The Impact of Measurement Error in Auxiliary Variables on Model-Based Estimation of Finite Population Totals: A Simulation Study

Brady T. West[1]
[1]Michigan Program in Survey Methodology, Institute for Social Research, P.O. Box 1248, Ann Arbor, MI 48106

Abstract
Model-based prediction theory for finite population sampling and inference (Valliant et al., 2000) largely assumes that auxiliary variables are available for all units in the target population. These auxiliary variables play many important roles in prediction theory: they are used to 1) select samples balanced on the auxiliary variables that, when combined with (theoretically) appropriate prediction models based on the auxiliary variables, produce bias-robust estimators of population totals with minimum variance; 2) fit models to sample data which enable making predictions for non-sample cases based on the sample observations; and 3) ultimately make predictions on key survey variables for non-sample units and compute appropriate model-based standard errors for estimated totals based on the predictions. Error-free measurement of the auxiliary variables on the population frame would thus seem critical for making appropriate finite population inferences. Unfortunately, there has been very little research examining the impact of measurement error in the auxiliary variables on the properties of these model-based estimators. This simulation study empirically examines the properties of selected model-based estimators of finite population totals when available auxiliary variables are measured with varying levels of error, and assesses the impact of the measurement error on theoretical expectations for the estimators. Increased variance in measurement errors is shown to increase both the bias and variance of selected model-based estimators, and simulation results show that careful attention to the selection of samples with weighted balance on reasonable powers of the auxiliary variable, as described in Valliant et al. (2000), provides protection against the bias that can be introduced by measurement error in unbalanced samples and maximizes efficiency among competing estimators in the case of measurement error (as expected by theory in the case of no measurement error). R code enabling users to perform similar simulations is provided.

Key Words: Model-based Prediction, Measurement Error, Finite Population Sampling, Estimation of Totals, Auxiliary Variables

1. Introduction

Model-based prediction theory for finite population sampling and inference (Valliant et al., 2000) largely assumes that auxiliary variables are available for all units in the target population. These auxiliary variables play many important roles in prediction theory: they are used to 1) select samples balanced on the auxiliary variables that, when combined with (theoretically) appropriate prediction models based on the auxiliary variables, produce bias-robust estimators of population totals with minimum variance; 2) fit models to sample data which enable making predictions for non-sample cases based on the sample observations; and 3) ultimately make predictions on key survey variables for non-sample units and compute appropriate model-based standard errors for

estimated totals based on the predictions. Error-free measurement of the auxiliary variables on the population frame would thus seem critical for making appropriate finite population inferences. Unfortunately, there has been very little research examining the impact of measurement error in the auxiliary variables on the properties of these model-based estimators (e.g., Bolfarine, 1991). This simulation study aims to empirically assess the properties of selected model-based estimators of finite population totals when available auxiliary variables are measured with varying levels of error, and to determine the impact of the measurement error on theoretical expectations for the estimators.

The impact of measurement error in predictor variables on the bias of estimated regression coefficients in linear regression models has been well-established (e.g., Biemer and Trewin, 1997; Fuller, 1987; Berkson, 1950): the true relationships of predictor variables with dependent variables under some model will be attenuated toward zero when the predictors are measured with error, and this attenuation will increase with additional variance in the measurement errors. More specifically, in some finite population, let X_i represent the value of an auxiliary variable for unit i, measured with error. If T_i represents the true value of this auxiliary variable, which is assumed to be a latent random variable with some mean and variance, then X_i can be defined as

$$X_i = T_i + \varepsilon_i,\ T_i \sim N(\mu_T, \sigma_T^2),\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

The random measurement errors denoted by ε_i are assumed to be independent. For a sample of size n from this finite population, a simple linear regression model with a dependent variable Y regressed on the predictor variable T can be written as

$$Y_i = \beta_0 + \beta_1 T_i + e_i,\ e_i \sim N(0, \sigma_e^2),\ i = 1, ..., n$$

If the errors in this regression model are independent of the measurement errors that define the values on X, and the dependent variable Y is regressed on X (instead of T) by applying ordinary least squares to the sample data (to predict values on Y for non-sample cases, for example), the estimated regression coefficient for X will not be equal to β_1 in expectation, but instead will be attenuated toward zero (Fuller, 1987):

$$Y_i = \gamma_0 + \gamma_1 X_i + e_i,\ e_i \sim N(0, \sigma_e^2),\ i = 1, ..., n$$

$$E(\hat{\gamma}_1) = \left(\frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2} \right) \beta_1$$

The ratio determining the degree of attenuation in the simple linear regression coefficient above, $\sigma_T^2(\sigma_T^2 + \sigma_\varepsilon^2)^{-1}$, is known as the *reliability ratio*, measuring the proportion of the total variance in X that is due to the variance of the true values. Therefore, as the variance of the measurement errors increases, the degree of attenuation toward zero will increase as well. Given the importance of models in computing predictions for non-sample units in prediction theory for finite population inference, the simple case above suggests that measurement error in the auxiliary variables could lead to poor (or less than optimal) predictions for the non-sample units, and increased bias in the resulting estimates of finite population totals.

The selection of samples with either simple or weighted balance on moments of the auxiliary variables (Valliant et al., 2000) also plays a critical role in the theoretical properties of model-based estimators of finite population totals. Measurement error in the auxiliary variables could thus lead to the selection of samples that do not have optimal balance properties. For instance, samples may be selected that have nearly optimal balance properties in terms of the auxiliary variable measured with error. However, predictions based on this sample may be far from optimal, given that population units with true values on the auxiliary variable that may have led to

better balance in the sample may have been excluded from the sample due to the measurement error.

The objective of this study is to use simulations to empirically examine the impact of measurement error in auxiliary variables on the bias and variance properties of selected model-based estimators of finite population totals. Simulations will be based on a known population data set (Hospitals) provided by Valliant et al. (2000), and theoretical properties of the estimators will be examined both when the auxiliary variables are error-free and when varying levels of measurement error are applied to the variables. Specifically, this study seeks answers to the following research questions:

1. Does measurement error in auxiliary variables lead to bias in estimates of finite population totals, and is the amount of bias a function of the type of estimator used?
2. What are the impacts of measurement error in auxiliary variables on robust estimators of standard errors for estimated totals, and the resulting inferences for the totals?
3. Do the effects of measurement error in auxiliary variables on the performance of model-based estimators of finite population totals vary in samples that are balanced and unbalanced on moments of the auxiliary variables?

2. Methods

2.1 Data Source

This study will analyze data from the "Hospitals" population in Appendix B.2 of Valliant et al. (2000), obtained from a national sample of short-stay hospitals with fewer than 1,000 beds in 1968 (N = 393). The auxiliary variable X known for the full population will be the number of beds in each hospital, and the survey variable of interest Y will be the number of patients discharged from each hospital. Figure 1 below presents a scatter plot showing the relationship of Y to X in the Hospitals population, along with the fit of a working model [$M(0,1,1 : X^2)$, or a model including X and X^2 as predictors, with variance in Y proportional to X^2] that has been shown in previous studies (Section 3.5, Valliant et al., 2000) to be a reasonable model for this population.

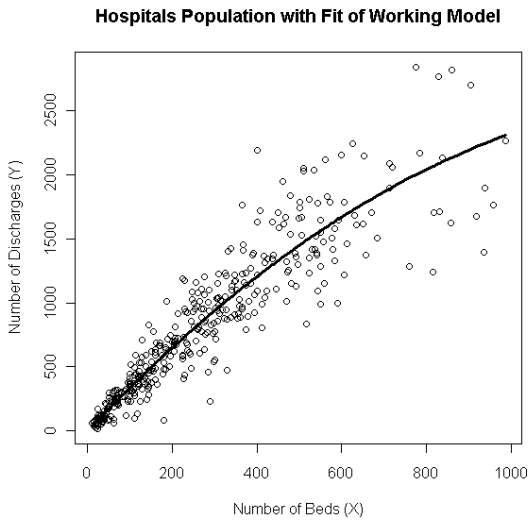

Figure 1: Scatter plot of the data in the Hospitals population, including the fit of the working model $M(0,1,1 : X^2)$ that will be assumed to hold for the simulation study.

The known total number of discharges for the Hospitals population is 320,159. This will be the target parameter T for all estimation procedures in the simulation study, and the relative bias (%) of a given estimator $\hat{T}$ based on a set of S = 1,000 simulated samples (defined as $RB = 100 \times \left(\left[\sum_{s=1}^{S} \hat{T}_s / S \right] - T \right) / T$) will be computed based on this value. Figure 1 clearly illustrates the strong relationship of the auxiliary variable measuring number of beds with the number of discharges in this population (reported by Valliant et al., 2000, to be r = 0.91); relationships between available auxiliary variables and survey variables of interest will likely not be this high in practice, and this study should be replicated in populations where auxiliary variables do not have this strong of a relationship

The present study will simulate the selection of samples and the estimation of total number of discharges based on those samples, considering scenarios with and without measurement error in X. The measurement error in X will be introduced prior to sampling from the Hospitals population as described in the section below.

2.2 Introduction of Measurement Error

In practice, a survey statistician will construct a sampling frame and / or analyze previous samples where the key auxiliary variables may be measured with error. Measurement error may arise, for example, when linking auxiliary variables measured with error in a commercial household database (e.g., Experian) to a sampling frame representing lists of household addresses. Measurement error may also arise in the collection of administrative data that is eventually linked to sampling frames (Davern, 2006).

To simulate the impact of measurement error in auxiliary variables on model-based sample selection, estimation and inference, the first step in each simulated sample will be to either use the Hospitals population as it currently exists, assuming no measurement error in the auxiliary variable X, or introduce measurement error in the number of beds on the sampling frame according to a pre-specified algorithm. In simulations where this auxiliary variable is measured with error, the error will be introduced as follows:

- Assign a random draw from the *UNIFORM*(0,1) distribution to each of the 393 units in the Hospitals population.
- *Low Error Simulations*: These simulations will assume that 80% of hospitals in the population will have the number of beds enumerated correctly (i.e., there is little error in the collection of the number of beds for administrative purposes). If the random draw from the *UNIFORM*(0,1) distribution falls between [0.1, 0.9), including 0.1, the number of beds will be measured without error. If the random draw falls between [0.05, 0.10) or [0.9, 0.95), the number of beds will be either 10% lower or 10% higher than truth, respectively (rounded to provide an integer count). If the random draw falls between [0, 0.05) or [0.95, 1], the number of beds will be either 25% lower or 25% higher than truth, respectively.
- *High Error Simulations*: These simulations will assume that only 10% of hospitals will have the number of beds enumerated correctly (an admittedly extreme case); if the random *UNIFORM*(0,1) draw falls between [0.45, 0.55), the number of beds will be measured without error. If the random draw falls between [0.25, 0.45) or [0.55, 0.75), the number of beds will be either 25% lower or 25% higher than truth, respectively (rounded to provide an integer count). If the random draw falls between [0, 0.25) or [0.75, 1], the number of beds will be either 50% lower or 50% higher than truth, respectively.

- The *reliability ratio*, defined by Biemer and Trewin (1997) as the variance in the true values of X divided by the total variance in X (the variance of the true values plus the variance of the errors) and shown in the Introduction to represent the degree of attenuation of the coefficient for X in a simple linear regression of Y on X, will be computed for each simulated sample based on the error introduced in X. Means of the reliability ratio will be computed for each set of simulated samples, and the relationship of estimation error (differences between estimates of the total number of discharges and the true total) and the reliability ratio will be examined.

By design, this algorithm does not introduce any kind of systematic measurement error in the auxiliary variable (e.g., larger hospitals consistently have more measurement error). The objective of this algorithm is to explore the impact of changing reliability ratios on model-based estimates of finite population totals according to different combinations of sampling methods and estimators, as described in the next section.

2.3 Sample Designs, Estimators, and Variance Estimators

Six (6) unique combinations of sampling methods, estimators, and variance estimators will be considered for each of the three measurement error conditions defined above, resulting in a total of 18 unique simulations. In each simulation, S = 1,000 samples of size n = 50 will be selected from the population of N = 393 hospitals. Based on the 1,000 samples, the relative bias (%), empirical root mean squared error (RMSE, defined as $\sqrt{\sum_{s=1}^{S}(\hat{T}_s - T)^2 / S}$), 95% confidence interval coverage (based on a critical value of $t_{0.975,49} = 2.01$ and a specified variance estimator), and mean 95% confidence interval width for a given estimator of the population total (and corresponding variance estimator) will be computed. For samples where models are fitted to the sample data, outliers (or poorly fitted observations) will be defined as having standardized residuals that are larger than three (3) in absolute value. The total number of outliers will be computed for each of the samples and then averaged across the 1,000 samples, to see if measurement error has a tendency to introduce more outliers when using model-based approaches. This section describes the six combinations of sampling methods, estimators and variance estimators in detail, along with theoretical expectations for each combination in the case of no measurement error in the auxiliary variable.

1. Unbalanced Sampling / Expansion Estimator: In the combinations with unbalanced sampling, simple random samples without replacement (SRSWOR) will be selected from the Hospitals population, which by definition are not balanced in any way on values of the auxiliary variable X. For a given sample s, the expansion estimator will be computed as $\hat{T}_{\text{exp}} = N\overline{Y}_s$. The simple expansion estimator, which fails to incorporate any information in the auxiliary variable X, is theoretically justified by a simple homogeneous mean model with independent homoscedastic errors, which clearly does not hold for the Hospitals population (Figure 1). If a more general second-order polynomial model M holds for the Hospitals population (which is assumed to be the case in this study), then under unbalanced sampling, the expansion estimator will have bias of the form

$$E_M[\hat{T}_{\text{exp}} - T] = N\sum_{j=1}^{2}\beta_j[\overline{X}_s^{(j)} - \overline{X}^{(j)}]. \qquad (1)$$

Thus the amount of bias is a function of the discrepancies between the sample means for the number of beds and the squared number of beds and the true population means on these variables, and larger discrepancies tend to be more likely in unbalanced samples. The expansion estimator is expected to be unbiased over all simple random samples, due to negative biases cancelling with

positive biases. However, SRSWOR in general does not do well at achieving balance (Section 3.4.1, Valliant et al., 2000); as the sample size increases, the bias of the expansion estimator never becomes inconsequential. Also, given the lack of balance, the RMSE of the expansion estimator is expected to be higher than under more balanced sampling. A design-based variance estimator will be used for the expansion estimator in this case, which is expected to be biased under the specified working model, where the variance of Y is a function of X (rather than a constant). Results based on this combination (expected to be poor) will be used as a reference for comparison of performance with the other combinations below.

2. Unbalanced Sampling / Ratio Estimator: Under the assumed second-order polynomial model M for the Hospitals population, the bias of the ratio estimator, which incorporates the auxiliary information on X in the sample and is defined as $\hat{T}_R = N\bar{Y}_s \frac{\bar{X}}{\bar{X}_s}$, is

$$E_M[\hat{T}_R - T] = N\bar{X}\sum_{j=0}^{2}\beta_j\left[\frac{\bar{X}_s^{(j)}}{\bar{X}_s} - \frac{\bar{X}^{(j)}}{\bar{X}}\right]. \qquad (2)$$

This bias could be substantial in unbalanced samples, and an upcoming combination will consider the use of simple balance in samples to eliminate the bias. To protect against potential misspecification of the variance structure for Y under the assumed model for the Hospitals population (with the ratio estimator, the variance is proportional to X), the asymptotically consistent and robust jackknife variance estimator (Section 5.4.2, Valliant et al., 2000), which is guaranteed to have positive bias and will result in conservative inferences, will be used for this model-based estimator. Given the slight misspecification of the mean component of the model underlying the ratio estimator [$M(0,1: X)$], we expect overestimates of the sampling variance, and given that no precautions have been taken regarding the design of the sample for this combination, this robust variance estimator will not guarantee sound or conservative confidence intervals. Further, because individual points with large leverage may seriously inflate jackknife variance estimates and cause them to be become unstable (Valliant et al., 2000, p. 143), each simulation in this study that uses a robust jackknife variance estimator will compute the total number of points in each sample with leverage greater than $2p / n$, where p is the number of predictors in the model fitted to the sample data (e.g., $p = 1$ for the ratio estimator) that underlies the estimator of the total and $n = 50$.

3. Unbalanced Sampling / Minimal Model Estimator: The *minimal model* estimator for the Hospitals population includes as predictor variables the function of X that defines the variance of Y in the working model (X^2), along with the corresponding standard deviation of Y (as a function of X). Hence, the model includes X and X^2 as predictors of Y for non-sample cases. Estimates of the power used to define the variance of Y as a function of X in the minimal model are generally based on previous data or similar populations; using results from previous investigations of the Hospitals population, we will assume that the variance for this model is proportional to X^2. The minimal model is primarily used with samples having *weighted balance* (see combination 6 below), which per Theorem 4.2.1 in Valliant et al. (2000) will achieve a lower bound on the variance of the estimated total. More careful modeling is generally necessary in unbalanced samples; using the same minimal model estimator consistently will likely lead to more bias relative to the weighted balance condition, and a higher RMSE (in some samples, the model will simply not make any sense). This combination will be used for comparison with the weighted balance combination in cases involving measurement error, to see if balanced sampling in combination with a minimal model estimator still minimizes the variance of the estimated total in the case of measurement error. The same robust jackknife variance estimator will be used to estimate variances of this estimated total.

4. Simple Balanced Sampling / Expansion Estimator: Restricted random sampling (RSRS), as implemented in the `restrict.srs()` function of Valliant et al. (2000), will be used to select simple random samples that have simple balance on the first two moments of X (for this and the next combination). The E1 and E2 parameters of this algorithm (Section 3.4.4 of Valliant et al., 2000) will be set to 0.125, such that about 90% of samples will be rejected, leading to samples with reasonable balance on the first two moments of X (Herson, 1976). The expansion estimator will be unbiased in samples with simple balance, even despite the incorrect modeling. The robust jackknife variance estimator will be used for conservative variance estimation, given the balanced sampling and the clear misspecification of the mean structure for Y inherent to the expansion estimator (Section 5.6, Valliant et al., 2000; future work could derive the variance of the expansion estimator under the working model for the Hospitals population). This combination once again serves as more of a reference case to which results from other combinations will be compared in the case of measurement error in X, as one would typically not use the simple expansion estimator in the presence of a strongly predictive auxiliary variable like the number of beds.

5. Simple Balanced Sampling / Ratio Estimator: The bias of the ratio estimator in the case of unbalanced sampling (defined for the second combination above) can be removed by balancing on the j-th power of X in the general polynomial model (defined in Section 3.2 of Valliant et al., 2000), using a sample with simple balance (i.e., a sample that is balanced on means of every power term for X up to the j-th power in the general polynomial model). The use of the `restrict.srs()` function will aim at satisfying this condition for the first two moments of X, given the assumed model for the Hospitals population. In general, samples with simple balance will make both the expansion estimator and the ratio estimator unbiased under the general polynomial model. As in the case of unbalanced sampling, the robust jackknife variance estimator will be used for variance estimation. Per Section 5.6 of Valliant et al. (2000), robust variance estimators will tend to be conservative and over-estimate variance when the regression component of a model is incorrectly specified; in this case, the misspecification is not severe.

6. Weighted Balance Sampling / Minimal Model Estimator: By Lemmas 3.3.1 and 3.3.2 presented in Valliant et al. (2000), under *weighted* root(v) balance (as defined in Section 3.3 of Valliant et al., 2000, and implemented using restricted PPS sampling in their `restrict.pps()` function), where v is the function of X to which the variance of Y is proportional, the minimal model estimator defined above will be the best linear unbiased predictor (BLUP) of the population total. Further, by Theorem 4.2.1 (Valliant et al., 2000), a sample with weighted root(v) balance will also achieve minimum variance in the estimated total, and the minimal model estimator under root(v) balance has been shown to have lower RMSE than the ratio estimator under simple balance, per simulation results (Section 3.5.1, Valliant et al., 2000). To ensure theoretical results, when samples with weighted root(v) balance are selected in the simulations, the assumed variance for the minimal model estimator will be the variance v (i.e., X^2) used for selecting the root(v) balanced sample. The robust jackknife variance estimator will be used for variance estimation for this combination, given the potential that the variance of Y has been mis-specified in the working model. In simulations involving measurement error and weighted balance sampling, an R function named `gamma.fit()` (available upon request from the author) will be applied to the full Hospitals population to *estimate* the power of X to which the variance of Y is proportional, and means of these estimates will be computed across the 1,000 samples. Rounding of these estimates is described in the R code available from the author.

The objectives of the simulations presented in this study are to evaluate the properties of these various combinations of sampling methods and estimators in the presence of different levels of measurement error in the auxiliary variable *X*, and empirically examine whether the theoretical expectations described above are altered by the measurement error. Different levels of measurement error should not matter in the combinations where the expansion estimator is used; in these cases, only alternative sampling methods are expected to have an impact on the properties of the expansion estimator. As previously mentioned, the combinations involving expansion estimators are expected to produce poorer results to which other results will be compared.

2.4 Simulation Notes

Simulations were programmed using the R software (Version 2.9.2). The R code defining a flexible R function for running the simulations entitled `aux.me.sim()`, along with the code for all functions called by this function and results from applying the function for the 18 simulations, can be requested from the author. This function enables the user to input one of the six combinations of sampling method and estimator defined in the section above, parameters used to define measurement error (i.e., cut points for the random values from the UNIFORM distribution and magnitudes of measurement error in the specified brackets, e.g., the value of *X* will be 25% lower if the UNIFORM draw is between 0 and 0.05), if applicable, and the sample size and the number of samples to select. In all simulations in this study, 1,000 samples of size $n = 50$ were selected and analyzed under a given set of simulation parameters. Future work could examine the sensitivity of the results to variation in the sample sizes.

3. Results

Empirical results from the 18 simulations are presented in Table 1. Considering first the case of no measurement error in the auxiliary variable *X*, we see that the simulation results largely support theoretical expectations. For all six combinations of sampling method and estimator, there is very little bias in the estimate of the population total for the number of discharges *Y* (320,159). The expansion estimator in the case of unbalanced sampling has the most bias (although the bias is fairly minimal), and the bias is in a negative direction. The minimal model estimators under both unbalanced and weighted balance sampling have the least relative bias.

In the case of no measurement error in the auxiliary variable *X*, the largest differences between the six combinations of sampling method and estimator arise in the RMSE of the estimators and the properties of confidence intervals based on the robust jackknife standard errors. As expected based on theory, the minimal model estimator under weighted balance sampling has the smallest RMSE and mean confidence interval width, while the expansion estimators have by far the highest mean confidence interval width and the expansion estimator in the case of unbalanced sampling has the highest RMSE (which was also expected). Figure 2 shows box plots of the 1,000 estimates for each of the six combinations under no measurement error, where the increased efficiency due to the minimal model estimator in combination with a weighted balance sample is evident. The 95% confidence interval coverage for the combinations appears to be close to nominal, save for the expansion estimator in the case of balanced sampling; this could be due to a sub-optimal choice of variance estimator in this case. Finally, outliers do not appear to be substantial problems for models fitted to the sample data in the case of no measurement error, but a fairly large number of points with high leverage were detected on average across the samples (more than 10% of sample points on average in some cases). This could be a function of the small

Table 1: Empirical results from the 18 simulations, presented by the six combinations of sample balance / estimator and level of measurement error in the auxiliary variable *X*.

	No Measurement Error in *X*					Low Measurement Error in *X*						
Sample Balance / Estimator	**Rel. Bias (%)**	**RMSE**	**95% CI Cover.**	**Mean 95% CI Width**	**Mean Outlier / HL Ct.**	**Rel. Bias (%)**	**RMSE**	**95% CI Cover.**	**Mean 95% CI Width**	**Mean Reliab. Ratio**	**Mean Gamma Est.**	**Mean Outlier / HL Ct.**
Unbalanced / Expansion	-0.24	29924.85	0.95	122016.80	N/A	-0.17	31169.90	0.95	122338.00	0.98	N/A	N/A
Unbalanced / Ratio	0.10	13930.44	0.94	54815.16	0.38 / 5.66	0.22	14255.99	0.94	57664.38	0.98	N/A	0.40 / 5.83
Unbalanced / Minimal	0.03	12615.44	0.94	51264.66	0.38 / 3.24	0.16	12862.90	0.95	52867.96	0.98	N/A	0.42 / 3.31
Simple / Expansion	0.18	12452.82	1.00	124132.8	0.35 / 0.00	-0.01	12727.67	1.00	124202.70	0.98	N/A	0.39 / 0.00
Simple / Ratio	0.15	12005.67	0.97	55123.84	0.37 / 5.67	-0.01	12262.45	0.97	58068.69	0.98	N/A	0.42 / 5.81
Weighted / Minimal	0.03	9976.87	0.96	40384.44	0.51 / 4.11	-0.06	10321.60	0.96	42100.24	0.98	1.51	0.52 / 3.88

	High Measurement Error in *X*						
Sample Balance / Estimator	**Rel. Bias (%)**	**RMSE**	**95% CI Cover.**	**Mean 95% CI Width**	**Mean Reliab. Ratio**	**Mean Gamma Est.**	**Mean Outlier / HL Ct.**
Unbalanced / Expansion	-0.17	31169.90	0.95	122338.00	0.72	N/A	N/A
Unbalanced / Ratio	0.84	24287.69	0.96	101633.70	0.72	N/A	0.26 / 6.69
Unbalanced / Minimal	-0.90	23469.62	0.95	94522.43	0.72	N/A	0.36 / 3.65
Simple / Expansion	-0.11	18283.34	1.00	123955.40	0.71	N/A	0.36 / 0.00
Simple / Ratio	-0.08	17937.33	0.99	101700.10	0.71	N/A	0.24 / 6.73
Weighted / Minimal	-0.03	18255.97	0.95	72959.44	0.72	1.22	0.67 / 3.63

sample size or an indication that the jackknife variance estimator may not be the most stable choice among alternative robust variance estimators.

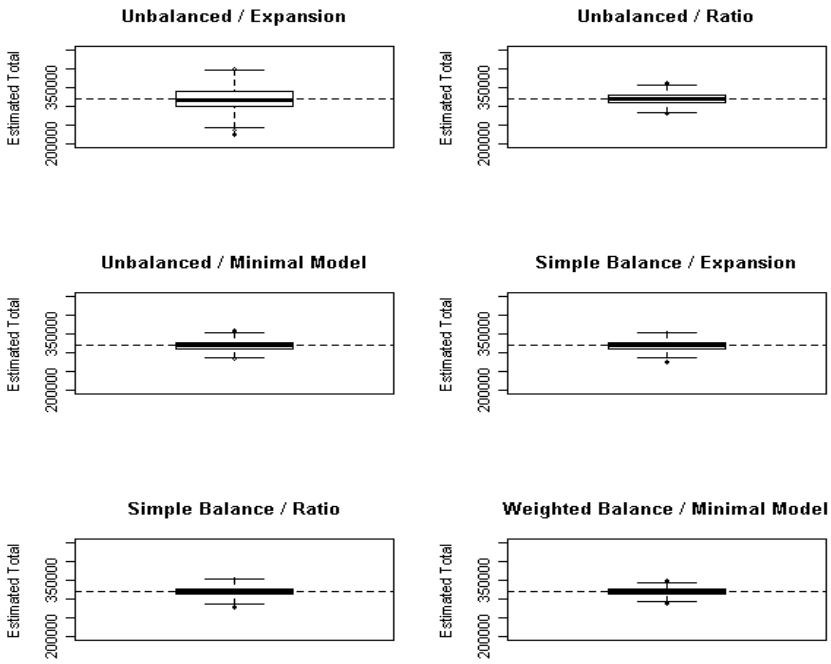

Figure 2: Distributions of sample estimates for 1,000 samples as a function of sampling method and estimator used, in the case of no measurement error (the dashed line shows the true population value for the total number of discharges).

For the sake of comparison and given the small sample size (n = 50), five of these six simulations were repeated in the case of no measurement error using the robust sandwich variance estimator (Section 5.3, Valliant et al., 2000), which like the jackknife variance estimator is consistent but does not involve dividing by a factor involving 1 minus the leverage (as is the case with alternative robust variance estimators, including the jackknife). For the final five combinations in Table 1 considering robust variance estimators, 95% confidence interval coverage under the sandwich variance estimator was 0.93, 0.93, 1.00, 0.96, and 0.94, respectively; and mean 95% confidence interval width was 53003.42, 47975.44, 122726.00, 53274.77, and 38985.08, respectively. Compared with the results in Table 1, these results suggest that inflation in the jackknife variance estimates due to points with high leverage was not severe.

Two of the primary research questions motivating this study concerned whether measurement error in auxiliary variables leads to bias in model-based estimates of totals and whether the amount of bias was a function of the combination of sampling method and estimator used. The results in Table 1 show that when the level of measurement error in an auxiliary variable is relatively low (mean reliability ratio around 0.98), the relative biases for the six combinations of sampling method and estimator are largely similar to those in the case of no measurement error (nothing is expected to change in the cases of the expansion estimators when measurement error in X is present, and this is shown). However, in the case of relatively high measurement error in the auxiliary variable X (mean reliability ratio around 0.72), more interesting results emerge. In the case of unbalanced samples under the high measurement error condition, the measurement

error introduces a relatively large amount of bias in the model-based ratio and minimal model estimators, in a positive direction for the ratio estimator and a negative direction for the minimal model estimator (potentially due to the curvature in the model being fitted for the minimal model estimator). This bias is much larger than the bias in the expansion estimator under unbalanced sampling, and the use of simple balanced or weighted balance sampling appears to *remove* the bias introduced by measurement error. These findings suggest that attention to reasonable sample balance on moments of the auxiliary variable can provide a form of protection against the potential bias introduced by measurement errors in model-based estimators of finite population totals.

A third research question under investigation in this study concerned the impacts of measurement error in auxiliary variables on the variability of estimates and corresponding inferences for the finite population total. The RMSE, 95% confidence interval coverage, and mean 95% confidence interval width for the six combinations of sampling method and estimator were largely similar for the low measurement error condition compared to the no measurement error condition. These properties of the estimators all increased slightly in the case of low measurement error, suggesting that measurement error may have an impact on the efficiency of the estimates and corresponding inferences. Once again, more interesting results emerged in the high measurement error condition. Interestingly, the RMSE of the minimal model estimator under weighted balance sampling was very similar to that of the expansion estimator and the ratio estimator under simple balanced sampling, and in general reductions in RMSE for the other five estimators relative to the combination of unbalanced sampling and the expansion estimator were not as high as in the case of no measurement error in X. Of note, the use of the robust jackknife variance estimator in conjunction with the minimal model estimator under weighted balance sampling led to much narrower 95% confidence intervals on average relative to the other five combinations. These results suggest that although measurement error can have a negative impact on the precision of these model-based estimates of finite population totals, use of the minimal model estimator in conjunction with weighted balance sampling once again results in more efficient inference relative to the other five combinations, as expected by theory.

Finally, Figures 3 and 4 present scatter plots of the simulation results that demonstrate a negligible relationship of the reliability ratio with estimation error for each of the six combinations of sampling method and estimator, under both low and high measurement error conditions. Figure 3 provides another visual representation of the increased efficiency in the estimates under simple or weighted balance sampling in the case of low measurement error, and Figure 4 shows how the effect of sample balance on efficiency (relative to Figure 3) is decreased as measurement error increases.

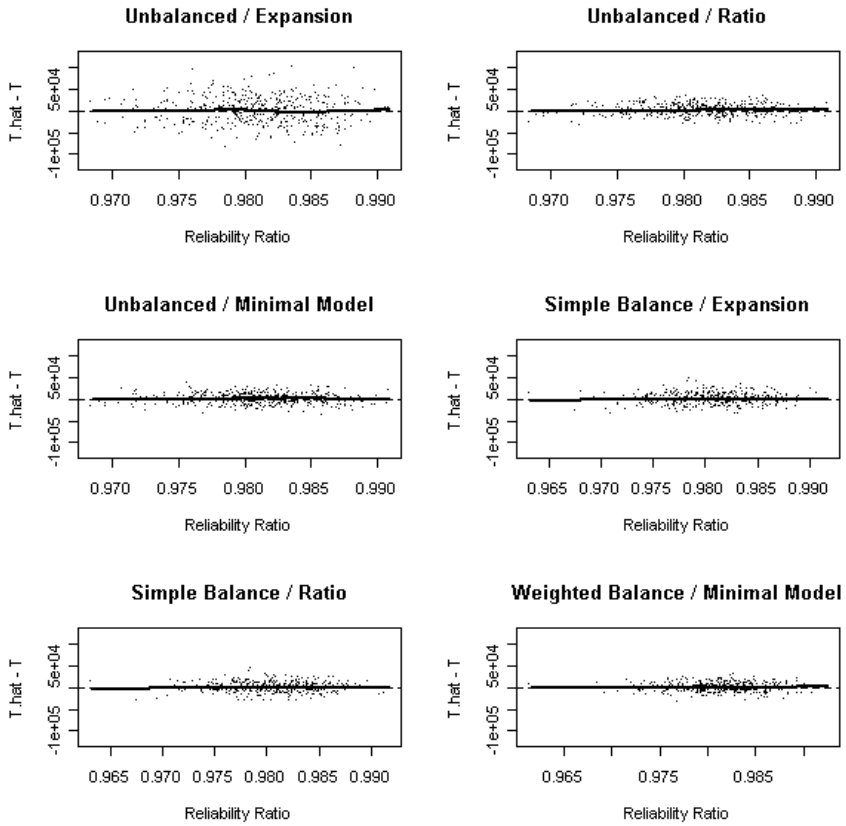

Figure 3: Scatter plots demonstrating a minimal relationship of reliability ratio with estimation error (labeled "T.hat – T" on the y-axis) in the low measurement error condition, for each combination of sampling method and estimator across the 1,000 samples (the thick black line is the fit of a Lowess smoother to the plotted points, which largely lays on the horizontal line at 0).

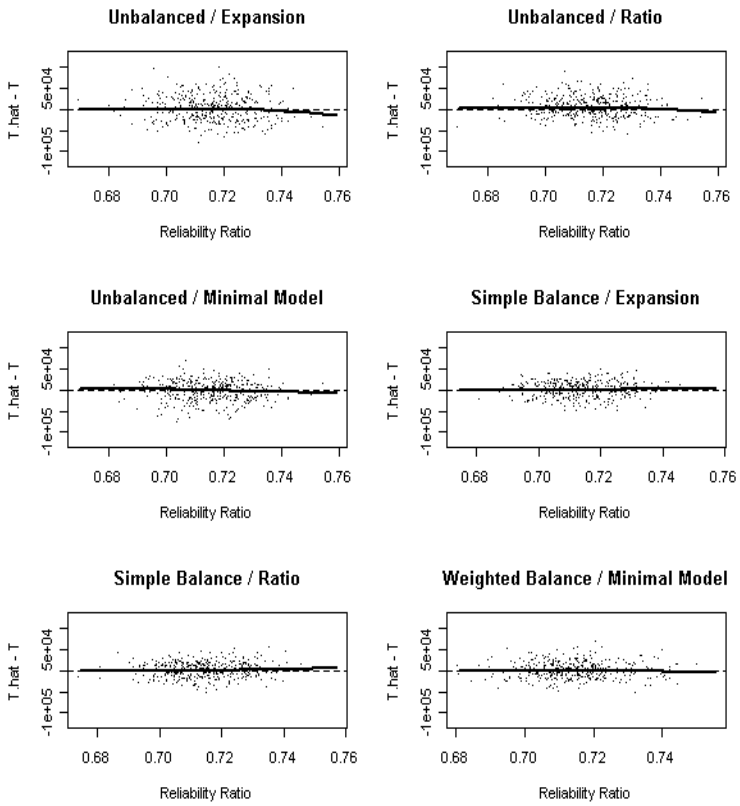

Figure 4: Scatter plots demonstrating a minimal relationship of reliability ratio with estimation error (labeled "T.hat – T" on the y-axis) in the high measurement error condition, for each combination of sampling method and estimator across the 1,000 samples (the thick black line is the fit of a Lowess smoother to the plotted points, which largely lays on the horizontal line at 0).

4. Summary

The results of this simulation study have shown that measurement error in auxiliary variables can have a negative impact on both the bias and precision of model-based estimators of finite population totals. Careful attention to the selection of samples balanced on reasonable powers of the auxiliary variable, as described in Valliant et al. (2000), has been shown empirically to provide protection against the bias that can be introduced by measurement error in unbalanced samples, and minimal model estimators in combination with samples having weighted balance have been shown to maximize efficiency in cases involving measurement error, as expected by theory. The simulation results also suggest that there is a negligible relationship between the reliability ratio for an auxiliary variable measured with error on a sampling frame and the resulting estimation error for all combinations of sampling methods and estimators that were studied.

This work could be extended by studying the effects of measurement error in multiple auxiliary variables, including categorical variables, on the performance of these combinations of sampling method and estimator, and considering overall effects of measurement error on the estimation of totals for multiple *Y* variables in multi-purpose surveys. In addition, future work should also consider the possibility that the amount of measurement error may be greater for auxiliary

variables with higher values (i.e., there is a correlation between error and the size of the value on the auxiliary variable, which was not considered in these simulations). Future work could also consider larger populations and auxiliary variables having weaker relationships with key survey variables. Larger populations would enable more study of the impact of altering sample sizes on the results presented in this study.

References

Berkson, J. (1950). Are There Two Regressions? *Journal of the American Statistical Association*, 45(250), 164-180.

Biemer, P.P., and Trewin, D. (1997). A Review of Measurement Error Effects on the Analysis of Survey Data. Chapter 27 in *Survey Measurement and Process Quality*, Editors Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, and Trewin.

Bolfarine, H. (1991). Finite-Population Prediction under Error-in-Variables Superpopulation Models. *The Canadian Journal of Statistics*, 19(2), 191-207.

Davern, M. (2006). Incorporating Linked Survey and Administrative Data Files into Policy Research. Presented to the Federal Committee on Statistical Methodology, November 2006.

Fuller, W. (1987). Chapter 1: A Single Explanatory Variable. *Measurement Error Models*. Wiley.

Herson, J. (1976). An Investigation of Relative Efficiency of Least-Squares Prediction to Conventional Probability Sampling Plans. *Journal of the American Statistical Association*, 71, 700-703.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley.