Small Area Estimation from Multiple Overimputation*

James Honaker † Eric Plutzer ‡

April 1, 2016

Abstract

Researchers of state and local politics often want to uncover relationships between local-level attitudes, and local policy implementation. This requires measuring mean attitudes in small geographies (states, counties, school districts) when the only survey data is from a national sample.

We show how to estimate these "small area" means using multiple overimputation (MO) and compare to other approaches. MO is a framework that unifies treatment of missing data and measurement error. We reconceptualize the objective as a classic missing data problem where demographic or Census data exists for individuals in every geography, but attitude responses exist for only a relatively small number of observations taken from national polls. Treating small areas as a missing data problem allows us to apply the large body of statistical theory from that literature. When additional aggregate data is available to inform these estimates, we show how MO can implement a second stage model to create more precise small area estimates.

We demonstrate this with a model estimating the local support for teaching evolution in schools, using national polls, Census microsamples, and aggregate variables measuring local religious participation.

*We would like to thank Michael Alvarez, Matthew Blackwell, Garrett Glasgow, Jeffrey Kraus, James Lo, Julianna Pacheco and Boris Shor for helpful comments and discussions.

†Senior Research Scientist, The Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street CGIS Knafel Building, Room 350, Cambridge, MA 02138 jhonaker@iq.harvard.edu; http://hona.kr

‡The Pennsylvania State University, Department of Political Science Pond Laboratory University Park, PA 16802; exp12@psu.edu

1 Local Opinion and Theories of Representation

Do democratic systems provide a meaningful role for ordinary citizens to influence public policy? This question grounds a large empirical research program often referred to as "policy responsiveness" or "policy congruence" studies. These studies tend to follow one of three traditional approaches:

- The ***dyadic representation*** tradition examines the congruence between constituent public opinion and the roll call votes cast by individual legislators. Miller and Stokes's (1963) pioneering work launched this field, followed by a key replication by Erikson (1978) that laid the groundwork for more contemporary work. Bartells (2008) and Gilens (2005) work on inequality in representation is in this tradition, as is Karols (2007) innovative use of *Literary Digest* data to explore the impact of public opinion polls on representation.

- The ***systemic representation*** tradition examines the congruence between aggregate public opinion and system level outputs such as laws. These include time series of a single polity (Erikson, MacKuen and Stimson 2002); and cross sectional analysis of subnational units such as nations (Brooks and Manza 2006), states (Erikson, Wright and McIver 1993; Brace et al. 2002; Norrander 2000; Lax and Phillips 2009; Berkman and Plutzer 2009); counties (Percival, Johnson and Niemann 2009); or school districts (Berkman and Plutzer 2005).

- The ***representative bureaucracy*** tradition is somewhat smaller–at least if restricted to studies that seek to examine substantive representation as opposed to the impacts on symbolic representation on administrative outcomes. This work in the substantive tradition includes those on implementation of welfare policies (Weissert 1994; Fording, Soss and Schram 2007), science curriculum (Berkman and Plutzer 2012), and race differences in incarceration rates (Percival 2010; see also Percival 2004).

Common to all empirical efforts across all three traditions is the challenge of measuring public opinion. Such measurement usually runs afoul of one or more of three different problems. The most fundamental is a "small area" problem. National public opinion polls and major academic surveys are typically designed to have a small margin of error (in the range of plus or minus 4% or less) for estimates of national opinion. When subset to major demographic groups such as those based on race or sex, the margin of error can become substantially larger but remains useful. But when the survey data are used to estimate small geographies such as cities, counties, or even medium-sized states, the sample sizes are simply too small to generate estimates with informative confidence intervals (though see Karol's, 2007, exploitation of the massive Literary Digest polls of the 1930s). Indeed, for cities, counties or school districts, a majority of these small polities will not have a single respondent in even the largest political or opinion surveys. Small area estimation refers to various techniques intended to use auxiliary information to generate valid and reliable estimates under these circumstances. Auxiliary

information most typically includes the demographic composition of the small polity, or informative priors based on Bayesian shrinkage models.

Two secondary problems often arise as side effects of solving the first. Many scholars pool data from multiple sources to generate large sample sizes. In doing so, they often combine data that is incommensurate due to differing question wording, sampling schemes, or modes of interview. To sidestep the incommensurate measurement problem, scholars are attracted to large repeated surveys such as the General Social Survey (GSS) and the National Election Studies (NES). These tend to use identical question wording and similar sampling methods repeatedly over many years. Pooling the data from several cross sections of these surveys can increase statistical power, but generates a third challenge. These data sets were designed to be nationally representative and the multi-stage cluster sampling schemes are highly efficient at the national level. But they virtually ensure that state subsamples will be biased. For example, if metropolitan areas serve as primary sampling units that are selected with some probability, then in a state with two large metropolitan areas, it is possible for both, just one, or even none of the metropolitan areas to be selected. The national-level survey weights provided with the data and calculated by the original investigators are not designed to–indeed cannot–correct for such state-specific biases.

Cummulatively, these three challenges have spurred a variety of innovative approaches to small area measurement of public opinion. The earliest efforts to simulate opinion based solely on demographics (Pool, Abelson, and Popkin 1965; Weber, Hopkins, Mezey and Munger 1972) enjoyed very limited success and two decades would pass before major advances on two fronts. In the time series tradition, Stimson's measurement of policy mood (1991) overcame the problem of incommensurate wording by extracting common elements of opinion change from different collections of variables tapping into the same concept (see also Erikson et al 2002). This provided an estimate of latent opinion on broad ideological scales, but these scales had no natural metric (such as percent approving a particular policy). At about the same time, Erikson, Wright and McIver (1993) aggregated thirteen years of media polls in order to generate highly valid and reliable measures of state-level ideology and partisanship. This approach was later extended to individual policy preferences by many scholars including Norrander (2001) and Brace and his colleagues (2002). Aggregation has been used to measure public opinion for smaller geographies, but typically only within a particular state (e.g,, Percival et al. 2009).

2 Current and Emerging Approaches: MrP and Multiple Imputation

More recently, political scientists have utilized a variety of different modeling approaches to achieve more reliable and valid measures using far less data. Increasingly popular is multi-level modeling with imputation and post-stratification (Gelman & Little 1997; Park, Gelman & Bafumi 2006), which is more recently referred to as Multi-level Regression with Post-stratification, or, "MrP" (recent applications include Lax and Phillips 2009a, 2009b;

Pacheco 2008; Berkman and Plutzer 2010, chapter 3). The MrP approach creates small area estimates by synthetic simulation via a three step process.

First, Bayesian multilevel modeling is used on available survey data of public opinion to fit responses to a question of interest as a function of place (typically a US state) and a series of demographic variables. In the second step, fitted values from this model can be calculated for people types that represent the combination of place and demographic variables. Park, Bafumi and Gelman (2006) divide the US into 3,264 types based on 50 states (plus the District of Columbia), 2 sex categories, 4 education categories, 4 age categories and 2 race categories, For example, a college educated, black, man, aged 66+, living in Alabama would constitute one of the types. To account for uncertainty in these estimates, one can sample from the posterior distributions of the multilevel regression parameters to get many alternative estimates for each set of person types (Lax & Phillips typically generate 10,000 different samples) and this constitutes the imputation phase. In the third step, these fitted values can be combined as a weighted average with the weights based on actual Census counts for each person type (post-stratification).

The matrices below show the process schematically: In the left most dataset, observations (rows) are survey respondents, who reside in state s, and have demographic d, and express opinion y. The second dataset creates all possible person-types that is, all possible combinations of s and d, and calculates a predicted opinion, $\hat{y}$, for such a hypothetical person from an analysis in the previous dataset. Weights w are then added to this dataset to reflect the known numbers of such individuals from Census data, and the predicted values weighted together within each state to get state means, $\bar{y}_s$.

The results of such a process have been very impressive. Park et al. (2006) show that they can replicate Erikson, Wright and McIvers analyses based on a fraction of the sample size. Pacheco's simulations show that post-stratification can reduce bias due to factors like state-specific differential non-response in surveys (2008). In previous work, we employed a single imputation variation on this approach and showed that it could be effectively applied to units as small as school districts (Berkman & Plutzer 2005). Our estimates of local preferences for school spending levels not only predicted actual government spending, but property values as well (after controlling for income levels).

3 Small Area Estimation as a Missing Data Problem

We argue that the three-step MrP procedure can be reconceptualized as a very large missing data problem. We can then consider and contrast solutions from that literature, particularly the widely used statistical literature of multiple imputation. Multiple Imputation (MI) is a commonly employed statistical solution to deal with missing data and nonresponse, particularly in survey data. Under the assumption that any unobserved latent value can be predicted from the relationships present in the partially observed data (known as the Missing at Random regularity condition), MI models correctly combine all the partially observed information through iterative techniques such as Markov Chain Monte Carlo or Expectation Maximization algorithms, into one coherent set of sufficient statistics which depict all the

$$
\left(\begin{array}{rrrr}
obs & s & d & y \\
1 & 1 & 0 & 3 \\
2 & 1 & 1 & 1 \\
3 & 1 & 1 & 2 \\
4 & 2 & 0 & 3 \\
\vdots & \vdots & \vdots & \\
n\text{–}m\text{–}2 & k & 0 & 2 \\
n\text{–}m\text{–}1 & k & 1 & 1 \\
n\text{–}m & k & 1 & 3
\end{array}\right)
\Rightarrow
\left(\begin{array}{rrrr}
obs & s & d & \hat{y} \\
1 & 1 & 0 & 2.7 \\
2 & 1 & 1 & 1.9 \\
3 & 2 & 0 & 1.2 \\
4 & 2 & 1 & 1.7 \\
5 & 3 & 0 & 1.5 \\
6 & 3 & 1 & 2.3 \\
\vdots & \vdots & \vdots & \vdots \\
2k\text{–}3 & k\text{–}1 & 0 & 2.1 \\
2k\text{–}2 & k\text{–}1 & 1 & 1.4 \\
2k\text{–}1 & k & 0 & 1.9 \\
2k & k & 1 & 2.6
\end{array}\right)
\Rightarrow
\left(\begin{array}{rrrrl}
obs & s & \hat{y} & w & \\
1 & 1 & 2.7 & 4 & \left.\begin{array}{l}\\\end{array}\right\} \bar{y}_{s=1} = 2.19 \\
2 & 1 & 1.9 & 7 & \\
3 & 2 & 1.2 & 3 & \left.\begin{array}{l}\\\end{array}\right\} \bar{y}_{s=2} = 1.58 \\
4 & 2 & 1.7 & 9 & \\
5 & 3 & 1.5 & 3 & \left.\begin{array}{l}\\\end{array}\right\} \bar{y}_{s=3} = 2.03 \\
6 & 3 & 2.3 & 6 & \\
\vdots & \vdots & \vdots & \vdots & \\
2k\text{–}3 & k\text{–}1 & 2.1 & 2 & \left.\begin{array}{l}\\\end{array}\right\} \bar{y}_{s=k-1} = 1.40 \\
2k\text{–}2 & k\text{–}1 & 1.4 & 8 & \\
2k\text{–}1 & k & 1.9 & 4 & \left.\begin{array}{l}\\\end{array}\right\} \bar{y}_{s=k} = 2.36 \\
2k & k & 2.6 & 7 & \\
 & & \overline{\sum w = N} & &
\end{array}\right)
$$

Table 1: **MRP:** *In the left most dataset, observations (rows) are survey respondents, who reside in state s and have demographic d, and express opinion y. The second dataset creates all possible* person-types *that is, all possible combinations of s and d, and calculates a predicted opinion $\hat{y}$ for such a hypothetical person from an analysis in the previous dataset. Weights are then added to this dataset to reflect the known numbers of such individuals from Census data, and the predicted values weighted together within each state to get state means, $\bar{y}_s$.*

possible relationships between all the variables in the dataset. These estimated relationships can then be used to make predicted simulations of the missing values in the dataset.

The basic setup for the MI approach to Small Area Estimation is illustrated in the second set of matrices. In the left most matrix, the same survey data as originally used in MRP is stacked together with the previously unused incomplete survey observations and all the individuals in the Census to form one large dataset with missing values (denoted NA). The first $(n - m)$ observations are survey respondents for whom all observations are present, the next m observations are survey respondents missing a demographic variable, and the next N observations are Census participants for whom we do not have survey responses (obtained, for example, from person-level records in the 5% public use micro sample, hereafter PUMS). This incomplete dataset is multiply imputed using missing data algorithms to fill in (multiple times, although for simplicity only once in the figure) all the missing values with a distribution that reflects both the best guess and uncertainty in that missing value. With opinion values filled in for all Census participants, the mean response in any state (or lesser geography) can be calculated by averaging the imputed opinion response for all Census takers in that geography.

From the MI perspective, some observations in the small area problem are complete–those from the original survey where we know opinion, place and demographics. Other

		obs	s	d	y
complete survey data		1	1	0	3
		2	1	1	1
		3	1	1	2
		4	2	0	3
		⋮	⋮	⋮	⋮
incomplete survey data		$n-m-2$	k	0	2
		$n-m-1$	k	1	1
		$n-m$	k	1	3
		$n-m+1$	1	NA	2
		⋮	⋮	⋮	⋮
		$n-1$	k	NA	1
		n	k	NA	3
Census data		$n+1$	1	0	NA
		$n+2$	1	1	NA
		$n+3$	1	1	NA
		$n+4$	1	1	NA
		$n+5$	2	0	NA
		⋮	⋮	⋮	⋮
		$n+N-4$	$k-1$	1	NA
		$n+N-3$	k	0	NA
		$n+N-2$	k	1	NA
		$n+N-1$	k	1	NA
		$n+N$	k	1	NA

$\Rightarrow$

obs	s	d	y	
1	1	0	3	
2	1	1	1	
3	1	1	2	
4	2	0	3	
⋮	⋮	⋮	⋮	
$n-m-2$	k	0	2	
$n-m-1$	k	1	1	
$n-m$	k	1	3	
$n-m+1$	1	0.2	2	
⋮	⋮	⋮	⋮	
$n-1$	k	0.9	1	
n	k	0.8	3	
$n+1$	1	0	1.6	
$n+2$	1	1	2.1	$\bar{y}_{s=1} = 1.98$
$n+3$	1	1	2.0	
$n+4$	1	1	2.2	
$n+5$	2	0	1.2	
⋮	⋮	⋮	⋮	
$n+N-4$	$k-1$	1	1.3	
$n+N-3$	k	0	1.9	
$n+N-2$	k	1	2.3	$\bar{y}_{s=k} = 2.05$
$n+N-1$	k	1	2.1	
$n+N$	k	1	2.2	

Table 2: **MI:** *In the left most dataset the same survey data as originally used in MrP is stacked together with the previously unused incomplete survey observations and all the individuals in the Census to form one large dataset with missing values (denoted NA). The first $(n-m)$ observations are survey respondents for whom all observations are present, the next m observations are survey respondents missing a demographic variable, and the next N observations are Census participants for whom we do not have survey responses. This incomplete dataset is multiply imputed using missing data algorithms to fill in (multiple times, although here for simplicity only once) all the missing values with a distribution that reflects both the best guess and uncertainty in that missing value. With opinion values filled in for all Census participants, the mean response in any state (or lesser geography) can be calculated by averaging the imputed opinion response for all Census takers in that geography.*

observations–those from the Census–are incomplete, as we only know the place and demographic variables, but not opinion. The multiple imputation approach is to draw simulations of all the missing values that complicate the analysis, so that the analyst can ignore the missing data problem and analyze the fundamental model of interest as if no missing data were present. Thus multiple imputation would result in multiple versions of the Census data where all of the demographic variables remain as collected by the Census, and all the missing opinion data are drawn from a distribution that reflects the best guess and uncertainty in how that individual in the Census would have answered the opinion survey.

By reconceptualizing small area estimation as a missing data problem all three steps in MrP can be unified into one estimator, allowing us to consider all the algorithms for multiple imputation as interchangeable or alternate rivals for the multilevel hierarchical approach. Most importantly it brings the literature and results (particularly the well understood regularity conditions) from the missing data literature to bear on the problem of small area estimation. Multiple imputation approaches are specifically engineered to be flexible to large numbers of patterns of missingness, and can easily deal with partially incomplete data, distributed across questions. Thus one advantage of these algorithms is that we can avoid selection bias from only using fully observed observations in the original survey (demographic variables are generally complete in more expensive face-to-face surveys but often incomplete in telephone surveys). This also allows the researcher to use multiple opinion questions or combine different opinion questions across different surveys. However, in standard multiple imputation algorithms, this flexibility comes at the price of extra modeling assumptions such as multivariate normality and linearity.

4 Empirical Demonstration

The example we employ in this paper concerns the teaching of evolution in public schools. Previously, we have demonstrated that even though each state provides content standards regarding the teaching of evolution, the implementation of those standards varies considerably within each state (Berkman and Plutzer 2010). Presumably, much of this intra-state variation is due to the influence of local public opinion, so developing valid and reliable measures of local sentiment can help specify this relationship, allowing more detailed testing of the mechanisms that enhance or diminish policy responsiveness.

In addition to having some substantive import, the available polling data are appropriate for us to demonstrate the use of the MI approach and compare it with results obtained by MrP – the current state of the art. Doing so requires assembling a data set that combines several available opinion polls, and assembling a data set of individual Census records. The data sets will be combined to enable the MI estimation of local opinion; for MrP, the Census records will serve as the basis for post-stratification weights.

4.1 The Polling Data

We utilized data from every national poll or academic survey that met three conditions: (1) the survey contained a standard question asking specifically about teaching evolution, (2) the survey recorded the state of residence of each respondent, and (3) the original data records were available to us to analyze. In total, we were able to utilize nine different studies from 1998 through 2005 that included 11,262 respondents. These included three polls conducted by the Pew Center for People and the Press in 2005 and 2006, the 2005 Virginia Commonwealth University Life Sciences Survey, a 1998 Southern Focus Poll conducted by the University of North Carolina, and a 2004 CBS/New York Times poll. The details of differences in question wording are detailed elsewhere (Appendix 2 in Berkman and Plutzer 2010) and outlined in the appendix to this paper.

	1998	1999	2004	2005	2006
Pew Relig	0	0	0	2000	2003
Gallup	0	1016	0	0	0
CBS NYT	0	0	885	0	0
UNC	1257	0	0	0	0
VCU	0	0	0	1002	0
Newsweek	0	0	1009	0	0
Harris	0	0	0	1000	0
Pew Press	0	0	0	1090	0

Table 3: *Sample sizes and dates across polls: There are 11262 total individual responses across nine surveys to some question format measuring whether evolution should be taught in high school.*

For each question (or question pair, as in the case of the Pew and CBS polls), we coded respondents 1 if they rejected all forms of creationism, including Intelligent Design (either along with evolution or instead of evolution) and 0 for all other combinations of responses, yielding a measure that connotes support for teaching evolution and only evolution. Differences in question format are coded and modeled so that estimated probabilities can be expressed in terms of any particular question format. We have typically followed the convention of expressing estimated probabilities based on the University of North Carolina single question about teaching creationism along with evolution. Across many other polls, this position is rejected by about 30% of the US public and this becomes the scale of reference for all analyses in this paper. Other scales can be used with the support for teaching evolution only being lower for all other commonly employed questions (see Berkman and Plutzer 2010, chapter 2).

We merged these surveys to the lowest common aggregation level of question format. The demographics that are collected are broadly similar across polls. Thus we had to collapse education and age to four common categories across surveys, as well as collecting gender and racial identifiers for Black, Asian and Hispanic respondents, and a dummy variable for

marital status. Income was collected in all surveys, but on different scales. We collapsed income to four common categories (0-30k,30-50k,50-75k,75k+), as well as identifying two supplemental dummy variables for very low income respondents (below 15k) and very high income respondents (above 100k). In surveys where the scaling was less precisely measured, one or both of these latter dummy variables could not be determined for those respondents and thus were missing data.

5 State Level Inference

To examine the feasibility of re-conceptualizing small area estimation as one large missing data problem, and to demonstrate our multiple imputation approach, we first examine a baseline model. We first estimated state means of support for teaching evolution using MrP. Similarly, we treated the combination of the Census and polling data as one large incomplete dataset, imputed the evolution opinions for the observations originally from the Census and averaged these by state. We intentionally used only the 9321 complete observations in the polling data (82.7% of the original total observations across the polls) and treated all of the allocated values in the Census as if they were real observed values, so that both methods used exactly the same set of observations.

The estimates of mean support for teaching evolution (and only evolution) in each state are shown in the figure below. The sizes of the plotted points are proportional to the number of complete observations (listwise n) from that particular state in the original polling data. The two sets of estimates line up nicely, indicating that our MI method recovers very similar estimates in this scale of dataset when using the same observations as the more standard approach of MrP . However, it can be seen that small states (small circles with fewer opinion responses) are slightly shrunk towards the national mean (quite dramatically for Hawaii and Rhode Island) in the MrP estimates, whereas the larger states line up in their estimated positions very directly between the two methods. From this we conclude that the added hierarchical structure of the MrP model is not changing the model estimates significantly from the simpler multivariate normal assumptions of the MI model.

5.1 Incorporation of partially observed observations

The imputation method easily and appropriately handles missing values from both the polling data and the Census, and thus can use more of the valuable polling information available. The fraction of observations which are only partially complete in the model, depends on the response rates, but also the set of variables included in the model. If incomplete observations need to be dropped from the dataset, as in the MrP model, there is a balance between adding additional variables to improve the fit of the model, and the resulting decrease in the sample size from missingness in those additional variables. In figure 3 we start from a very simple model of support for teaching evolution as a function of question wording and state of residence, and sequentially add in additional control variables, until we arrive at a full model using all available demographics. However, in these models, we now treat

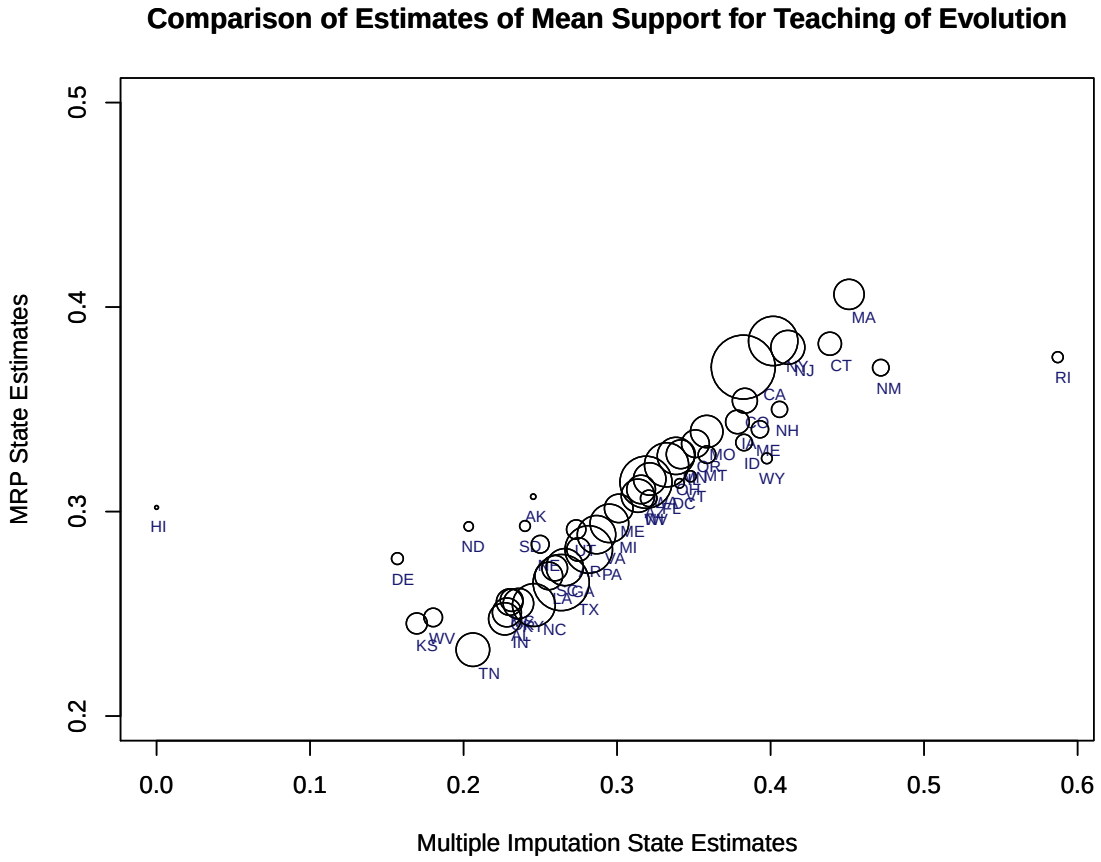


Figure 1: *Comparison of estimated state level support for teaching evolution, between the MI and MRP models. The area of each circle is proportional to the number of observations present in the combined polling data. With 9321 polling observations, the two models estimates line up strongly, except in small states with very small observations where the MRP estimates show shrinkage to the national mean.*

all *allocated* Census data as missing values, and allow the multiple imputation model to use the partially complete observations. Figure 2 shows the patterns of missingness across the stacked polling dataset, and 50000 samples from the Census, with clearly the absence of the evolution question on the Census being the largest block of missingness.

Figure 2: *The patterns of missingness in the stacked polling and Census data.*

Each graph in figure 3 plots the state level estimates from the simplest model–including only a state-level effect, and dummy for question wording–along the x-axis against the state

level estimates from a more complicated model that includes additional demographic covariates. From left to right, the models include more variables, starting with education and age, adding in gender and marital dummies, racial characteristics, and finally income. In the top row we show how these models behave for the MI approach. Increasing covariates does not greatly change the estimated state level means. Obviously the estimated parameters are changing across these models, but the final estimated fraction of state supporting teaching evolution remains steady. In the model on the x-axis these fractions are entirely predicted by state-level fixed effects, while as the models grow more complicated, individual level covariates are substituting in explanatory power. However, across the top row, for the MI models, the estimates of mean state support remain steady regardless of the set of included covariates.

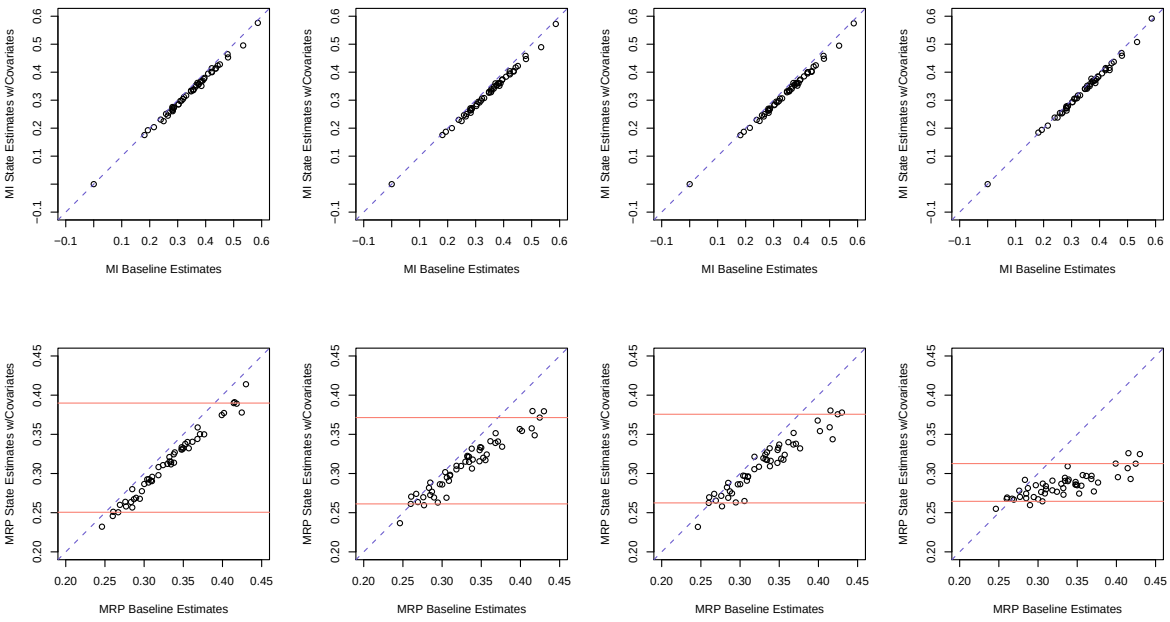
In the bottom row we see this comparison for the MrP model. Here, as we include more covariates, the number of complete observations –both in the polling data and the Census– drop and the estimated state means change dramatically. The red bands represent intervals which contain 90 percent of the state level means, and we can see that as we include more covariates this interval collapses. The states are increasingly estimated to be alike. As the sample size gets smaller, the multilevel shrinkage on the state-level parameters shrinks all the state estimates to a narrow band. Increasing the available covariates should increase the fit of the model, but the small sample estimates are collapsing to the national mean.

6 County Level Inference

We now arrive at our chief objective, and investigate extending the previous analysis to the county level. There are 3139 counties in the United States. As before, we estimate both the MrP and our MI approaches to create estimates of the fraction supporting teaching only evolution across these small areas.

To preserve confidentiality, individuals in the Census data are geographically located by "public use microdata areas" or PUMAs, which are contiguous areas of at least one hundred thousand individuals. Many PUMAs are entirely located within counties, in which case we immediately know the county of the Census respondent. For individuals in PUMAs that cross multiple counties we obtained the fraction of individuals in the PUMA who belonged to each county, and used these fractions to randomly assign individuals to counties by the appropriate multinomial distribution. After mapping every Census respondent in the 5% microsample to a county, the median county has 1150 individuals from which to aggregate predicted opinion. The largest of these, Los Angeles county, has 330000, while the smallest, Harding, South Dakota, has 28 individuals. Overall, 97 percent of counties contain more than a hundred PUMS Census individuals, while 76 percent have more than 500 respondents.

We treated all Census allocated data as observed data, which aids the MrP approach, but in order to see variation *within* states, included all demographic covariates, thus reducing the amount of polling data available to MrP. The estimated rates for each county are plotted in figure 4. As can be seen –and following the intuitions carried over from figure 3– the MrP estimates are collapsed into a narrow band across all counties. The MI estimates have



	State-level effects, Question wording effect , Gender, Education.	State-level effects, Question wording effect , Gender,Education, Age, Marital Status.	State-level effects, Question wording effect , Gender, Education, Age, Marital Status, Racial dummies.	State-level effects, Question wording effect , Gender, Education, Age, Marital Status, Racial dummies, Income.
	Poll total obs.			
n,	11,626	11,626	11,626	11,626
	Poll complete obs.			
n,	9346	7418	7320	5059

Figure 3: *Model stability in MI and MRP as an increasing function of number of included demographic covariates.*

a larger range. The state-level fixed effects have a large effect, even after accounting for demographics. This can be seen in this graph as clusters of diagonal bands, each of which represent counties across a state that have varying demographics.

The continued importance of state-level fixed effects can be seen more directly by examining figures 5(a) and 5(b). Here we show choropleth maps of the estimated support for evolution for the two modeling strategies. The blue and violet end of the rainbow represent high support for teaching evolution and the red end of the spectrum the lowest levels of support. In the MI estimates in figure 5(a) we see substantial variation across states, and some moderate variation across counties within states. In the MrP estimates below, whose colors are mapped on the same scale, we see all estimates compressed within the same range.

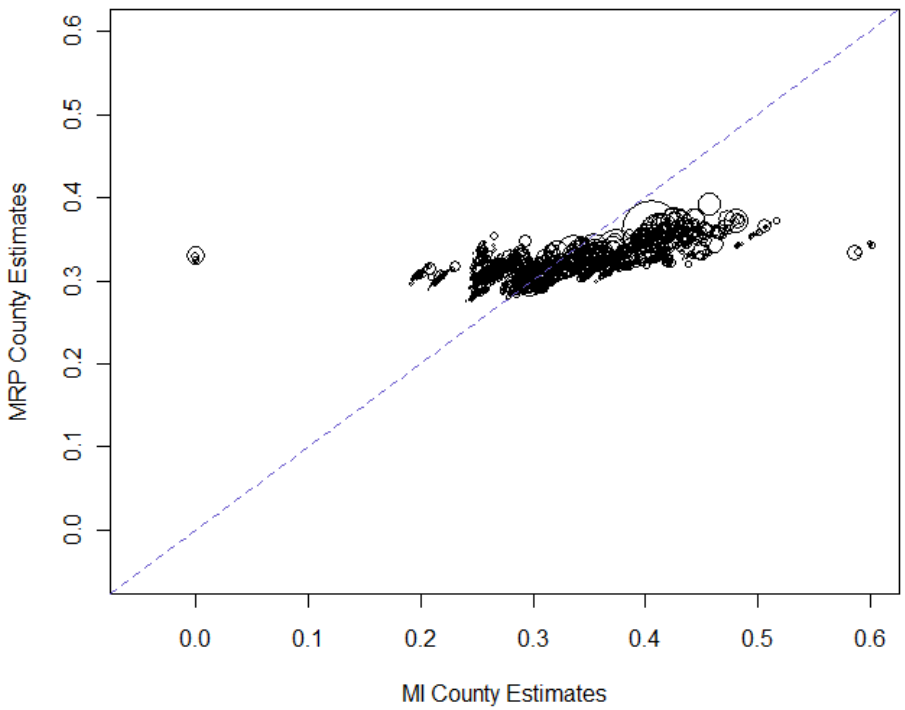
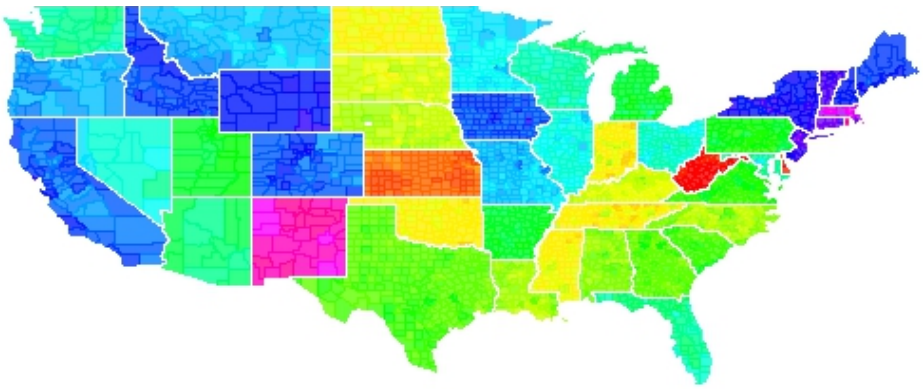


Figure 4: *Comparison of county level small area estimates of support for teaching evolution from* MrP *and* MI *models.*

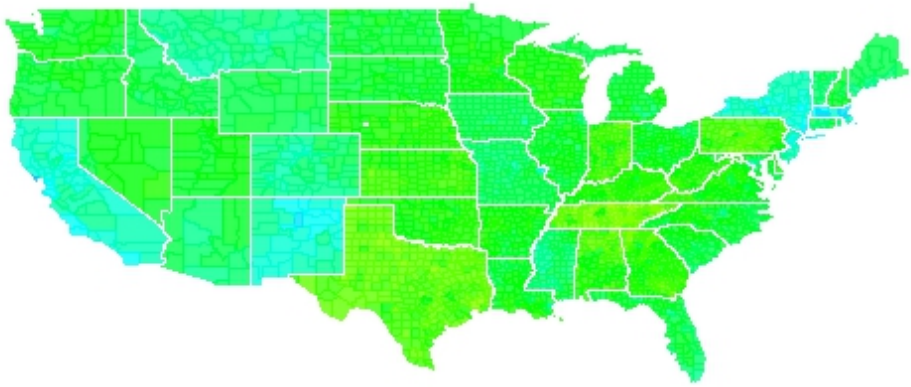
Kansas is a state with very low support for evolution in the MI estimates, while California and New York have among the highest relative levels of support in both maps.

6.1 Validation with non-opinion data

As a second illustration of how these approaches perform for county-level estimation, we also examine small area estimates of the percentage of adults who are currently divorced. This is a percentage that can be estimated by polls, varies by geography, and can be verified using Census data. While the Census asks individuals whether they are currently divorced, we omit this data from all our estimations, and measure the divorce status of individuals solely in the polling data, to mimic the MrP and MI procedures previously used for modeling attitudes towards teaching evolution. The Census measurements of divorce in the PUMS data are treated as validation data to compare the model estimates to. In figure 6 below, we compare these validation values aggregated by county, to the county level predictions generated by MI and MrP. Across the various polls, the mean level of divorce is between 10



(a) *MI estimates of Preferences for Teaching Evolution*



(b) *MrP estimates of Preference for Teaching Evolution*

Figure 5: Choropleth maps of estimated preferences.

and 18 percent, while it is 8 percent in the 10 million PUMS observations. We see that both sets of model estimates are systematically positively biased in estimating the rate of divorce.

Table 4 shows the root mean squared error across all county estimates, and we can see that error for the MI model is 11 percent lower than for the MrP model (or 9.1 percent lower when these errors are weighted by county population). The left most graph in figure 9 plots the absolute error for each MI county estimate against each MrP county estimate. We can see that the reduction in mean squared error in this example is because the majority of counties have more error in the MrP estimates than the MI estimates (that is, are above the $y = x$

	Root Mean Squared Error		Pearson Correlation	
	MI estimates	MrP estimates	MI estimates	MrP estimates
County Divorce Rates Using Polling Data	0.0742	0.0830	.689	.598
County Divorce Rates Population Weighted	0.0738	0.0812	.715	.668
County Divorce Rates Using Census Subsample	0.0120	0.0120	.804	.804

Table 4: Root mean squared error in estimating divorce data, using polling data and treating a subsample of the Census as if it were collected by a poll.

diagonal), and this is not simply driven by poor estimation of a small set of observations or a few outliers. The error in both model estimations, however, seems largely driven in this example by differences in the sampling frame and exact question wordings between the polling data and the Census data. To demonstrate this with a synthetic example, we subsample 10000 observations from the Census, and treat them *as if* these were the polling data, and treating the rest of the Census observations as the Census data. In this simulation of a polling universe, both models have very small error as seen in the right most figure in 9 and table 4.

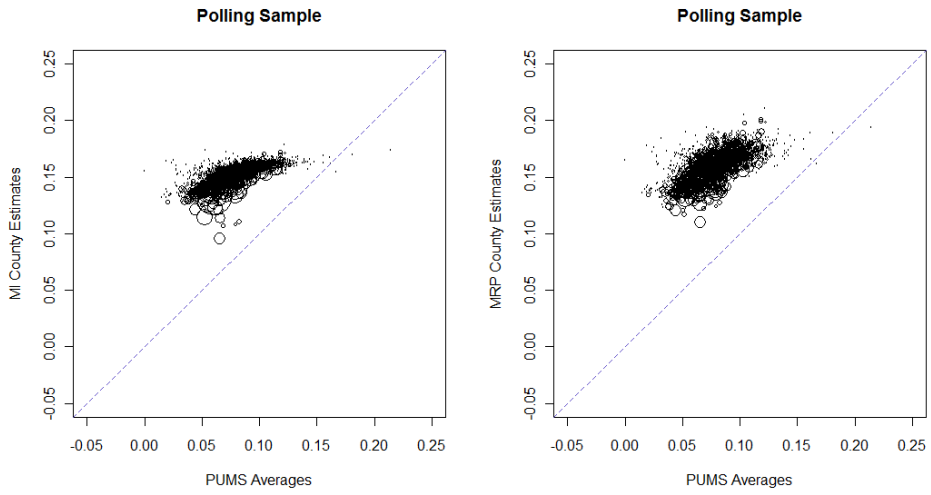


Figure 6: *Actual levels of divorce from PUMS data versus model estimates.*

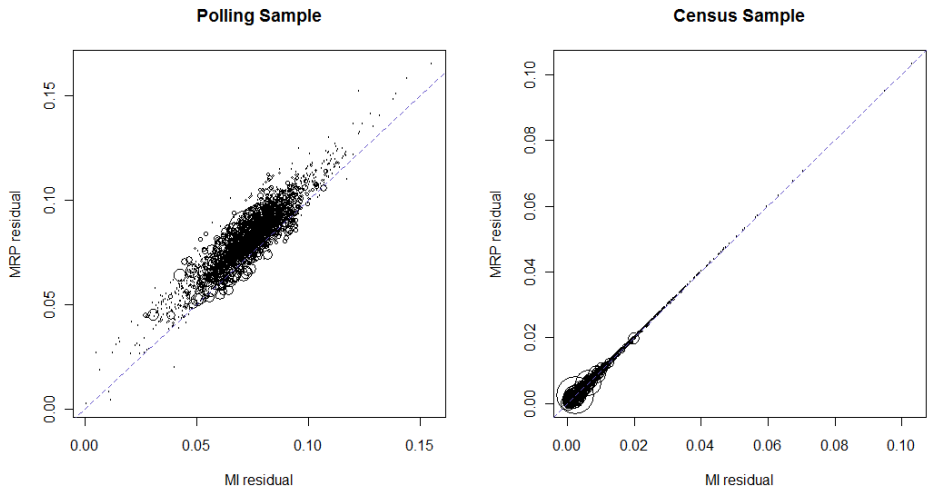


Figure 7: *The absolute error in each county estimate from each model.*

7 Use of Estimates in Second Stage Models

While we have introduced a number of individual level covariates into these models to forecast individual opinion, it may commonly be that there exists important information aggregated to the local level geography that can help to predict the aggregate small area quantities that our estimation strategies generate. How to appropriately include such variables has been both a pragmatic and theoretical question in the small area literature.

We could simply include geographically aggregated variables into the individual level model, if we believed the variables had a context effect, such that the average value in the geography or community leads to individuals themselves engaging in different behavior. If however, we have geographically aggregated variables that do not effect our quantity of interest through the individuals, but rather are helpful forecasts of the small area aggregates we are attempting to predict, then we should introduce these variables at the appropriate level of the model.

Succinctly, individuals have individual level variables, and thus individual level forecasts of behavior. The aggregate of all this individual level behavior leads to the estimated small area quantity. However, if we have additional aggregated variables that are also good forecasters of the aggregated small area quantity, we want them to help refine the aggregate small area estimates, even though we can not appropriately introduce them into the individual level model.

The small area estimates –because they are estimates– have associated uncertainty. This uncertainty reflects both that we do not know (but rather estimate) the parameters that predict behavior from individual level covariates, and also that the estimates of the population means are generated by samples from those populations. The combined uncertainty

leads to estimates that have measurement error from the true values.

In recent work we have shown that measurement error can be directly addressed within the multiple imputation framework (Blackwell, Honaker and King, Forthcoming a,b). This approach successfully treats many forms of measurement error in data, and robustly corrects their bias. In this setting, mismeasured observations are treated as missing values of the latent or true data. The observed mismeasured or proxy variable gives us prior information about where the true value is located, and so is used as a prior. The mean of the prior is value of the proxy variable, and the variance is a direct function of the estimated measurement error. The mismeasured data is *overimputed* that is, replaced or overwritten with multiple draws reflecting the posterior of the true latent data, which is what the analyst would have used if available to avoid the measurement problem. In our small area application, the estimated small areas are mismeasures of the latent data, and any other aggregated covariates can be introduced as predictors in a second imputation model where the unit of observation is the county. What we need is to construct priors on the value of the true latent small area means, from the mismeasured estimates generated in the last imputation model at the individual level.

In Honaker and King (2010) we show how to implement priors on individual missing cell-values in an incomplete matrix, within an EM or MCMC algorithm. These priors are generally normal distributions with some given mean and variance. The prior mean reflects the best guess of missing value, and the variance the strength of that belief.[1]

Calculations of total uncertainty in a quantity of interest from a multiple imputation procedure are well known (for discussions see, Schafer 1997, King et al 2001). In the example here, if q_{sj} is the estimated small area quantity of interest in geography s from the jth imputed dataset, the standard errors of the estimated small area averages, generated by the described multiple imputation procedure in the previous sections are:

$$\mathrm{SE}(q_s)^2 = \frac{1}{m}\sum_{j=1}^{m}\mathrm{SE}(q_{sj})^2 + S_q^2\,(1+1/m)\,. \tag{1}$$

where $\mathrm{SE}(q_{sj}) \approx \sqrt{q_{sj}(1-q_{sj})/n_s}$ is the uncertainty in any one small area estimate from one imputed dataset, driven by sampling size and population variance, and the second term:

$$S_q^2 = \sum_{j=1}^{m}(q_{sj} - \bar{q_s})^2/(m-1) \tag{2}$$

reflects the disagreement across imputed datasets driven by estimation uncertainty in the parameters of the imputation model.

[1] Because the normal distribution is itself conjugate normal, both the contribution of the cell value to the sufficient statistics, and the posterior distribution for that cell value from which imputations might be drawn are the product of the normal prior and the observed data posterior. In the limit, as the prior variance collapses to zero, the cell contributes to the sufficient statistics of the dataset as if it were observed at the prior mean, and the imputations that are generated for that cell shrink to the prior mean. As the prior variance becomes increasingly large, the cell contributes very little to the sufficient statistics, and in the limit behaves as any other missing observation with no prior information.

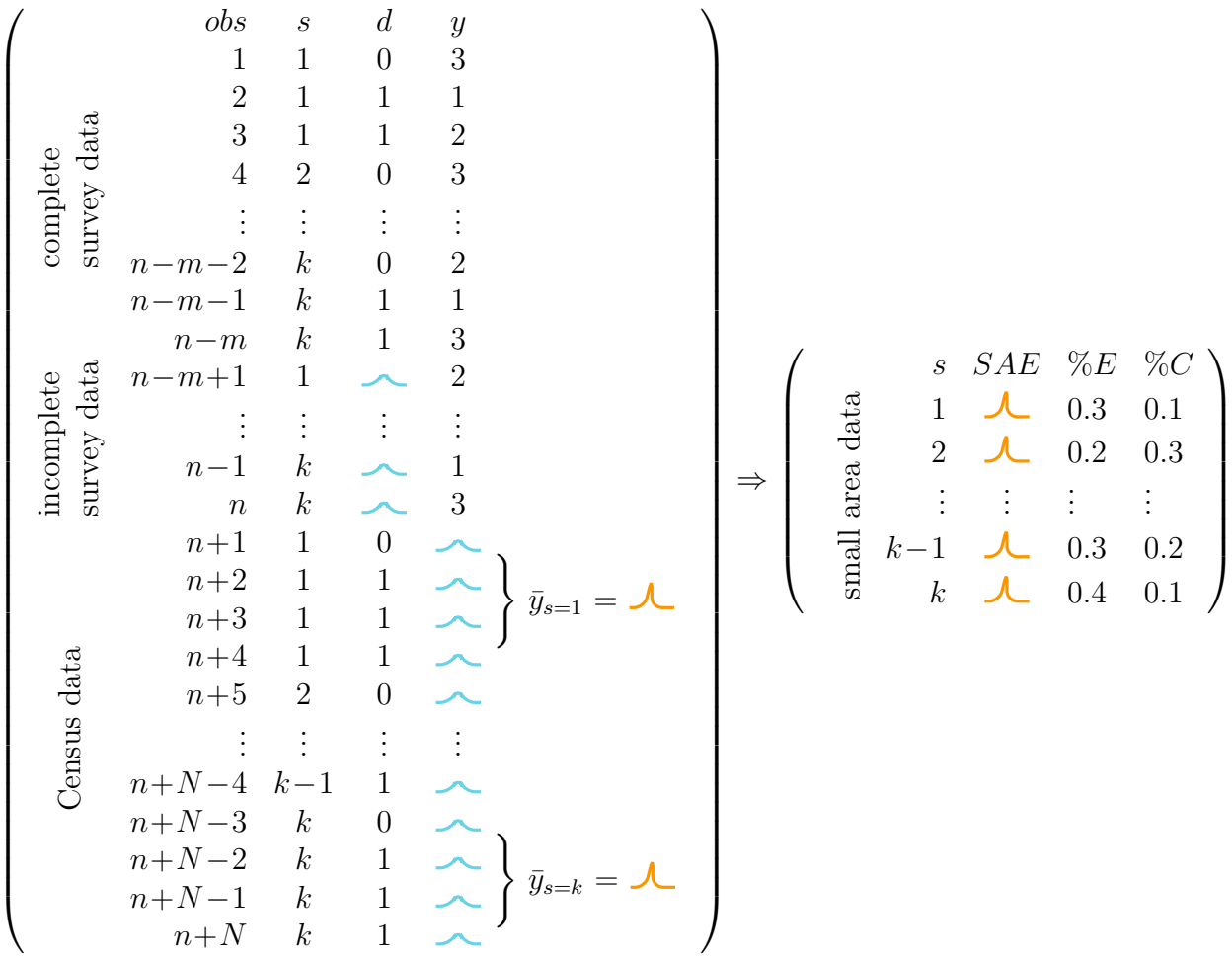


Table 5: *The stacked dataset on the left imputes all missing values of individual level data to form a* ***distribution of imputed values*** *as in Table 2. Aggregating the imputed survey responses, creates a distribution of county level* ***small area estimates*** *. These normal distributions form priors for the location of the latent true county average, which are imputed by Multiple Overimpuation in the dataset on the right, where the unit of observation is the county, and which includes county level variables as covariates in the Overimputation model. Here, the county level covariates %E and %C are the religious adherence rates in the county of Evangelicals and Catholics.*

We use $\text{SE}(q_s)^2$ to set the normal prior for any estimated small area quantity as $\mathcal{N}(q_s, \text{SE}(q_s)^2)$. We then include county level variables that forecast the small area quantity in this imputation model and generate new imputations of the latent level of small area quantity of interest. In our example, we have county level measures of the rate of advanced degrees in the county and the adherence rates of evangelical Christians and of Catholics, which we believe are predictive of the level of support for teaching evolution in that county. An overview of the creation of the multiple overimputation model, and the two levels of imputation, is given in table 5.

7.1 Replication of Local Level Policy Implementation Study

In an analysis of instruction by high school biology teachers, Plutzer and Berkman (forthcoming) examine the degree to which behavior in the classroom is driven by the competing pressures of state level curricular standards and local community sentiment and pressure. They construct a scale of five survey items of the "competing goals for science instruction" among a survey of high school biology teachers, that measures the practices, importance and centrality of teaching evolution in a high school biology curricula *vis-à-vis* teaching creationism. They predict this scale by a measure of state curricula standards, and local level community attitudes towards teaching evolution.

We replicate the simplest analysis they present using the different measures of local community attitudes towards teaching evolution studied here, that is, those estimated with MrP the simplest MI approach, and the two level MO model using additional aggregated covariates. There are 907 surveyed biology teachers, and their community sentiment towards teaching evolution is measured by using the small area estimate of support for teaching evolution in the county of the high school in which they teach. We replicate table 1, column 1 of Plutzer and Berkman with our three different small area estimates, and present the results in table 6. We see small improvements in the performance of the small area measures, as measured by their t-values as we move from the MrP to the MI estimates. We see additional improvement in fit in predicting behavior when we additionally use the aggregated county advanced degree and religious adherence rates in the MO model to overimpute the mismeasured estimates of support for evolution.

8 Conclusion

Measuring attitudes, preferences and behavior in small local geographies is central to studies of representation. However national random samples require additional auxiliary information and an often complex estimation strategy to construct these measures. By using individual level demographic variables from Census microsamples, we conceptualize the "small area" estimation problem as a problem of missing data, which can be estimated within the multiple imputation framework–commonly already employed in these surveys for more straightforward issues of item non-response. Similarly, using aggregate level variables to improve forecasts of aggregate small area quantities of interest can be implemented in the multiple

	β (se)	**t**	β (se)	**t**	β (se)	**t**
Local support for evolution	2.39	**3.12**				
Estimated with MRP	(.767)					
Local support for evolution			.819	**3.83**		
Estimated with MI			(.214)			
Local support for evolution					1.07	**4.52**
Estimated with MO					(.236)	
Rigor of state standards	.038	2.49	.038	2.83	.036	2.67
	(.015)		(.013)		(.013)	
Constant	-.371		.125		.040	
	(.248)		(.074)		(.082)	
n	907		907		907	

Table 6: Estimates of the effects of local level community sentiment and state curricula standards on teacher behavior in instruction on evolution in high school biology classes. The three measures have slightly different scales, especially the MrP estimates which we have seen are much more collapsed in range. However, the t-values improve from MRP to MI to MO suggesting improved fit. (Estimated random effects for states omitted.)

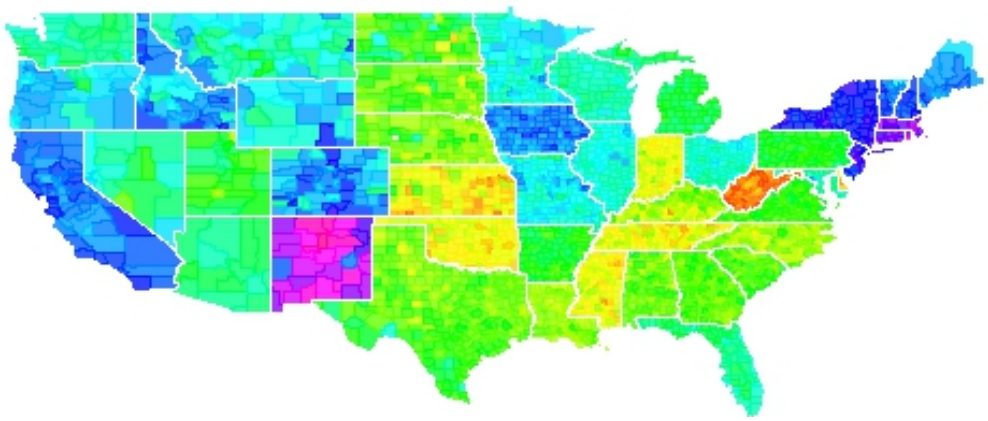


Figure 8: *Estimated county level support for teaching evolution after the MO missing data/measurement error model.*

overimputation framework, a generalization of imputation that incorporates and treats for measurement error in variables. Imputation approaches perform comparably to current best

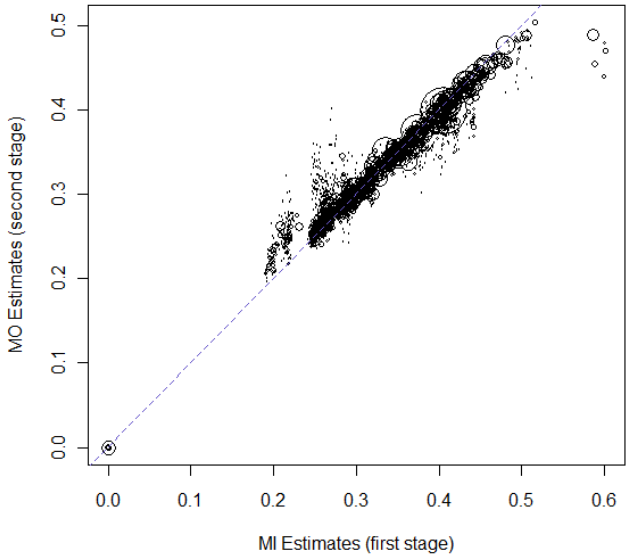


Figure 9: *Initial estimates after imputation (*MI*), and second stage imputation estimates after correcting for measurement error (*MO*). The largest changes in the estimates are for the smallest counties, which would have the most measurement error.*

practice such as MrP , with small improvements seen in a forecasting example with validation data, improvements in the performance of a measure in a replication, and the ability to seamlessly handle missing data problems in the survey data at the same time as generating the small area estimates.

References

Bartells, Larry. 2008. *Unequal Democracy: The Political Economy of the New Gilded Age*. Princeton University Press.

Berkman, Michael B. and Eric Plutzer. 2009. Public Opinion and the Teaching of Evolution in the American States. *Perspectives on Politics* 7 (September): 485-500.

Berkman, Michael B. and Eric Plutzer. 2005. *Ten Thousand Democracies: Politics and Public Opinion in Americas School Districts*. Washington D.C.: Georgetown University Press.

Berkman, Michael B. and Eric Plutzer. (forthcoming) "Local Autonomy vs. State Constraints: Balancing Evolution and Creationism in U.S. High Schools." *Publius*.

Berry, William D., Evan J. Ringquist, Richard C. Fording, and Russell L. Hanson. 1998. "Measuring Citizen and Government Ideology in the American States, 1960-93." *Amer-*

ican Journal of Political Science 42(1): 337-348.

Blackwell, Matthew, James Honaker and Gary King. *Forthcoming a.* "A Unified Approach to Measurement Error and Missing Data: Overview and Applications" *Sociological Methods and Research* Copy at http://hona.kr/papers/files/measure.pdf

Blackwell, Matthew, James Honaker and Gary King. *Forthcoming b.* "A Unified Approach to Measurement Error and Missing Data: Details and Extensions" *Sociological Methods and Research* Copy at http://hona.kr/papers/files/measured.pdf

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

Brooks, Clem, and Jeff Manza. 2006. "Social Policy Responsiveness in Developed Democracies." *American Sociological Review* 71 (3): 47494.

Burstein, Paul. 2003. "The Impact of Public Opinion on Public Policy: A Review and an Agenda." *Political Research Quarterly* 56: 29-40.

Erikson, Robert S. 1978. "Constituency Opinion and Congressional Behavior: A Reexamination of the Miller-Stokes Representation Data." *American Journal of Political Science* 22(3): 511-535.

Erikson, Robert S., Michael B. MacKuen, and James A. Stimson. 2002. *The Macro Polity.* New York: Cambridge University Press.

Erikson, Robert S., Gerald C. Wright, John P. McIver. 1993. *Statehouse Democracy: Public Opinion, Politics and Policy in the American States.* New York: Cambridge University Press.

Fording, Richard C., Joe Soss and Sanford F. Schram. 2007. "Devolution, Discretion, and the Effect of Local Political Values on TANF Sanctioning." *Social Service Review* 81(2): 285-316.

Gelman, Andrew and Thomas C. Little. 1997. "Poststratification Into Many Categories Using Hierarchical Logistic Regression." *Survey Methodologist* 23(December): 127-135.

Gilens, Martin. 2005. "Inequality and Democratic Responsiveness." *Public Opinion Quarterly* 69 (5): 778-796.

Honaker, James and Gary King. 2010. "What to Do about Missing Values in Time-Series Cross-Section Data" *American Journal of Political Science* 54(2):561-581.

Karol, David. 2007. "Has Polling Enhanced Representation? Unearthing Evidence from the Literary Digest Issue Polls." *Studies in American Political Development* 21: 16-29. doi:10.1017/S0898588X07000144

Keiser, Lael. Forthcoming. "Understanding Street-Level Bureaucratic Decision Making: Determining Eligibility in the Social Security Disability Program." *Public Administration Review.*

Keiser, Lael R. 1999. "State Bureaucratic Discretion and the Administration of Social Welfare Programs: The Case of Social Security Disability." *Journal of Public Administration Research and Theory* 9 (1): 87-106.

Keiser, Lael R. and Joe Soss. 1998. "With Good Cause: Bureaucratic Discretion and the Politics of Child Support Enforcement." *American Journal of Political Science* 42 (4): 1133-1156.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" *American Political Science Review* 95(1):49-69.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53(1): 107-21.

Maestas, Cherie. 2000. "Professional Legislatures and Ambitious Politicians: Policy Responsiveness of State Institutions." *Legislative Studies Quarterly* 25(4): 663-690.

Manza, Jeff and Fay Lomax Cook. 2002. "A Democratic Polity? Three Views of Policy Responsiveness to Public Opinion in the United States." *American Politics Research* 30 (6): 630-667.

Miller, Warren E. and Donald Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57(1): 45-56.

Norrander, Barbara. 2000. "The Multi-Layered Impact of Public Opinion on Capital Punishment Implementation in the American States." *Political Research Quarterly* 53(4): 771-793.

Norrander, Barbara. 2001. "Measuring State Public Opinion with the Senate National Election Study." *State Politics and Politics Quarterly* 1 (1): 111-125.

Pacheco, Julianna. (Forthcoming) "Using National Surveys to Measure Dynamic State Public Opinion: A Guideline for Scholars and an Application." *State Politics and Policy Quarterly.*

Park, David K., Andrew Gelman, and Joseph Bafumi. 2006. "State-Level Opinions from National Surveys: Poststratification Using Multilevel Logistic Regression."In Jeffrey E. Cohen, ed., *Public Opinion in State Politics.* Stanford, California: Stanford University Press. (209-228)

Percival, Garrick L. 2010. "Ideology, Diversity, and Imprisonment: Considering the Influence of Local Politics on Racial and Ethnic Minority Incarceration Rates." *Social Science Quarterly* 91:1063-1082.

Percival, Garrick L. 2004. "The Influence of Local Contextual Characteristics on the Implementation of a Statewide Voter Initiative: The Case of California's Substance Abuse and Crime Prevention Act (Proposition 36)." *Policy Studies Journal* 32: 589-610.

Percival, Garrick L., Martin Johnson and Max Neiman. 2009. "Representation and Local Policy: Relating County-Level Public Opinion to Policy Outputs." *Political Research Quarterly* 62 (March): 164-177.

Pool, Ithiel de Sola, Robert P. Abelson, and Samuel L. Popkin. 1965. *Candidates, Issues And Strategies: A Computer Simulation Of The 1960 And 1964 Presidential Elections.* Cambridge: MIT Press.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman

and Hall.

Soroka, Stuart, and Christopher Wlezien. 2010. *Degrees of Democracy.* Cambridge University Press.

Weissert, Carol S. 1994. "Beyond the Organization: The Influence of Community and Personal Values on Street-Level Bureaucrats' Responsiveness." *Journal of Public Administration Research and Theory* 4 (2): 225-254.

Appendix

Sources and question wording:

Virginia Commonwealth University Life Sciences Survey (9/14/20059/29/2005) [Data provided directly by Survey and Research Evaluation Laboratory, Virginia Commonwealth University]

Regardless of what you may personally believe about the origin of biological life, which of the following do you believe should be taught in public schools?Evolution onlyevolution says that biological life developed over time from simple substances. Creationism onlycreationism says that biological life was directly created by God in its present form at one point in time. Intelligent design onlyintelligent design says that biological life is so complex that it required a powerful force or intelligent being to help create it. Or some combination of these?

[If "some combination"] *Which approaches do you think should be* taught?

University of North Carolina, Southern Focus Poll (2/4/19982/24/1998) [Data acquired from the Odum Institute, University of North Carolina]

Would you generally favor or oppose teaching creation along with evolution in public schools?

CBS/New York Times Poll (11/18/0411/21/04) [Data acquired from the Roper Center, University of Connecticut]; Pew Typologies Callback Survey (3/17/053/27/05) [Data acquired from the Pew Research Center]; Pew Religion and Public Life Poll (7/7/057/17/05) [Data acquired from the Pew Research Center]; Pew Religion and Public Life Poll (7/6/067/19/06) [Data acquired from the Pew Research Center];

Would you generally favor or oppose teaching creation along with evolution in public schools?

Would you generally favor or oppose teaching creation instead of evolution in public schools?