REWEIGHTING A NATIONAL DATABASE TO IMPROVE THE ACCURACY OF STATE ESTIMATES

Allen L. Schirm, Mathematica Policy Research, and Alan M. Zaslavsky, Harvard University
Allen Schirm, Mathematica Policy Research, 600 Maryland Ave., S.W., Suite 550, Washington, DC 20024-2512

KEY WORDS: Small Area Estimation, Simulation

1. INTRODUCTION

State or substate estimates are often required to inform critical policy decisions or administer important public programs. However, national surveys and even national samples of administrative records can typically support only imprecise direct estimates for states or substate areas because state sample sizes are small. Borrowing strength with an empirical Bayes or similar indirect estimator is a common solution to the problem of imprecise direct estimates, and has been used successfully in many applications. For example, an indirect estimator has been used for several years to derive state estimates for allocating federal funds under the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). Similar estimators have also been used to obtain state and county estimates of poor school-aged children for allocating federal Title I funds for compensatory education in elementary and secondary schools. The difference between these indirect estimators and a direct estimator is that the direct estimator uses data from only the domain of interest—as defined, typically, by geography and time—whereas an indirect estimator uses data from other areas or other time periods to derive an estimate for the domain. Using data from many domains and, often, from multiple data sources enables an indirect estimator to borrow strength and improve the precision of estimates for each domain.

The estimators used in the WIC and Title I applications are suitable for deriving a single estimate or a few closely related estimates for each geographic area. However, they are not suitable for deriving large numbers of estimates—for example, filling in a large table—for each area. The problem is that the modeling undertaken for empirical Bayes or similar estimation is specific to the estimates being produced, and it would not be practical to develop a model for each cell of a large table.

To address this problem, Schirm and Zaslavsky (1997) proposed a method for reweighting a national survey or administrative records database to borrow strength and improve precision. A Poisson regression model is fitted to obtain an estimated prevalence in each state (or other small area) of every household type in the database, where types are defined by households' characteristics. This model is specified to control important aggregates at the state level, and the prevalences are expressed as a matrix of weights, with each household having a weight for every state. After this Poisson regression model is estimated, no further modeling is required. Any estimates sought for a state are obtained using all the households in the database, not just the households actually in that state. By applying the appropriate weight for each household, the database is weighted to look like the state, rather than the whole country. This reweighting method can be combined with empirical Bayes methods by using empirical Bayes estimates as control totals in the reweighting.

Schirm and Zaslavsky (2001) describe two previous applications of reweighting: (1) estimating the numbers and characteristics of children without health insurance and (2) microsimulation of proposed food stamp and welfare program reforms. In both of these applications, the reweighting method was appropriate because of the need to develop large tables of estimates for each state from databases with small state samples.

We describe the reweighting method in detail in the next section. In Section 3, we describe the design of the evaluation that we have conducted to determine, principally, whether indirect estimates from a reweighted database are more accurate than direct estimates. Then, in Section 4, we present results from the evaluation. Throughout this paper, we will assume that our objective is to obtain estimates for states, although the reweighting method could also be used to derive estimates for other subnational areas.

2. REWEIGHTING TO BORROW STRENGTH

2.1 Basic Ideas

The basic idea of the reweighting approach is to use—that is, give weight to—households from many states when deriving estimates for any one state. How reweighting can be used to borrow strength is illustrated by comparing (1) the direct estimator that uses the original sample weights and does not borrow strength with (2) the indirect estimator that uses reweighted data and does borrow strength.

To calculate an estimate for Virginia, for example, the direct estimator uses only the sample households for Virginia and their original sample weights. Observations for other states are ignored. This is equivalent to using all the observations in the database weighted by "Virginia weights" that equal the original sample weights for households in Virginia but are zero for households in all other states.

In contrast, for indirect estimation, nonzero Virginia weights would be assigned to households in not only Virginia but also other states. Taking a specific example, suppose that the only two states in our database are Virginia and Maryland. Suppose also that each state has a sample household with similar income and other characteristics, but the only child in the Virginia household is age 4, while the only child in the Maryland household is age 5. If the age distributions in the two states are similar, each of the households (or ones like them) could about equally well have appeared in the other state's sample. In other words, the presence of the household with the 4 year old in Virginia's sample rather than Maryland's sample reflects sampling variability. Thus, if a count of the number of households with children under age 5 is needed or if some program provision being simulated with a microsimulation model is triggered by the presence of a child under age 5, better estimates could be obtained for both states by giving each state a copy of each of the two households but with half as much weight. On the other hand, if we have evidence that the age distribution of children in Maryland is shifted upward relative to the age distribution in Virginia, we might want to give the household with the older child slightly more weight in Maryland and slightly less weight in Virginia.

Giving weight to out-of-state households introduces some bias—that is, persistent error across samples attributable to modeling—because these households may differ from

households in the state of interest in ways that are not captured by the reweighting model. However, using many more observations that are similar except for state of residence should substantially improve precision, that is, reduce variance, which is nonpersistent error attributable to purely random sampling fluctuations. The objective of reweighting and, more generally, indirect estimation is to enhance accuracy as measured by a standard like mean squared error (MSE) that reflects the tradeoff between bias and variance.

Under our proposed approach, we derive a matrix of state weights. Every household in the database gets as many weights as there are states (51 counting the District of Columbia as a state, as we will throughout this paper). For every state, there is a weight for each household in the database, although some weights may be small or (by design) zero. To derive estimates for any one state, we use all households in the database—regardless of actual state of residence—and apply the appropriate set of weights. Virginia weights are used to derive estimates for Virginia, Maryland weights are used to derive estimates for Maryland, and so forth. Thus, Virginia borrows strength from other states that have households with nonzero Virginia weights.

Using a Poisson regression model, our reweighting method assigns a Virginia weight to a household according to how prevalent that "type" of household is in Virginia. The more prevalent it is, the more Virginia weight it gets. A household's type is defined by all the characteristics in the database, some of which are measured directly while others are calculated or simulated. A household's prevalence is determined, under the model, by a set of household characteristics that (1) are policy-relevant (e.g., those determining program eligibility and benefits), (2) capture the key dimensions along which households in different states are different, and (3) have about the same meanings across states. This third property has important implications when a characteristic—such as an indicator of cash welfare receipt—is highly relevant to the estimates being produced, but means something different in different states because of differences in state policies. In Schirm and Zaslavsky (2001), we discuss how our reweighting approach needs to be modified when this third property is seriously violated because state welfare programs differ substantially.

The variables included in the reweighting model serve as control variables, and households are reweighted so that weighted sums (indirect estimates) equal specified control totals. These totals can be direct estimates, indirect estimates smoothed using empirical Bayes methods, or administrative totals. For example, if the number of children in the household is a control variable, the total state child population is a control total. Similarly, if the number of poor people in the household is a control variable, the total number of poor people in the state is a control total. If a zero-one indicator that the household has earned income is a control variable, the number of households with earnings in the state is a control total.

With the original (national) weights, the database looks like the entire United States. With Virginia weights, the database looks like Virginia in terms of some key aggregates (the control totals). We then conjecture that the reweighted database resembles Virginia in terms of many other relevant aggregates for which we cannot control, including, for example, the main estimands of a microsimulation model. Our evaluation addresses the extent to which this is accomplished.

2.2 The Formal Model

The reweighting model is:

$$w_{hs} = \gamma_{hs} e^{\beta_s' x_h + \delta_h},$$

where w_{hs} is the expected number of households of type h in (the population of) state s. A type is, practically speaking, unique on the database because no two households are exactly alike. Therefore, each household in the database represents its own type, and w_{hs} is the weight that will be given to household h when deriving estimates for state s. γ_{hs} is an indicator set by the modeler to one if state s is allowed to borrow from the state in which household h actually resides, and zero otherwise. Although we will assume for the evaluation presented in this paper that each state borrows from every other state, Schirm and Zaslavsky (2001) describe an application in which the extent of borrowing is restricted. x_h is a column vector of I control variables, that is, household characteristics for household h. β_s is a vector of I unknown parameters to be estimated for each state. δ_h is an unknown parameter to be estimated for each household. The first term in the exponent on the right side of the model reflects the prevalence in state s relative to other states of households of the same general type as household h, that is, with the same vector of observed characteristics. The second term reflects the national prevalence of households of specific type h.

The β_s and δ_h parameters are estimated by maximum likelihood and satisfy two constraints that are the first order conditions for maximum likelihood estimation:

$$\text{Constraint 1: } \sum_s w_{hs} = W_h \text{ for each } h,$$

where W_h is the control weight, that is, the original sample weight or national weight of household h, and:

$$\text{Constraint 2: } \sum_h w_{hs} x_{hi} = X_{si} \text{ for each } s \text{ and } i,$$

where X_{si} is the control total for control variable i in state s. According to the first constraint, reweighting does not change the total weight given to a household across all states, that is, at the national level, ensuring that the household contributes the same to a national estimate after reweighting as before. According to the second constraint, all control totals are satisfied for every state.

These two constraints do not determine, by themselves, unique weights. However, if we first distribute each household's total sample weight uniformly across the states, and then alter those initial state weights the least amount that is necessary to reproduce the control totals (Zaslavsky 1988), we obtain the reweighting model given at the beginning of this section, which in combination with the constraints makes the weights unique. Our estimation algorithm—an iterative two-step procedure described in Schirm and Zaslavsky (2001)—takes this approach to obtain unique weights.

One way to understand the general philosophy underlying our approach to modeling is to think of the entire database as a high-dimensional contingency table with dimensions that are defined by the many household characteristics contained in the database, including state of residence. If we describe this table by a log-linear model, our modeling assumption is that some margins and low-order interactions of the household characteristics—the ones that appear in the vector x_h—are interacted with state, while the high-order interactions of the household characteristics are not interacted with state and, therefore, are the same in each state. In other words, the model

assumes that the ways some household characteristics interact with each other are similar across states. Fitting a model that includes low-order interactions and excludes high-order interactions is a standard approach to smoothing a contingency table. The estimated table that is fit under the model is smoother, that is, less affected by sampling variability than the sample table that is obtained by tabulating every state separately with the original sample weights, which corresponds to the (fully saturated) model in which every interaction among household characteristics is also interacted with state so that no interactions are excluded. By using weights fitted under the model with only low-order interactions, rather than the original sample weights, we reduce the variances for the entries in the high-dimensional table that is our database and, hence, for estimates calculated from the table.

In a simple case, the control variables are all dummy (indicator) variables for membership in a set of strata that fully partition the population of households—that is, the strata are disjoint and exhaustive. In that case, our reweighting procedure reduces to synthetic estimation of state means because the model simply weights the households in each stratum to represent the prevalence of that stratum in the state for which weights are being derived. Our procedure is more flexible than synthetic estimation, however, allowing us to use more variables as controls. For example, our procedure can control for 10 dichotomous controls by fitting 10 model parameters in each state. A purely synthetic approach would have to estimate the prevalence of each of 1,024 categories formed by cross-classifying the 10 variables.

3. EVALUATION DESIGN

We have designed an evaluation of the reweighting method to answer the following main questions:

- Are indirect estimates from a reweighted database more accurate than direct estimates? That is, does the reduction in variance from reweighting more than offset the bias introduced? If indirect estimates are more accurate, how much more accurate are they?

- Among several reweighting models with fewer or more control variables, which give the most accurate estimates for various estimands?

The fundamental problem in answering these questions and assessing the relative accuracy of alternative estimates is not knowing the truth, that is, the true values of the quantities that we are trying to estimate.

We have taken two approaches to evaluating the error of direct and indirect estimates. One approach uses estimates of the variance and bias of an estimator that are internal to the sample at hand. This was the approach that we took in Olsen, Schirm, and Zaslavsky (2000), where the sample at hand was from the Survey of Income and Program Participation. Although variance estimation with SIPP—or Current Population Survey (CPS)—data is difficult due to the complex sample design, replication methods can be used to estimate the variances of the direct and indirect estimators. We can then estimate biases of indirect estimators using the formulae presented in Olsen, Schirm, and Zaslavsky (2000). This approach is well adapted to application in a production setting because it uses only the data in the sample that is already being used for production. A deficiency of this approach is that the estimate of the bias of an indirect estimator is obtained from complex formulae involving several terms measured with error, and is often unstable and highly sensitive to the accuracy of the variance estimates used.

Therefore, for the evaluation presented in this paper, we have taken a second approach based on simulation. With this approach, we construct a known artificial population that is similar to a real population. In other words, we specify what the truth is. Then, we can compare direct and indirect estimates with the truth and measure their accuracy. To construct our simulation population, we combine households from CPS data for several recent years. Then, for each of many samples of households that we draw from the population, we calculate direct estimates for various estimands of interest. We also calculate indirect estimates for each of the reweighting models that we have fit to the sample data. Finally, we calculate the difference between each estimate and the true value for the estimand, and measure the accuracy of the direct and indirect estimates according to mean (across samples) absolute or squared error. In this section, we describe in greater detail these steps in our evaluation. We present results from our evaluation of the reweighting method in the next section.

3.1 Constructing a Population and Drawing Samples

3.1.1 Selecting Households for the Population. Our evaluation measures the accuracy of an estimate by its difference from the true value of the estimand under consideration. We calculate this true value using all of the households—properly weighted—in the population that we have constructed. As described in Schirm and Zaslavsky (2001), we constructed this population by combining nearly 151,000 interviewed households in 24 nonoverlapping rotation groups from the March CPS samples for 1996 to 2000.

3.1.2 Weighting Households in the Population. Each interviewed household in the CPS has a sample weight. Although we refer to our collection of 151,000 households as a "population," we assigned weights to the households. We will refer to the assigned weights as "population weights" to distinguish them from the weights assigned when we draw samples from our population.

The initial population weight assigned to each household was its original CPS sample weight. Then, we raked the population weights for each of the 24 rotation groups in our population to one state population total and 43 national population totals. Our approach to raking is similar to the approach used in weighting the March CPS sample, although we used a smaller set of national controls.

3.1.3 Drawing Samples from the Population. Because an actual CPS sample consists of 8 rotation groups, we drew samples of 8 rotation groups from the 24 rotation groups in our population. We used simple random sampling with replacement to draw 200 samples. To obtain sample weights, we multiplied each household's population weight by the product of three and the number of times that the household's rotation group was selected for the sample. The factor of three reflects the fact that we select 8 of the 24 rotation groups. We drew 200 samples because the coefficient of variation of an estimated variance based on 200 draws is about 10 percent (assuming that estimates

are approximately normally distributed), which we regarded as adequate precision for the comparisons to be made.

3.2 Estimands

We specified 40 estimands for which we would compare the accuracy of direct and indirect estimates for states. Sixteen of the estimands are counts of persons. Each of the other 24 estimands gives the percentage of persons in a specified "denominator" group—persons ages 0 to 4, for example—who are also members of a smaller "numerator" group—persons ages 0 to 4 who are at or below 185 percent of poverty—that is a subset of the denominator group. The numerator of each of the percentage estimands is a count estimand, and each count estimand appears once or twice as a numerator. A rationale for considering related count and percentage estimands is that counts are often more directly interpretable and policy-relevant, but percentages are more comparable across states of different sizes. All of the estimands that we consider are potentially interesting to policymakers or the administrators of food and nutrition and welfare programs or programs for providing health insurance to children in low income families. In addition, the values of each estimand vary substantially or at least nontrivially across states. When developing the list of estimands to consider, we also sought to include some estimands for which direct estimates would be very imprecise for most, if not all, states and other estimands for which direct estimates might be relatively precise for at least some states. All 40 estimands are defined in Schirm and Zaslavsky (2001).

3.3 Reweighting Models

For each of the 200 samples drawn from the population, we fit four nested reweighting models that we have dubbed "null," "small," "medium," and "large." Among all possible reweighting models, the model that allows the most borrowing of strength has no control variables. With our estimation procedure, such a model would spread each household's national weight equally across all states. Then, however, each state would have the same estimated population (the national population divided by 51), which would render meaningless any comparisons of estimated to true counts. Thus, we want any reweighting model to include at least one control variable that enables us to obtain a sensible estimate of each state's total population or some other relevant measure of size. Our null model, so named because it controls almost nothing, controls for the total number of people and the total number of people ages 16 and over. The latter was included because it is the one state-level control used in weighting the CPS (and in raking the weights for our population).

We can think of our small model as a "demographic" model because it includes controls pertaining to the composition of households and states by age, race, and Hispanic origin. The small model also controls in each state for the number of households that are in large central cities with substantial black or Hispanic populations and the number of households that are not in such areas. The main purpose for including these two variables is to restrict somewhat the borrowing of strength from reweighting. Specifically, for the one state (the District of Columbia) with no households outside large central cities with substantial black or Hispanic populations, no weight is given to a household if it is not from such a central city. Likewise, for the 21 states that have no large central cities with substantial black or Hispanic populations, no weight is given to a household from such a central city in another state. For example, no Wyoming weight is given to a household from New York City.

To the small model, the medium model adds three control variables pertaining to the economic status of households. The corresponding control totals allow us to control for the income distribution in a state, as measured by the number of people at or below 100 percent of poverty, the number above 100 percent but at or below 130 percent of poverty, and the number above 130 percent but at or below 185 percent of poverty. Because we control to these three totals and the total population, we also control the number of people above 185 percent of poverty. The income thresholds that define the control variables and control totals in the medium model are the same as the income thresholds that define many of the estimands, so adding such controls may improve the accuracy of indirect estimates.

The large model adds still more controls that are related to the characteristics used to define estimands. Adding such controls may enhance accuracy by making the characteristics of reweighted state populations more similar in relevant ways to the characteristics of the true state populations. The added controls are listed in Schirm and Zaslavsky (2001).

As noted before, we fit each of our four reweighting models to each of the 200 samples. For each sample, we obtained control totals for the four models by direct estimation.

3.4 Measuring Accuracy

We assess accuracy with two commonly used measures of error: mean absolute error (MAE) and mean squared error (MSE). We calculate MAEs for the count estimands and MSEs for the percentage estimands. For a given estimator—the direct estimator or one of the four indirect estimators (null, small, medium, or large)—and a given count estimand, the MAE for a state is obtained by summing absolute errors over the 200 samples and dividing by 200. We obtain an average MAE by calculating a weighted sum of the 51 state MAEs, with states weighted equally. The average MAE can be interpreted as the typical number of people in a category who are placed in the wrong state by a given estimator. State MSEs and average MSEs are similarly obtained from squared errors for percentage estimands. When calculating average MSEs, we consider three weighting schemes. One scheme weights states equally, and the other two give more weight to states with more people in total or in a group that is relevant to the estimand under consideration. In using "average" here and in the remainder of the paper, we are referring to an average across states (and, sometimes in the next section, across estimands as well). In contrast, when we use "mean," we are typically referring to a mean across the samples that we have drawn.

One attractive property of MSE as a measure of error is that in contrast to MAE, MSE can be decomposed into the contribution from variance, which measures how estimates vary about the mean estimate, and the contribution from bias, which measures how the mean estimate differs from the true value of the estimand. Specifically, MSE equals the variance of the estimates plus the bias squared. It is important to understand that despite its pejorative colloquial meaning, "bias" in this context does not imply any intent to be unfair to a state. Rather, it describes the imperfect fit of the models that underlie indirect estimation. Direct estimates are typically unbiased or nearly so,

meaning that a mean estimate over many samples approaches the true population value. The mean value of estimates from an indirect estimator may differ from the population value because a state happens to have more people in a group (e.g., more adults over age 65 and under 130% of the federal poverty level) than could have been predicted from knowing the control totals of the model and the distribution of relevant characteristics in other states. When squared bias is large compared with variance, the model does not predict the estimand in question very well and a model with more controls might perform better. When the contribution due to variance is the largest part of MSE, the model might be too complex and, therefore, the estimates might be made more stable by reducing the number of controls obtained by direct estimation.

4. EVALUATION RESULTS

Several interesting results emerge from our initial analysis of the estimated mean absolute and squared errors for the direct estimator and the four indirect estimators. We discuss these results and extensions of our analysis in this section.

We find that the estimates of relative MSE—MSE for an indirect estimator as a percentage of the MSE for the direct estimator—are very little affected by the choice of weighting scheme. This may reflect the fact that in the CPS, from which we constructed our population, sample sizes are fairly uniform across states, certainly much less variable than the states' populations. Hence, the amounts of sampling error in state estimates are not very much related to the sizes of the states. Because of this finding, we discuss in this paper the results that were obtained by weighting states equally.

According to the MAE and MSE estimates that are presented in Schirm and Zaslavsky (2001), each of the four reweighting models at least sometimes reduces error relative to the direct estimator for both count and percentage estimands. The large model most frequently dominates the direct estimator. It improves accuracy relative to the direct estimator for 14 out of 16 count estimands according to the MAE criterion. For the two remaining count estimands, the accuracy is almost identical. The large model reduces error relative to the direct estimator for 21 out of 24 percentage estimands according to the MSE criterion.

The models with fewer control variables than the large model sometimes improve accuracy relative to the direct estimator, but they do so less frequently than the large model. By the MAE criterion for count estimands, the null, small, and medium models are better than the direct estimator for 1, 7, and 12 out of 16 estimands, respectively, whereas the large model dominates the direct estimator for 14 estimands. The large model dominates the other three indirect estimators for all but three estimands, for which the medium model is most accurate. Similarly, by the MSE criterion for percentage estimands, the null, small, and medium models beat the direct estimator for 6, 13, and 19 out of 24 estimands, respectively, compared with 21 out of 24 for the large model. For 14 out of 24 estimands, the large model has the smallest MSE. The medium, small, and null models have the smallest MSEs for 7, 1, and 1 estimands, respectively, and for one estimand, the large, medium, and small models have the same MSE.

Considering ratios of indirect to direct average MSEs, we find that the average ratio (across estimands) is 254 percent for the null model, 126 percent for the small model, 99 percent for the medium model, and 72 percent for the large model. By a criterion based on these averages, the null and small models do worse than the direct estimator, the medium model does about as well as the direct estimator, and the large model improves moderately on the direct estimator (about as much as increasing the sample size by 40 percent). However, this method of summarizing the results might understate the benefits of model-based estimation. In fact, calculating ratios for each estimand with the MSE of the direct estimator in the denominator of the comparison ratio favors the direct estimator in three ways.

First, there is some random variation in the MSE estimates due to the fact that only 24 rotation groups were available for sampling. That is, although the number of samples drawn (200) was chosen to give adequate precision for estimates of MSE in the simulation, the simulation population itself is a random sample of only 24 rotation groups from the real population of the United States, and therefore differs from the real population randomly in ways that may affect the results for particular estimands. Because of the nonlinearity of a ratio as a function of the denominator, the estimated ratio obtained by taking the ratio of two unbiased estimators tends to be biased upwards, advantaging the denominator quantity.

Second, there is nonrandom variation in the ratio of MSEs as well as random variation, because a model might fit very well for some estimands while the direct estimator might estimate other estimands relatively precisely. Although there are many possible ways of summarizing the comparison of estimators when the MSE ratios vary, averaging the ratios of indirect to direct MSEs tends to advantage the direct estimator for much the same reason as with random variation. To illustrate this, suppose that for one estimand the indirect estimator has 10 times the MSE of the direct estimator, and for another estimand the direct estimator has 10 times the MSE of the indirect estimator. Despite the symmetry of the relationship, the mean indirect/direct MSE ratio is $(1/2) \times (1/10 + 10) = 5.05$ or 505 percent, making the indirect estimator appear far inferior. Inverting all of the ratios, we find that the mean of the direct/indirect ratio of MSEs is 85 percent, 122 percent, 141 percent, and 154 percent, respectively, for the null, small, medium, and large models. These averages make all but the null model appear superior to the direct estimator, and the effective improvement from the large model is like adding 54 percent to the sample size. A measure that is unaffected by ratio bias is the median indirect/direct MSE ratio across the 24 estimands, which is 142 percent, 89 percent, 76 percent, and 70 percent for the null, small, medium, and large models, respectively. Again, this summary indicates the superiority of all but the null model to the direct estimator.

Third, MSE ratios might tend to be systematically related to the sizes of the errors. For some estimands, the MSE of the direct estimator is much larger than for other estimands, perhaps because the relevant sample size is very small in some or all states or because of patterns of clustering of households with relevant characteristics. The MSE of an indirect estimator, on the other hand, is less affected by sample size than the MSE of the direct estimator and more affected by model fit, and there is no reason why the model fit should be systematically related to the sample size for an estimand. Therefore, we might expect that the benefit of indirect estimation would be greater for the "problem" estimands for which the direct estimator has unusually large variance. For the two estimands with the largest variance, the large model indirect estimator reduces the MSE

very substantially (to, respectively, 50 percent and 41 percent of the direct estimator's MSE). This issue is addressed by an alternative summary measure of the performance of the estimators that calculates a mean across estimands for each estimator before comparing estimators (by taking a ratio). This measure treats all estimators as equally important, and hence gives more weight to reduction of error for estimands with large estimation error. Because the estimands are combined before calculating a ratio, this procedure is relatively unaffected by the ratio biases described above. By this measure, the indirect/direct ratios of average MSE are 172 percent, 80 percent, 70 percent, and 58 percent, respectively, for the null, small, medium, and large models. Again, all indirect estimators but the null model outperform the direct estimator, and use of the large model is comparable to a 72 percent increase in sample size. For the two problem estimands, the root mean squared error—or RMSE, which is roughly interpretable as a standard error—is reduced from 12.6 percent to 8.9 percent for one and from 11.1 percent to 7.1 percent for the other.

In summary, the indirect estimators improve upon the accuracy of the direct estimator for most estimands, with the greatest improvements being obtained with the large model. The comparison of average performance of the various indirect estimators with the direct estimator depends to some extent on the method used for aggregating across estimands. Nonetheless, the indirect estimator based on the large model improves on the direct estimator even when they are compared using the standard that is most favorable to the direct estimator. With a standard that aggregates squared error across all estimands, the standard that is most neutral among methods, three indirect estimators—all but the null model—improve upon the direct estimator.

We can gain further insight into the performance of the alternative estimators by measuring the tradeoffs between bias and variance. We find that the estimated bias-variance tradeoffs are related in a predictable way to the sizes of the reweighting models, that is, the numbers of control variables in the models. For the direct estimator, almost the entire MSE is due to variance because the estimator is unbiased except for the ratio bias from taking the ratio of two unbiased estimators. Conversely, the MSE of the null model is almost entirely (98 percent on the average, across estimands) due to bias—representing the difference between the national average (adjusted only for state total population and population ages 16 and over) implied by this model and the various true values in the different states—because the variance of the national estimate is quite small. The share of MSE due to variance grows monotonically with the size of the model. The average share across estimands is 10 percent for the small model, 24 percent for the medium model, and 42 percent for the large model. The null model, which is the smallest model, and the direct estimator, which can be regarded as the largest model, also fit this pattern. This pattern is expected because a model with more parameters (control variables) fits better and has less bias, but the estimates have more variance because more parameters have to be estimated from the fixed number of sample observations. That is, adding more control variables to a model means that more control totals will be estimated directly from data for individual states, rather than (implicitly) data for the whole nation, reducing the amount of borrowing of strength across states. Even with the largest of our models (excluding the direct estimator), more error is due to bias than variance on the average and for 17 out of the 24 percentage estimands. That result and the previously discussed result that the largest model is the most accurate for 14 of the 24 estimands suggest that we may not have reached the break-even point where the increased variance from adding control variables offsets the decreased bias. Thus, an even larger model might yield further improvements in accuracy.

We consider the results of our initial analysis of the estimated MAEs and MSEs to be very promising for the use of indirect estimates derived using our reweighting approach. Thus, we plan to conduct further analysis.

One extension, which was just discussed, is to consider what control variables might be added to our large model to make an even larger model. Then, we would assess whether adding control variables improves accuracy.

We will also work to improve our understanding of factors affecting the performance of the various indirect estimators relative to the direct estimator. For example, we will examine the relationship between the size of the denominator population and, hence, the noisiness of direct estimates of percentages and the amount of improvement obtained from using the various models. In addition, we will look in detail at the likely associations between specific estimands and control variables that would affect the accuracy of the various direct and indirect estimates.

Finally, we will apply the evaluative tools that we have already used to assess the accuracy of composite estimators that combine direct and indirect estimates either in fixed proportions or in proportions that depend on the sample size for each state. Such an estimator, like the other estimators that we have considered, can also be represented by a set of state-specific weights for each observation. Composite estimators have a natural justification in terms of empirical Bayes modeling techniques, which provide an explicit rationale for weighting together direct and indirect estimators to minimize mean squared error. However, our emphasis will not be on deriving the optimal empirical Bayes estimator for each estimand. Rather, we will try to find estimators of this form that can broadly improve on the "pure" strategies described above for a wide range of estimands simultaneously, using a single set of weights.

REFERENCES

Olsen, Cara H., Allen L. Schirm, and Alan M. Zaslavsky. "An Evaluation of State Estimates Produced by Model-Based Reweighting of a National Database." *2000 Proceedings of the Section on Government Statistics and the Section on Social Statistics.* Alexandria, VA: American Statistical Association, 2000.

Schirm, Allen L. and Alan M. Zaslavsky. "An Evaluation of a Reweighting Method for Small Area Estimation." Washington, DC: Mathematica Policy Research, Inc., September 2001.

Schirm, Allen L. and Alan M. Zaslavsky. "Reweighting Households to Develop Microsimulation Estimates for States." *1997 Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association, 1997.

Zaslavsky, Alan. "Representing Local Area Adjustments by Reweighting of Households." *Survey Methodology*, vol. 14, no. 2, December 1988.