the present article, it seems likely that there would be substantial variability in the correlations between the MRP estimates and true values across the thirty-nine policies. Even if the translation of public opinion into public policy was truly equal across policies (equal responsiveness) one would observe unequal responsiveness. Therefore, when Lax and Phillips (2012) report that the relationship between MRP opinion estimates and policy adoption varies in strength across the policies, it is unclear whether it represents true variation in responsiveness, as Lax and Phillips (2012) conclude, or whether it only *appears* that responsiveness varies because of the variation in the quality of the MRP estimates across policies.

Looking forward, there are a variety of implications of the findings reported here for scholars considering the use of MRP with conventional national surveys. First, nothing in this article suggests that previous methods like DM should be used instead of MRP. In the instances where MRP performs poorly, DM tends to perform even worse, as documented in the Appendix. Second, although a national sample of typical size can produce respectable estimates of opinion, this will not always be the case. Of critical importance when using MRP with conventional national survey samples is the inclusion of geographic-level covariates that account for a substantial amount of the geographic variation in true opinion. Thus, we strongly endorse "optimizing an MRP model for a particular research question" (Warshaw and Rodden 2012, 218). Although it is relatively easy to apply the same set of geographic covariates for all opinion items in an analysis, as we have done with our eighty-nine items and Lax and Phillips (2012) do, when the goal is to produce consistently high-quality MRP estimates, particular care should be employed in selecting geographic predictors. The predictors that work well for cultural issues probably will not work well for other issue domains and vice versa.[21]

The difficulty posed by the need for strong geographic predictors is increased because in any applied MRP analysis, one only has a sample of opinions from the population. The population-level parameters necessary for assessing the quality of the MRP estimates are, of course, unavailable. Researchers cannot empirically determine the strength of their geographic-level covariates.[22] The implication is that when MRP is used with conventional national survey samples, a persuasive theoretical argument should accompany the description of the covariates included in the multilevel model. Researchers should explain why there is good reason to believe that the geographic covariates are strong proxies for the opinion being estimated.

Strong state-level predictors emerge as a necessary but not sufficient condition for MRP to perform well. The analysis of the eighty-nine items and Monte Carlo analyses also point to the importance of the population ICC. Substantial interstate heterogeneity in preferences relative to the amount of intrastate preference heterogeneity appears necessary for MRP to produce high-quality estimates. In the absence of a high population ICC, MRP often does not perform well. And, like the state-level r-squared, reliably estimating the population ICC from a single sample is not possible. This further adds to the difficulty of assessing the likely quality of the estimates produced by MRP.

Finally, the importance of sample size for the quality of MRP estimates should not be understated or overlooked. To see the effect of sample size, we repeated the analysis of the eighty-nine policy items setting the sample size to $N = 10{,}000$ rather than $N = 1500$, and a notable improvement in performance was evident.[23] The average correlation with true values increased to 0.75 from 0.58; the average standardized bias was reduced to 0.38 from 0.52; and the average MAE dropped to 0.49 from 0.71. The greatest improvement was typically for those items on which MRP performance was weakest, with samples of 1500. For those items where MRP performed well with small samples, there was still improvement in performance with larger samples, but the magnitude was more modest. Therefore, the value of larger sample sizes should be kept in mind when designing research, writing grant proposals, and collecting data.

8 Conclusion

The possibility that one might be able to generate reliable estimates of opinion across the fifty states, the 435 House districts, or other geographic/political units from a sample the size of a typical national survey is certainly appealing. Many scholars could conceive of a host of valuable research questions to address that would otherwise lie beyond the scope of systematic empirical analysis. MRP has the potential to unlock these possibilities. Relying on the work of Gelman and Little (1997) that was extended in Park, Gelman, and Bafumi (2004, 2006), Lax and Phillips (2009b) and Warshaw and Rodden (2012) take MRP into a different survey context—one where sample sizes are much smaller—and find instances where the performance of MRP remains strong. Although we agree that its performance exceeds that of other methods and that the margin of overperformance is greatest with smaller sample sizes, we also find substantial variation in how well MRP performs.

[21] To be sure, this may be easier said than done. In other analyses, we experimented with a variety of covariates that have not been used in previous studies but that may be more suitable for public opinion on economic issues. Including individual-level measures of income and state-level measures of unemployment, median household income, and poverty rate does produce some improvement in the MRP estimates, but substantial variability in performance remains.

[22] In supplemental analyses, we investigated the possibility of inferring the state-level r-squared value on the basis of having a single sample of data. For example, one could use the multilevel opinion model to determine the amount of state-level sample variation accounted for by the state-level covariates. But, even if the state-level covariates account for all the population variation in opinion, they account for considerably less of the sample variation because, with small samples, variation across states is driven by population variation and sampling error. Moreover, we found that although there is a relationship between the state-level r-squared and the amount of state-level sample variation accounted for by the state-level covariates, it is not sufficiently strong to infer reliably the population state-level r-squared value.

[23] But, hardly any improvement was evident when we increased sample sizes to 3000.

After extensive analysis with actual and simulated survey data, the key factors we have identified that determine how well MRP performs are the strength of the geographic-level covariates included in the multilevel model of opinion and the ratio of opinion variation across geographic units relative to opinion variation within units. When these values are sizable, then MRP will often produce reliable estimates from national surveys of conventional size. However, the empirical analysis suggests that often these conditions will not be satisfied. Certainly scholars should not assume that they are met. One should therefore not presume that the properties of the MRP model are sufficient to produce the desired opinion estimates from conventional national survey samples.

Appendix

Opinion Items

As described in the main text, we identified eighty-nine opinion items for analysis that were asked of at least 25,000 respondents in the National Annenberg Election Studies (NAES) and the Cooperative Congressional Election Studies (CCES). The items, with variable names in parentheses, are as follows:

NAES 2000: cutting taxes v. strengthening social security (cbb05), health care spending for uninsured (cbe02), universal health care for children (cbe08), poverty a problem (cbp01), social security spending (cbc01), invest social security in stock market (cbc05), military spending (cbj07), tax rates a problem (cbb01), prescription coverage for seniors (cbe05), right to sue HMOs (cbe14), abortion restrictions (cbf02), death penalty (cbg01), gays in military cbl01) job discrimination (cbl05), school vouchers (cbd02), handgun licenses (cbg05), restrict gun purchases (cbg06), underpunished criminal problem (cbg12), job discrimination (cbm01).

NAES 2004: reduce taxes (ccb13), aid to schools (ccc40), income inequality (ccc41), military spending (ccd03), invest social security in stock market (ccc32), abortion ban (cce01), marriage amendment (cce21), school vouchers (ccc39), gun control (cce31), free trade agreements (ccb82), homeland security spending (ccd57), Patriot Act (ccd67), rebuilding Iraq spending (ccd34), American troops in Iraq (ccd35).

NAES 2008: tax rates-a (cbb01), tax rates-b (cbb01), immigrant path to citizenship (cdd01), border fence with Mexico (cdd04), abortion availability (cea01), same-sex marriage (cec01), environment v. economy (cfb01), American troops in Iraq (cdb01).

CCES 2006: minimum wage (v2072), social security private accounts (v3024), minimum wage (v3072), capital gains tax rates (v3075), taxes v. spending (v4040), taxes v. spending v. borrowing (v4044), abortion (v3019), late-term abortion (v3060), stem cell funding (v3063), illegal immigrant citizenship (v3069), environment (v3022), affirmative action (v3027), free trade—CAFTA (v3078), military use—oil supply (v3029), military use—terrorist camps (v3030), military use—genocide (v3031), military use—spread democracy (v3032), military use—protect allies (v3033), military use—help UN (v3034), Iraq troop withdrawal (v3066).

CCES 2008: balanced budget (cc309), privatizing social security (cc312), minimum wage (cc316b), health insurance for children (cc316e), assistance for housing crisis (cc316g), taxes v. spending (cc420), abortion (cc310), stem cell research (cc316c), gay marriage (cc316f), jobs v. environment (cc311), affirmative action (cc313), eavesdropping without court order (cc316d), free trade-NAFTA (cc316h), bank bailout (cc316i), carbon tax (cc422), Iraq troop withdrawal (cc316a), military use—oil supply (cc418_1), military use—terrorist camps (cc418_2), military use—genocide (cc418_3), military use—spread democracy (cc418_4), military use—protect allies (cc418_5), military use—help UN (cc418_6), internet absentee voting (cc419_1), election day registration (cc419_2), voter eligibility (cc419_3), vote by mail (cc419_4), automatic registration (cc419_5), photo ID to vote (cc419_6).

MRP versus DM

The highly variable and not uncommonly poor performance of MRP reported in the main text should not be construed as implying that the conventional method of estimating opinion across geographic units—DM—is a better alternative. Consistent with Lax and Phillips (2009b) and Warshaw and Rodden (2012), we find that with national samples of conventional size, MRP routinely outperforms DM, with one exception described below.

To compare the performance of MRP and DM with national samples of 1500 respondents, we return to the eighty-nine policy items analyzed in the main text and the performance measures. Across the eighty-nine items, the average correlation between MRP estimates and true state values is 0.56, compared to 0.36 for the average correlation between DM estimates and true state values. On eighty-one of the eighty-nine items, the average correlation for the MRP estimates exceeded that of the average correlation for the DM estimates. With regard to the average standardized MAE, the MRP estimates also typically outperform the DM estimates. For the eighty-nine policy items, the average standardized MAE for the MRP estimates is 0.71; the average standardized MAE for the DM estimates is almost three times as large, 1.97. And, for none of the eighty-nine items is the average standardized MAE for the DM estimates smaller than that for the MRP estimates.

The exception to the pattern of MRP outperforming DM is with respect to bias. The average absolute standardized bias for the MRP estimates is 0.52, more than three times that for DM 0.14. Further, the MRP estimates underperform the DM estimates on every one of the eighty-nine policy items. The explanation for DM outperforming MRP on the bias measure underscores the limitation of the bias measure for applied researchers. Recall that for a single state on a single policy item, bias is the difference between the *expected* value of an estimate for the state and the true value for the state. We approximate the expected value of a state estimate with the average estimate across the two hundred samples. Had we drawn an infinite number of samples, the expected value of the DM estimate would equal the true state value because the DM estimate is simply the state mean for each sample, and the mean of an infinite number of sample means is equal to the true mean. Thus, DM estimates, on average, should reproduce the true state mean and therefore have no bias. Although we do find some, which we attribute to the fact that we approximated the expected value with two hundred (rather than an infinite number of) samples, it is not surprising that DM would outperform MRP on the bias measure. But, as discussed in the main text, because bias is based on the expected estimates and applied researchers will have only a single sample from which to derive estimates, the other performance measures are more informative. And, as discussed above, MRP does typically outperform DM on those.

References

Arceneaux, Kevin. 2001. The "gender gap" in state legislative representation: New data to tackle an old question. *Political Research Quarterly* 54:143–60.

Berkman, Michael B., and Eric Plutzer. 2005. *Ten thousand democracies: Politics and public opinion in America's school districts*. Washington, DC: Georgetown University Press.

Buttice, Matthew K., and Benajmin Highton. 2013. Replication data for: How does multilevel regression and poststratification (MRP) perform with conventional national surveys? Dataverse Network. http://hdl.handle.net/1902.1/22001 (accessed September 17, 2013).

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 1993. *Statehouse democracy*. New York, NY: Cambridge University Press.

Erikson, Robert S. 2006. Constituency influence in Congress. *American Political Science Review* 100:674–4.

Gelman, Andrew, and Thomas C. Little. 1997. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* 23:127–35.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Kastellec, Jonathan P., Jeffrey R. Lax, and Justin H. Phillips. 2010a. Estimating state public opinion with multi-level regression and poststratification using R. Unpublished manuscript, Princeton University.

———. 2010b. Public opinion and senate confirmation of Supreme Court nominees. *Journal of Politics* 72:767–84.

Lax, Jeffrey R., and Justin H. Phillips. 2009a. Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review* 103:367–86.

———. 2009b. How should we estimate public opinion in the states? *American Journal of Political Science* 53:107–21.

———. 2012. The democratic deficit in the states. *American Journal of Political Science* 56:148–66.

Miller, Warren E., and Donald E. Stokes. 1963. Constituency influence in Congress. *American Political Science Review* 57:45–56.

Pacheco, Julianna. 2011. Using national surveys to measure dynamic U.S. state public opinion: A guideline for scholars and an application. *State Politics and Policy Quarterly* 11:415–39.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* 12:375–85.

———. 2006. State-level opinions from national surveys: Poststratification using multilevel logistic regression. In *Public opinion in state politics*, ed. J. E. Cohen. Palo Alto, CA: Stanford University Press.

Selb, Peter, and Simon Munzert. 2011. Estimating constituency preferences from sparse survey data using auxiliary geographic information. *Political Analysis* 19:455–70.

Shapiro, Robert Y. 2011. Public opinion and American democracy. *Public Opinion Quarterly* 75:982–1017.

Warshaw, Christopher, and Jonathan Rodden. 2012. How should we measure district-level public opinion on individual issues? *Journal of Politics* 74:203–19.