Using Multiple Imputation to Estimate the Effect of Data Entry Errors

Barry W. Johnson , R. Louise Woodburn, and Vicki A. Cutwright
Internal Revenue Service

The Statistics of Income Division (SOI) of the Internal Revenue Service (IRS) is charged with collecting data from tax returns for use by government and private researchers in the analysis of tax policy and the economy. As a part of its ongoing quality improvement efforts, SOI has recently instituted a double sample quality review (QR) system to measure data capture errors. Data from a small sample of returns are transcribed by two different workers, the two versions are computer matched and the resulting discrepancies are noted and then corrected.

This paper will examine the use of error data captured by a double sample quality review system to estimate the bias and variance due to data capture errors. First, the Federal estate tax data used for the investigation will be described. Next, a general discussion of non-sampling error will follow, with details on the SOI error data. The strategy used will then be described, along with the methodological issues. Fourth, results and conclusions will be presented. Finally, suggestions for future research will be discussed.

■ The Federal Estate Tax Data

The Statistics of Income Division conducts research on the characteristics of the Nation's top wealthholders through studies of Federal estate tax returns. As a part of this research, SOI has been collecting data for returns filed since the inception of the estate tax in 1916 (see McCubbin, 1990). These returns contain detailed demographic and financial data for wealthy decedents, including complete listings of their assets, liabilities, and beneficiaries. This database, known as the Intergenerational Wealth Study (IWS), will be a rich resource with which to study the effects of intergenerational transfers on the accumulation of wealth, the composition of wealth, and the ways they have changed over time.

The data are collected from the estate tax returns by workers at IRS service centers nationwide. The workers transcribe data from the return into an on-line relational database using ORACLE software. Because some of the financial data can be very detailed and complex, workers must exercise a great deal of judgement in categorizing portfolio components. As data are transcribed, tests are performed to ensure that entered values are consistent with those already keyed, as well as with predetermined parameters. If an error is detected, a corrective action must be taken immediately in order to continue entering data. However, due to the complex nature of the data, not all errors can be successfully caught using these tests.

The data used in this paper are a sample of estate tax returns collected during the years 1991 and 1992 for decedents dying between 1950 and 1981. The original SOI sample is a simple stratified sample with three stratifiers, age at death, total gross estate (TGE) and year of death.

■ Non-Sampling Errors

It is well known that there are basically two sources of error inherent in any data collected through statistical samples: sampling error and non-sampling error. Sampling error is present whenever a subset of records is used to represent a population. Much is already known about the nature of this component of error, including well developed theories of how to measure and control its effect (Wolter, 1985; Cochran, 1977).

Non-sampling error constitutes the remaining component of error in a database. In fact, in a well-constructed sample, it can be the most important source of error. It includes non-response, wrongly conceived definitions, missing data, keying errors, errors in interpreting responses, and errors in interpreting transcription/editing instructions. Often these types of errors can be reduced by introducing greater structure into the data capture system or by increasing worker motivation. There is not, however, a comprehensive theory for assessing the effects of non-sampling error (Lessler and Kalsbeek, 1992).

Early SOI efforts to measure non-sampling error due to keying or interpretation errors involved having a

second worker, usually a supervisor, review a sample of the transcribed records by comparing captured data to the source document. Mistakes were reported to the original worker, who would make the corrections. Recently, SOI has developed automated double sample quality review systems, similar to those used by other statistical and survey institutions, including the Bureau of the Census (Biemer and Forsman, 1992).

Double sample quality review involves the daily computerized selection of a sample of completed work for re-transcription by a second worker. The second worker is not given access to data from the original version. Once complete, the original and second versions are computer matched and a listing of mismatches is generated. These mismatches, or discrepancies, are then resolved by a supervisor and corrections to the database are made. Information is kept on the magnitude of errors and their possible causes.

The importance of preventing the second worker from accessing the original database values is emphasized in a study of resurvey methods employed in conjunction with the 1960 Census (Bailar, 1968). This research concluded that the best reinterview procedure is one which is close in time to the original survey and one in which the reinterviewers are not given access to the original responses. The study found that error rates for dependent reviews were about half those found using an independent review method. A comparison of error rates calculated for the IWS data using both types of reviews yielded similar results.

It is well-recognized that a blind double sample review does not detect all errors in the database, because there is the potential for the second worker to make as many, or more, mistakes during the re-transcription of the record as were made during the original transcription (West and Winkler, 1991). It is also possible that all of the workers will transcribe data in the same way, but that their interpretation of the transcription instructions may not be correct. It is, however, possible to study the existence of bias even when the review process is also subject to error (Lessler and Kalsbeek, 1992). We make the simplifying assumption that the distribution of errors described by the quality review system represents the distribution of all the errors in the database.

■ Strategy and Methodolgy

A special database is used for this investigation, which contains both the original and corrected values for each variable. The basic strategy is to select sub-samples to be treated as the reviewed (corrected) data. We use this 'correct' data to investigate three different strategies for modeling the errors. The models are applied to the remaining 'uncorrected' portion of the database and corrected data values are imputed. Thus, we are able to evaluate the success of our efforts by comparing the imputed estimates to the 'correct' answer. Different review sample sizes are investigated to determine their effect on the efficiency of the models.

We apply the imputation methodology customarily used for missing information to our data. Just as imputation methods use the complete data to estimate a model for the missing values, we use the corrected data from the quality review sample to estimate a model for correct values in the remaining data. The main difference is that we have the additional task of determining which records are in error for the uncorrected portion of the data set. Little (1988) makes the point that imputation methods should be based on the predictive distribution of missing values, taken in the context of all the observed values for a particular case. Basing the model of missing values on the distribution of known values is preferable to filling in missing values with some conditional mean, because it avoids distortion of the distribution of the data (Little, 1990). By imputing multiple values, the uncertainty of the modeling process can be measured (Rubin, 1987).

The first method that we investigate is to model the errors directly via ordinary least squares regression. Results for this approach are summarized in the next section. The second method that we investigate examines the error as a percentage of the total gross estate. We then develop a model for the distribution of the percentages. Errors are drawn randomly from this distribution and used to compute the corresponding 'correction.' For the third method, we employ hotdeck imputation and select errors from a 'neighbor,' as defined by total gross estate categories (see Hinkins and Scheuren, 1986).

We narrowed our investigation to the value of corporate stock. For the estate tax data, a majority of the errors arise in the classification of different types of assets, and, thus, some of the errors are dependent. We ignore this possibility here by studying the error for a single variable. However, the collinearity of the errors with other asset values is examined in the next section.

We investigate the following questions: Do our corrections decrease the bias of the estimate? What is the

model uncertainty (evaluated by examining the variance of repeated applications of a model to the same QR sample)? What is the variation between the different modeling methodologies (evaluated by comparing estimates from the different models for a particular set of QR samples)? What is the sampling variability of the QR sample and correction process (evaluated by comparing estimates for each of the models for different QR samples sizes)?

Regression Analysis

The following analysis will determine, using regression analysis, whether any relationship exists between variables present on estate tax returns and the errors made during data capture. If there is a relationship, at what degree are the errors explained by the formulated model(s)?

The dependent error variable of interest (ERROSTK) represents the difference between the final value of corporate stock (after quality review) and the original value. For example, the first worker determines an asset on a return was stock and enters an amount of $5,000. Then, during the QR process, a second worker determines the stock field should instead be $6,000. Therefore, the value for ERROSTK = $6,000 - $5,000 = $1,000; the original value was increased by $1,000. So, ERROSTK could be either positive, negative, or zero.

The independent variables included tax return items such as total gross estate, debts, cash, stock, etc. The *original* values of database variables were used in the regression equation.

Model Assumptions

Since the data were from a stratified sample of tax returns, we cannot assume they are independent and identically distributed (IID). In order to handle this concern, we included the weight of each observation as an independent variable in the regression.

Initially, our model using the original untransformed data violated the assumption of equal error variance, with the residuals' variance increasing with Y. After using the logarithmic transformation, the assumption of equal error variance was satisfied. Approximate normality of the residuals, or, symmetry, is sufficient to satisfy the assumption of normality, since the ε_i are estimates of the theoretical error E, where $E_i \sim N(0,\sigma^2)$. Again, after using the log transformation, our residuals were fairly symmetric. In a regression analysis, multicollinearity concerns the relationship (correlation) of the independent variables to one another. Multicollinearity causes numerical problems, including the inaccurate computation of (1) estimates of the regression coefficients, (2) estimates of standard errors, and (3) hypothesis test statistics. No unreasonable levels of correlation were observed between any independent variable pair, as determined from the Pearson correlation coefficient matrix. It should also be noted that the stronger the correlation between the dependent and independent variables, the more powerful the regression model. ERROSTK was not particularly well correlated with any of the independent tax return item variables.

Logistic Analysis

In the quality review process, it is quite plausible that the primary worker's value for a specific item will be the same as the reviewer's value, with ERROSTK = 0. In fact, 89% of the data actually had zero values for ERROSTK. Fitting a regression line in this case was practically futile. Instead, we employed logistic regression to predict whether or not there was an error, and, if so, used ordinary least squares (OLS) regression to determine the error magnitude. The logistic regression produced the following model, where Logerr = 0 if ERROSTK = 0 and Logerr = 1 if ERROSTK = 0. ***Logerr=2.1182 - 8.71e^{-8} *Total Gross Estate - 1.25e^{-6} * Cash***, at significance level .10. Several analogous measures of OLS R^2 have been proposed for logit models. Aldrich and Nelson (1984) suggest pseudo $R^2 = \chi^2 / (\chi^2 + n)$. The model then results in an R^2 value of .008, with only .8% of the variation explained by the model. Though this measure does not support a straightforward interpretation, as does true R^2, the suggested model is very poor. We must conclude that the specific items on the tax form have little effect on the odds that an error will be made.

OLS Regression Analysis

Now, assuming that an error is present, what is the magnitude and direction (positive or negative) of that error? In this model we eliminated all cases for which there were no errors (ERROSTK = 0). Three regression models were estimated: first, using only cases where ERROSTK < 0; second, only cases where ERROSTK > 0; and third, using all cases, taking absolute values of ERROSTK. For independent variables with 0 values, the value was changed to 1 before taking natural logs.

To examine the distribution of each variable of interest, a simple histogram and box plot were produced. The distributions are highly skewed. Before proceeding with any analysis, the natural logarithmic transformation was used to normalize the variables. Therefore, the regression model is of the type $ln(y) = \beta_o + \Sigma\beta_j ln(x_j) + \varepsilon$.

Using the stepwise regression procedure, the best model was selected with all variables significant at the .15 level. Figure A displays the model parameters, and their associated R^2 explanatory power. The parameter estimates for these 3 groups do not differ much, except for the intercept term, noting that the intercept term is not significant in Model 1 and Model 3. From this, we can conclude that there is not much difference in whether a worker makes a positive or negative error. Also, and more importantly, the considerably low R^2 values show that the specific items on the tax form are not good predictors of transcription errors; the errors are in effect random. In fact, only about 10% of the variation in data errors are explained by any of the 3 models. Thus, we do not use the method further in the analyses.

Figure A. OLS Regression Coefficients and Estimated R^2 Values

	Intercept	ln (TGE)	ln (DEBTS)	R^2
Model 1 ERROSTK < 0	-0.5893	0.6361	0.0985	0.0806
p-value	0.8045	0.0014	0.1289	
Model 2 ERROSTK > 0	-5.5010	0.9990	–	0.1318
p-value	0.0364	0.0001		
Model 3 ERROSTK = 0	-1.9210	0.7047	0.1130	0.1013
p-value	0.2860	0.0001	0.0322	

Error Model Method

For the Error Model Method, we work with the error as a percentage of the total gross estate, (PERRTGE) rather than the raw errors themselves. This forces all computed errors into a feasible range, thus satisfying our editing checks. The first step was to estimate a model for PERRTGE. A histogram of the percentages for the entire file appeared to follow a normal distribution. Thus, we assumed a normal distribution, but computed the mean and variance for each quality review sample used. The following three steps were then employed to impute 'corrected' values for the unreviewed portions of the data set:

(1) Determine which records were in error:

Using the QR sample the proportion of records having errors was computed (PROPERR). Returns not included in the QR sample were assigned a random uniform number between 0 & 1. Those falling below PROPERR were selected for correction.

(2) For the records to be corrected, select a random percent error = PERRTGE:

The mean (m), and standard deviation (std), of PERRTGE are computed using the qr sample. These were used as the parameters of the normal distribution, in order to estimate the distribution of PERRTGE for the entire data set, i.e. $N(m,(std)^2)$. Thus, the randomly selected PERRTGE = m + std*N(0,1), where N(0,1) is a standard normal deviate. Also, since the error cannot be larger than the total gross estate, we bounded PERRTGE, $-1 <= PERRTGE <= 1$.

3) 'Correct' the unreviewed data:

The 'correct' value for stock is then computed as CORSTOCK = TGE*PERRTGE + Original Value. This, again, must be bounded, due to the data constraints, $0 <= CORSTOCK <= TGE$.

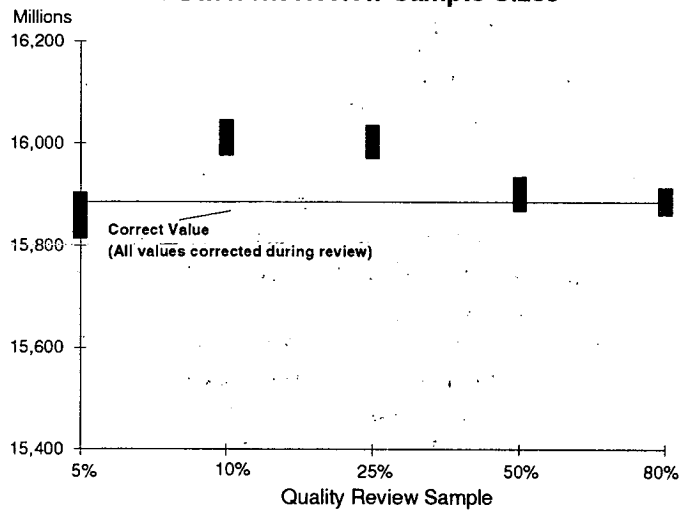
Results

Initially, corrected values were imputed 50 times for a single review sample. A mean value was calculated in a stepwise fashion as each new imputation was completed, in order to see how many imputations were needed to produce a relatively stable estimate. After 15 imputations, each additional imputation had little effect on the mean value of the estimate, even for review samples which were small relative to the uncorrected portion of the database. Figure B shows the estimates of total stock generated by applying the model a total of 15 times for each of 5 successively larger QR samples. The uncertainty caused by the model (spread of the points on the graph for each QR sample) is fairly small and decreases as the QR sample sizes increase.

In order to examine the effect of the quality review sample on the imputed estimates, we repeated the imputation process, using 5 different review samples for each of 5 sample sizes (a total of 25 sets of imputations). The results are shown on Figure C. Here, the 5% sample estimates show the most variability, although clearly centered around the 'true' value. Overall, the variability of the estimates due to the QR sample, itself, decreases

somewhat with the larger sample sizes. The model, however, tends to somewhat over-estimate the true total.

Figure B. Model-Based Imputation: 15 Imputations for Each of 5 Different Review Sample Sizes

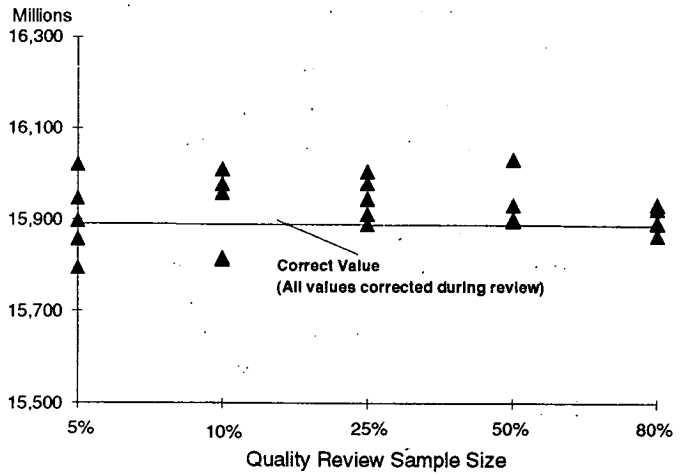


Hotdeck Imputation

The third method of imputation employed the hotdeck procedure within adjustment cells. Records for which errors were to be imputed were matched to donor records in the same adjustment cell. The unobserved error was estimated using the observed value of the error from the donor record.

Earlier attempts to model editing errors revealed that the size of an error was somewhat dependent on the total value of the estate. This is not surprising since most consistency tests, performed as data are captured, limit data values, or sums of related values, to this total.

Figure C. Model-Based Imputation: 5 Different Samples at Each Sample Size



Because of this, adjustment cells were created based on the size of the gross estate. The data were first divided into 3 categories, based on their original SOI sample strata: gross estate under $1 million, $1 million under $5 million, and $5 million or more. Each of these three groups was further divided into the following percentiles, based on the weighted univariate distribution of gross estate values for the corrected sample: 5, 10, 25, 50, 75, 90, 95, 99 and over 99 (these values were used due to software limitations). This resulted in 27 adjustment cells. The uncorrected data were then grouped using the same boundaries.

A value of ERROSTK (final - original value) was then drawn randomly, with replacement, from a donor cell, for each record in a corresponding cell in the uncorrected data set. Thus, in the imputed data set, the 'corrected' values were the original value + ERROSTK. Again, this was bounded, due to the data constraints, 0<=CORSTOCK<=TGE. It should be mentioned that it was possible for a donor record to be error free (ERROSTK = 0); in such cases, no adjustment was made in the uncorrected 'neighbor' cell.

Results

Again, multiple imputations of a single review sample were examined to determine the number of imputations required to produce a stable mean value. As in the error model method, 15 imputations were adequate. Figure D shows estimates of total stock for 5 progressively larger QR samples. The uncertainty introduced by the model, indicated by the spread of the points, is substantially greater than that seen in the earlier model. The variance does decrease slightly as the QR samples increase in size.

Figure D. Hotdeck Imputation: 15 Imputations 5 Different Review Sample Sizes

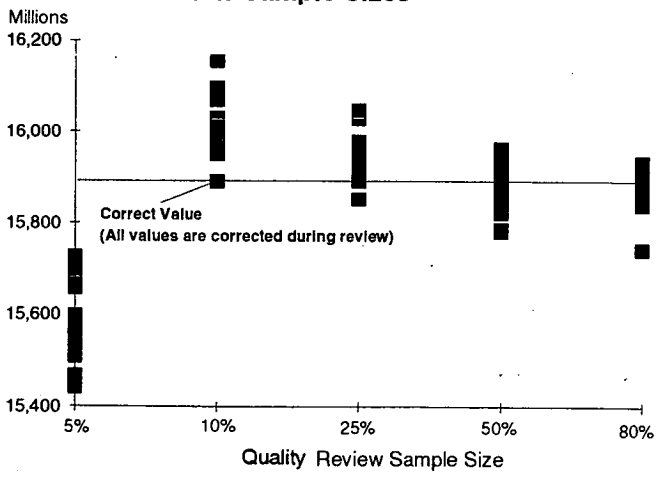
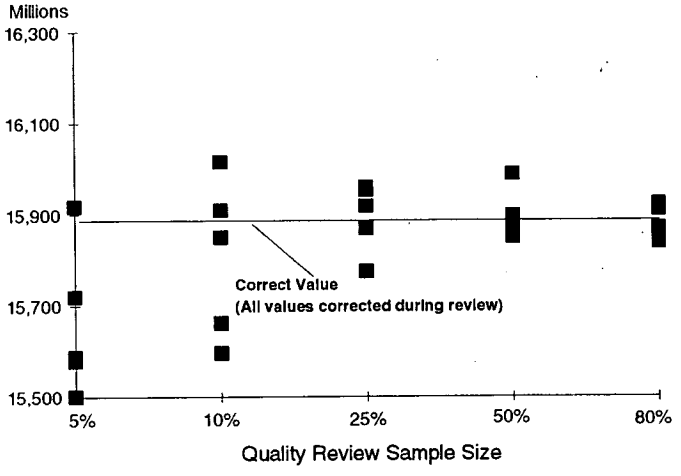


Figure E shows the mean values of 15 imputations for 5 different QR samples at each of 5 sample sizes (the same 25 QR samples used in the previous section). As before, the variability of the estimates declines as the sample size increases. The imputed values are well within the range of the 'correct' value for sample sizes of 10 percent or more.

Figure E. Hotdeck Imputation: 5 Different Samples at Each Sample Size

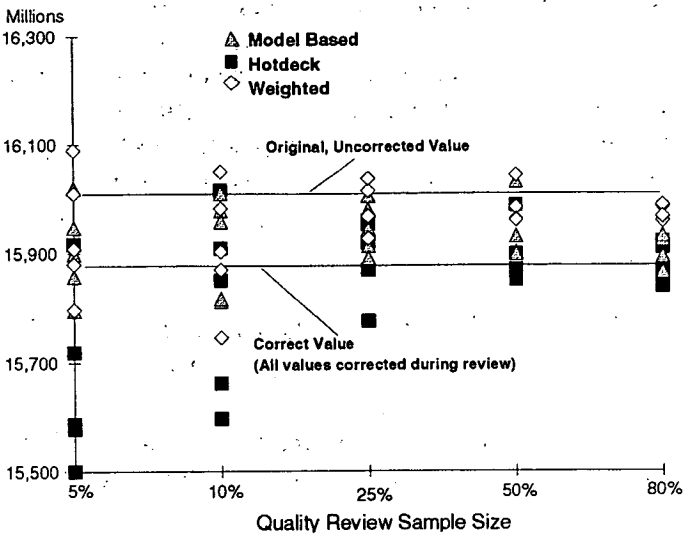


■ Conclusions

Editing errors for the variable examined proved to be random in nature and, thus, attempts to relate the occurrence and value of the errors to other data items using Logistic and OLS regression were unsuccessful. Thus, we resorted to two methods of imputation, one which used the normal distribution as the underlying model and the other employing hotdecking. Figure F shows a comparison of both the model-based and hotdeck imputation methods already seen in figures C and E, respectively, with point estimates derived from the QR sample weights for each of the 25 review samples previously examined. For sample sizes of at least 25 percent, estimates using imputed microdata were closer to the 'true' value than were the weighted estimates. In fact, the weighted estimates increased the bias more than either imputation method.

Of the two imputation methods examined, the more model-based approach is most attractive, largely because the uncertainty created by the model was substantially less than that introduced using hotdecking. The variance effect of the QR samples was similar between the two methods. The model-based approach also required fewer computer resources to implement than hotdecking.

Figure F. Weighted vs. Imputed Estimates



Hotdecking, however, was more effective in decreasing the bias of the final estimate. This preliminary investigation focused on errors in one variable, ignoring the interrelationships between fields on a record. It is possible that including these interrelationships in the constraints of the model may improve its performance.

Implications

The results of this very preliminary analysis suggest several issues which deserve further investigation. First, very small review sample sizes may not be sufficient to estimate the error remaining in a database. Our analysis focused on a data item which was present on most records--and in error a relatively large number of times, because it is often confused with other financial instruments, such as corporate bonds. It seems likely that it would be even more difficult to estimate the accuracy of data items which appear less frequently in the database. This suggests that quality review samples should emphasize key data elements and should be stratified to include sufficient numbers of records for sub-populations of particular interest.

Secondly, it is our practice to correct errors discovered during the quality review process before creating a final database for users. This practice, although well-intentioned, may be misguided. In 5 cases, or 20 percent of the time, correcting only the reviewed values resulted in a final database which produced estimates which were actually worse than the original, uncorrected values. The nature and importance of the bias which is introduced by this practice certainly bears further investigation.

Finally, we need to continue efforts to identify all of the factors relating to data capture errors. Other factors that we could not include in our study, such as worker experience, environmental factors, quality of the source document, instruction manuals, and training all have a bearing on data quality. We have been collecting some of these data for large errors and need to incorporate them into future research.

■ References

Aldrich, J.H. and Nelson, F.D. (1984). *Linear Probability, Logit, and Probit Models*. Newbury Park, CA: Sage.

Bailar, B.A. (1968). Recent Research in Reinterview Procedures. *Journal of the American Statistical Association*, 63: 41-63.

Biemer, P. and Forsman, G. (1992). On the Quality of Reinterview Data with Application to the Current Population Survey. *Journal of the American Statistical Association*, 87: 915-923.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

Hinkins, S. and Scheuren, F. (1986). Hotdeck Imputation Procedure Applied to a Double Sampling Design. *Survey Methodology*, 12: 181-196.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Non-Sampling Error in Surveys*. New York: John Wiley and Sons, Inc.

Little, R.J. (1990). Editing and Imputation of Multivariate Data: Issues and New Approaches. In Gunar E. Liepins and V.R.R. Uppuluri (Eds.). *Data Quality Control, Theory and Pragmatics*. New York: Marcel Dekker, Inc.

Little, R.J. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics*, 6, Number 3: 287-301.

McCubbin, J.G. (1990). The Intergenerational Wealth Study: Basic Estate Data, 1916-1945. *Statistics of Income Bulletin*, 9, Number 4: 79-114.

Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: John Wiley and Sons, Inc.

West, M. and Winkler, R.L. (1991). Database Error Trapping and Prediction. *Journal of the American Statistical Association*, 86: 987-996.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag Inc. ■