Developing an Optimal Approach to Account for Late-Filed Returns in Population Estimates

Cynthia Belmonte, Brian Raub, Paul Arnsberger, Charles Day[1]
[1]IRS, Statistics of Income, 1111 Constitution Ave NW, Washington DC 20024

Abstract
Estimates for populations of interest for Statistics of Income (SOI) programs are produced by drawing stratified, random Bernoulli samples of tax and information returns as they are filed, over predetermined sampling periods that often span multiple years. While this methodology results in the inclusion of the majority of targeted returns, a small number of returns for each study are filed beyond the data collection period, potentially introducing non-response bias into the population estimates. For a given sampling period, the paper will analyze historical filing patterns to develop an approach for accounting for late-filed returns. This research will assess the weight adjustment approach currently used in SOI's estate tax study and will provide a basis for application of a similar approach in each of the exempt organizations and private foundations studies.

Key Words: non-response bias, population estimates, post-stratification, Bernoulli sampling

1. Data Sources and Background

The Statistics of Income (SOI) division of the Internal Revenue Service (IRS) collects and disseminates detailed data based on samples of administrative records, including tax and information returns. The SOI sampling frame for any given study consists of tax or information returns posted to the appropriate IRS return transactions processing system within a designated time period. Often, this time period is the statutory period within which taxpayers are required to file. For other studies, in which taxpayers may file returns over many years, sampling occurs over a designated time period in which past experience tells SOI statisticians all but a small fraction of returns will be filed. In either event, some taxpayers may file returns for the period of interest after sampling for a study has ended. Over the years, SOI has taken several approaches to adjusting for the incompleteness of its sampling frames, some on a case-by-case basis and others more uniform in nature.

Building on previous research, this paper describes three SOI studies covering tax and information returns for estates, private foundations, and exempt organizations and briefly outlines current practices for handling late-filed returns [1]. Next, the authors describe two models for predicting the proportion of late-filed Estate tax returns using several covariates. Using the 2004 year-of-death sample, the authors will then apply and evaluate the new adjustment factors by comparing results to known population totals and previous estimates derived using existing adjustment factors.

1.1 The Estate Tax Study

With its annual Estate Tax study, SOI extracts demographic, financial, and asset data from Federal estate tax returns. The annual study allows production of a data file for each filing, or calendar, year. By focusing on a single year of death for a period of 3 filing years, the study allows production of periodic year-of-death estimates. A single year of death is examined for 3 years, as over 98 percent of all returns for decedents who die in a given year are filed by the end of the second calendar year following the year of death. Data included in this paper are for Year of Death 2004 and were obtained from returns filed in Calendar Years 2004-2006.

The estate of a decedent who, at death, owns assets valued in excess of the estate tax applicable exclusion amount, or filing threshold, must file a Federal estate tax return, *Form 706, U.S. Estate (and Generation-Skipping Transfer) Tax Return*. For decedents who died in 2004, the exclusion amount was $1.5 million. Alternate valuation may be elected only if the value of the estate, as well as the estate tax, is reduced between the date of death and the alternate date. The estate tax return is due 9 months from the date of the decedent's death, although a 6-month filing extension is allowed. In some cases, longer filing extensions may be permitted.

For the Year of Death 2004 Estate Tax study, there were 11,817 Form 706 returns in the sample selected from a population of 42,424. The SOI Estate Tax study is classified into strata based on year of death, the size of total gross estate, and age of the decedent. For the Year of Death 2004 study, there were a total of 57 sampling strata, with sampling rates ranging from 4 percent to 100 percent.

1.2 The Private Foundation and Exempt Organization Studies

The annual SOI studies of private foundations and exempt organizations collect detailed financial data, as well as information on charitable and grant-making activities and compliance with IRS regulations, from information returns filed by exempt organizations. Studies are conducted for a single tax year and include samples of returns filed and processed during the 2 calendar years immediately following the target tax year. Data discussed in this paper for the Private Foundation and Exempt Organization studies were obtained for Tax Year 2004 returns filed and processed to the IRS Business Masterfile during Calendar Years 2005 and 2006. While this 2-year sampling period ensures almost complete coverage of the target population, there are still a number of returns processed after the close of the second year (i.e., December 31, 2006 for the Tax Year 2004 study), which are generally excluded from the samples.

Private foundations and nonexempt charitable trusts are required to file Form 990-PF (*Return of Private Foundation or Section 4947(a)(1) Nonexempt Charitable Trust Treated as Private Foundation)* annually. Similarly, certain exempt organizations are required to file Forms 990 *(Return of Organization Exempt from Income Tax)* or Form 990-EZ *(Short Form Return of Organization Exempt from Income Tax)*. SOI conducts annual studies based on samples of Forms 990-PF, 990, and 990-EZ filed for a given tax year. These information returns are due 5 months after the close of the organization's accounting period, although a 3-month filing extension is allowed. In some cases, additional filing extensions may be granted.

For the Tax Year 2004 Private Foundation study, there were 7,805 Form 990-PF returns in the sample, selected from a population of 80,570. The SOI Private Foundation study is classified into strata based on the size of end-of-year fair market value of assets, with each stratum sampled at a different rate. Sampling rates ranged from 1 percent for private foundations with total assets less than $125,000 to 100 percent for private foundations with total assets of $10 million or more.

The Tax Year 2004 exempt organization sample of section 501(c)(3) filers comprised 15,070 Forms 990 and 990-EZ, selected from a population of 279,415. End-of-year book value of assets was the stratifying variable for the exempt organization study. Sampling rates ranged from 1 percent for exempt organizations with total assets less than $500,000, to 100 percent for those with total assets of $50 million or more.

2. Current Treatment of Late-Filed Returns

SOI's estate, private foundation, and exempt organization studies all share a common challenge in accounting for returns filed after the end of the designated sampling period. The Estate Tax study Year-of-Death estimates include weight adjustments for late-filed returns. Such adjustments were first developed in 1997 by Woodburn, and later updated in 2007 by Raub. Weight adjustment factors are calculated using historical data from the IRS Masterfile, and vary by size of estate, age of decedent, and tax status of return. The aim of using these weight adjustments is to improve the overall population estimates, as well as the estimates for the subpopulations of returns that have historically filed late with greater frequency. To the extent that late-filers create bias in the Estate tax estimates, this approach seems to be an effective strategy in mitigating this bias. Another strength of this approach is that the data used to calculate the adjustment factors are readily available in the IRS Masterfile.

In contrast to the estate tax study, population estimates for the private foundations study do not include standard adjustment factors to account for returns filed after the close of the 2-year sampling period. Instead, during file closeout, efforts are made to identify and include late-filed returns that would have been sampled at the 100-percent rate (i.e., organizations with fair market value of assets of $10 million or more). This allows for more complete coverage of the target population by including returns that would have been selected with certainty. This allows for time-series analysis of a specific organization (or panel of organizations). Potentially, this treatment can extend the two-year sampling period by 4 to 5 months, the typical length of time between the end of the normal sampling period (in December) and the creation of the final study file (in mid-May). This can introduce some inconsistency from year-to-year, since the slightest variation in the Master File processing cycle, file review schedule, or final delivery date can affect the sampling period from one year to the next. Additionally, this method does not specifically address smaller organizations, which account for the largest share of the late-filing population.

3. Methodology and Results

The goal of the current research is to determine whether the current estate tax study adjustment factors still accurately reflect taxpayer behavior. Additionally, the authors seek to develop and assess alternative methods of estimating adjustment factors on the estate tax study, and whether such methods can be applied to other studies (Private Foundations, Exempt Organizations) that are subject to similar late-filing challenges.

The authors propose adjusting the weights of the returns in the estate tax return sample by multiplying by the inverse of the predicted proportion of returns filed by the cutoff of sampling. In order to not overly inflate variance, it is desirable that a relatively small number of adjustments be applied to the returns. Rather than attempting to calculate an adjustment based on each return's values of selected covariates, the adjustment factors were calculated for specific categories that are either sampling strata, groups of strata, or subsets of a stratum. Such an adjustment accounts for returns that will be filed after the end of the sampling period for the estates of decedents who died during the reference year.

Discussions with the estate tax study analyst yielded three possible explanatory covariates: size of the estate (measured by the total gross estate value), age of the decedent, and taxability of the estate; that is, whether or not an estate tax was due before the application of credits. Taxability is naturally a categorical variable. While age is discrete, it can take on over 100 values, thus age categories, similar to the categories used in constructing sampling strata, were used as dummy variables, as were size categories. The categories were chosen to reflect marginal changes in late-filing behavior based on exploratory analysis. Precise category boundaries were then adjusted due to the desire to have them, when possible, match sampling stratum boundaries, and the need to have sufficient numbers of late-filing events in each cross classification (taxablilty × age × size) to support modeling.

3.1 Survival Analysis

Survival analysis, or time-to-event modeling, is a well-known technique for measuring the probability that some event (death in its original application) will occur within a given time period. It has been widely used since to model more general time-to-event problems. The survivor function estimates the probability of an event occurring at or after some time *t*. In this context, the event of interest is the filing of an estate tax return, and "survival" equates to making it to the end of the sampling period cut-off (3 years) without filing an estate tax return.

One method for forming such a model is Proportional Hazards (Cox) regression. Cox regression is a widely accepted type of survival analysis model. It allows the use of covariates to help explain differences in times to some event for different observations. For the estate tax study, age of the decedent and size of the estate are both important predictors of time to filing. Cox regression can also handle other important features of the estate tax study data.

In order for an estate to come into existence, someone must die. Prior to his or her death, and the formation of the estate, there is no risk of an estate return's being filed. SOI conducts a study of estates of decedents who die in every third year. Since the dates of death are distributed throughout the reference year, estates are formed and become subject to filing at different times. This is similar to a study of, say, cancer treatments, where subjects may enter the study at time of diagnosis and thus many subjects may become part of the study cohort at different times. The phenomenon of some subjects' beginning to experience positive probability of an event's occurring at a later time than others is called "delayed entry," and the observations for those subjects are referred to as "left-truncated." Cox regression can handle left-truncated observations.

Using Cox regression, the authors estimated the parameters of the survivor function conditional on the values of the covariates. For every adjustment stratum (shown in Table 1), the authors fit a model to the estate tax study year-of-death 2001 population data. In order to do this, the authors analyzed all of the possible combinations of the selected covariates for each stratum, keeping the best set of significant covariates for each stratum. The authors also used the year-of-death 2004 sample file to create a vector of all three covariates for each return. The authors then used the covariate vectors from the 2004 sample to predict a set of survival probabilities.

Table 1: Definition of Categories of Total Gross Estate and Age

Variable Name	*Lower Bound*		*Covariate*		*Upper Bound*
ageCats0	0	≤	Age	<	40
ageCats1	40	≤	Age	<	65
ageCats2	65	≤	Age	<	70
ageCats3	70	≤	Age	<	75
ageCats4	75		Age		or older
sizeCats0	$1.5 million	≤	Total Gross Estate	<	$2.0 million
sizeCats1	$2.0 million	≤	Total Gross Estate	<	$3.0 million
sizeCats2	$3.0 million	≤	Total Gross Estate	<	$5.0 million
sizeCats3	$5.0 million	≤	Total Gross Estate	<	$10.0 million
sizeCats4	$10.0 million		Total Gross Estate		or more

3.1.1 Survival Analysis Results

Table 2 presents new population estimates derived using the survival analysis approach as well as comparisons to known population totals and estimates using previous adjustment methods. The survival analysis model overestimated number of returns by about 6.5 percent and total gross estate by 10 percent.

Table 2: Year-of-Death 2004 Population Totals and Sample Estimates with Adjustment Factors Modeled Using Survival Analysis

Weight Adjustment Method	*Number of Returns*	*Percentage Difference*[1]	*Total Gross Estate ($ Millions)*	*Percentage Difference*[1]
Population total	**41,922**	**n.a.**	**149,430**	**n.a.**
Unadjusted estimate	40,453	-3.50	147,163	-1.52
Woodburn (1992)	40,785	-2.71	148,199	-0.82
Raub (2007)	40,867	-2.52	148,502	-0.62
Belmonte *et al.* (2010)	44,680	6.58	163,942	9.71

[1] Percent difference from known population total

The overestimation of both number of returns and total gross estate indicate that non-proportional hazards were not ignorable. The models were fit with time-dependent covariates to adjust for the effect of time on the effects of the different covariates. Many of the time-dependent covariates were highly significant. Also, their associated hazard ratios were greater than one, indicating that hazard, or risk, of filing increased as time passed. By ignoring the violation of proportional hazards, the hazards across time were essentially “averaged over”. This led to an underestimation of hazard, resulting in survival probabilities for late-filed returns higher than acceptable for the desired outcome.

3.2 Logistic Regression

Filing before or after the designated sampling cutoff can be modeled as a binary response variable. Logistic regression is a commonly used method for predicting the proportion of times an event occurs in a number of trials conditional on the values of some explanatory covariates [2, 3]. As in the previous model, the selected covariates were size of the estate (again, measured by the total gross estate value), age of the decedent, and taxability of the estate. Definitions of the selected categories are shown in Table 3.

Table 3: Definition of Categories of Total Gross Estate and Age

Variable Name	*Lower Bound*		*Covariate*		*Upper Bound*
ageCats0	0	≤	Age	<	40
ageCats1	40	≤	Age	<	65
ageCats2	65	≤	Age	<	70
ageCats3	70	≤	Age	<	75
ageCats4	75		Age		or older
sizeCats0	$2.0 million	≤	Total Gross Estate	<	$3.0 million
sizeCats1	$3.0 million	≤	Total Gross Estate	<	$5.0 million
sizeCats2	$5.0 million	≤	Total Gross Estate	<	$10.0 million
sizeCats3	$10.0 million		Total Gross Estate		or more

3.2.1 Logistic Regression Results

Table 4 shows the analysis of maximum likelihood estimates. All categories of all covariates are highly significant. Model development was guided in part by residual analysis, influence measures, and goodness-of-fit tests, but, as this paper is primarily concerned with good predictions and not explanation, these are omitted here.

Table 4: Analysis of Maximum Likelihood Estimates

Parameter[1]	*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > Chi-Sq*
Intercept	1	-2.4574	0.1316	348.8	< 0001
ageCats1	1	-0.5127	0.1311	15.3	< .0001
ageCats2	1	-0.7958	0.1385	33.0	< .0001
ageCats3	1	-0.9031	0.1356	44.3	< .0001
ageCats4	1	-1.3849	0.1286	116.0	< .0001
sizeCats1	1	-0.1191	0.0400	8.9	0.0029
sizeCats2	1	-0.2640	0.0525	25.3	< .0001
sizeCats3	1	-0.5813	0.0786	54.7	< .0001
Taxable	1	-0.1385	0.0413	11.2	0.0008

[1] The effect of the first category of each of the dummy variables for Age and Total Gross Estate is reflected in the Intercept.

Results from this method were quite good. Table 5 presents new population estimates derived using the logistic regression model as well as comparisons to known population totals and estimates using previous adjustment methods. This method produced an excellent estimate of total number of returns, the predicted value for which the method was designed. Additionally, the method resulted in a reasonable estimate of total gross estate.

Table 5: Year-of-Death 2004 Population Totals and Sample Estimates with Adjustment Factors Modeled Using Logistic Regression (for Returns with Total Gross Estate of $2.0 million and above)

Weight Adjustment Method	*Number of Returns*	*Percentage Difference*[1]	*Total Gross Estate ($ Millions)*	*Percentage Difference*[1]
Population total	**28,355**	**n.a.**	**161,007**	**n.a.**
Unadjusted estimate	27,701	-2.31	159,330	-1.04
Woodburn (1992)	27,926	-1.51	160,245	-0.47
Raub (2007)	27,981	-1.32	160,582	-0.26
Belmonte *et al.* (2010)	28,315	-0.14	162,213	0.75

[1]Percent difference from known population total

4. Future Steps

Estimates for the Estate Tax study benefit from a small adjustment to account for late-filed returns. As the research shows, logistic regression can be a useful method for calculating such adjustment factors. Results from logistic regression models are encouraging for the future development, assessment, and potential application of such models to adjust population estimates for other SOI studies. The authors recommend that efforts to develop similar models for each of the Private Foundation and Exempt Organization studies be undertaken as soon as possible.

5. Acknowledgements

The authors would like to acknowledge the management of the Statistics of Income Division, Internal Revenue Service, for their support and encouragement of this work. The views expressed in this paper are those of the authors and should not be attributed to the Internal Revenue Service or the U.S. Department of the Treasury.

6. References

[1] Raub, B., C. Belmonte, P. Arnsberger, M. Ludlum. The Effect of Late-Filed Returns on Population Estimates: A Comparative Analysis. In *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association, 2009.

[2] Agresti, A. *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New York, NY, 1996.

[3] Stokes, M., Davis, C., Koch, G. *Categorical Data Analysis Using the SAS System*, SAS Institute, Cary, NC, 1996.

The Effect of Content Errors on Bias and Nonsampling Variance in Estimates Derived From Samples of Administrative Records

Barry W. Johnson and Darien B. Jacobson
Barry W. Johnson, Statistics of Income RAS:S:SS, P.O. Box 2608, Washington, DC 20013-2608

Key words: Bias, Non-sampling error

The Statistics of Income Division (SOI) of the Internal Revenue Service (IRS) uses a number of methods for ensuring the quality and integrity of the data it produces for tax administration research. As a first line of quality assurance, codes and mathematically related data items are extensively tested as SOI employees enter them into computer databases. In addition, for a sub-sample of returns selected and processed in most studies, SOI assigns a second employee to reenter and edit the data. Values from the first and second edit are then computer-matched. A supervisor resolves discrepancies discovered during the match. The original value, second value, and correct values are all collected as a part of the quality review system, as are a set of codes that describe the cause of the error, in broad categories.

This paper will use quality review data from Federal estate tax returns (Form 706) selected into the Calendar Year 2002 SOI Estate Tax Study to estimate the effects of non-sampling error on estimates derived from the final data file.

Background

The Federal estate tax is levied on estates for the right to transfer assets from a decedent's estate to its beneficiaries; it is not an inheritance tax. A Federal estate tax return must be filed for every U.S. decedent whose gross estate, valued on the date of death, combined with certain lifetime gifts made by the decedent, equals or exceeds the filing threshold applicable for the decedent's year of death. A decedent's estate must file a return within 9 months of a decedent's death, but a 6-month extension is usually granted.

All of a decedent's assets, as well as the decedent's share of jointly owned and community property assets, are included in the gross estate for tax purposes and reported on Form 706. Also reported are most life insurance proceeds, property over which the decedent possessed a general power of appointment, and certain transfers made during life.

Expenses and losses incurred in the administration of the estate, funeral costs, and the decedent's debts are allowed as deductions against the estate for the purpose of calculating the tax liability. A deduction is allowed for the full value of bequests to the surviving spouse. Bequests to qualified charities are also fully deductible.

Data Description

The 2002 SOI Estate Tax Study was a stratified, random sample of returns filed in Calendar Year 2002 and was the second year in a 3-year study of Federal estate tax returns filed 2001-2003. The sample was designed for use in both estimating tax revenues in all 3 calendar years and personal wealth holdings for 2001 decedents. The 3-year sample period was devised to ensure that nearly all returns filed for 2001 decedents would be subjected to sampling, since a return could be filed up to 15 months after the decedent's death. The design had three stratification variables: size of total gross estate plus the value of most taxable gifts made during the decedent's life, age at death, and year of death. The year-of-death variable was separated into two categories, 2001 year of death and non-2001 year of death, in order to facilitate studies of 2001 decedents. Returns were chosen before audit examination and selected using a stratified random probability sampling method. A portion of the sample was selected because the ending digits of the decedents' Social Security Numbers (SSN) corresponded with those in the 1-percent Social Security Administration Continuous Work History Sample. However, the majority of returns were selected on a flow basis using the Bernoulli sampling method.

The sampling mechanism was a permanent random number based on an encryption of the decedent's SSN. Sample rates were preset based on the desired sample size and an estimate of the population. Sampling rates ranged from 3 to 100 percent, with more than half of the strata selected with certainty.

Data collection for the 2002 Estate Tax Study was conducted at the IRS Cincinnati Submission Processing Center. Employees entered the data from the estate tax return into a database using a Graphical User Interface (GUI) data entry system. Nearly 100 distinct data items were captured, with some balance sheet assets recurring hundreds, even thousands, of times, as assets were allocated to 32 different categories, such as stocks, bonds, and real estate. Tax returns ranged in size from a dozen to many thousands of pages, including appraisals, investment account listings, and legal documents. Tests embedded in the data entry system were used to validate entries and to ensure that mathematical relationships among variables were correctly preserved. There were more than 200 validation tests performed on each tax return included in the 2002 study.

While embedded testing can assure that codes are correct within a given range of values and that fields are mathematically consistent, many of the decisions that employees make when transforming tax return information into statistically usable data are not easily tested. For example, while several codes may be valid, determining the best code to describe a particular taxpayer's behavior or characteristics cannot always be automated. To address this problem, SOI developed a double entry quality review system. This system is a valuable tool for measuring both individual employee performance and overall data quality.

Quality Review System

A subsample of returns in the 2002 Estate Tax Study was subjected to additional review for quality assurance purposes. Returns were included in the quality review (QR) subsample through two different mechanisms, 100-percent review and product review. The 100-percent review consisted of all returns that were edited while an employee was in training. Product review was selected after the training period had been completed, and it comprised a 10-percent random sample of each employee's work. The product review sample was selected on a flow basis method using a pseudorandom number called the Transform Taxpayer Identification Number, or TTIN. The TTIN is a unique random number that is generated by mathematically transforming selected digits of the decedent's Social Security Number. The TTIN was then compared to the sample number, which represented the sample rate, in this case 10 percent. If the TTIN was less than the sample number, then the return was selected for product review.

Under the double-entry quality review system, one return was entered into the computer system twice by two different employees. The first employee did not know that a return was selected for review until after the first edit was complete, and the second employee was not allowed to see the first employee's entries. Therefore, each return had two versions in the database, the first edit and the second edit, and each was entered independently of the other.

When both employees finished editing a return, the computer compared the values from the original and QR versions. In some cases, the two versions matched perfectly; so, the return was released from the system, and the first edit data was treated as final and stored for later analysis. However, if mismatches between the two versions occurred, the discrepancies were stored in a separate data table to be reviewed by a supervisor.

The supervisor reviewed the discrepancies and charged the errors, assigning two codes to each discrepancy--one to identify the incorrect value and the other to describe the cause of the error. A discrepancy code was assigned to the error to explain which version was considered incorrect. Discrepancy codes were assigned to one of the following: the first version, the second version, both versions, or neither version. An error was assigned to both versions if both of the employees entered or interpreted the information from the return incorrectly. In this case, the supervisor was also required to supply the correct data value. In some cases an error was not assigned to either version, usually when the discrepancy was the result of a data processing peculiarity and not a true database error. After the error was assigned a discrepancy code, a numeric error resolution code was assigned to describe why the entry was incorrect. Error resolution codes indicate situations such as spelling errors, incorrect money amounts, or incorrectly assigned codes.

Once the supervisor reviewed all the discrepancies, each employee was given a list of the discrepancies, along with the discrepancy and error resolution codes, so that any first edit errors detected during quality review could be corrected prior to considering return processing complete. The feedback from the review also enabled employees to learn from their mistakes on each return and carry this knowledge into the editing of other returns. In the end, there is a database consisting of a table that includes all the values from the second edit of the return as entered, a quality review table containing a record of each discrepancy between the first and second edits (along with codes indicating who made the error and why), and a final data table containing the correct version of the return data that will ultimately be sent to customers.

For this paper, only a portion of the quality review data was used for analysis. First, data that were collected during periods of training, 100 percent review, were excluded. Second, only errors that were charged to the first edit or to both edits, meaning that the error required a correction to the final data set, were retained. This was done because these errors are more representative of errors that remain in the roughly 90 percent of the 2002 estate tax sample that was not selected for quality review. Third, errors that reflected idiosyncrasies related to the edit process itself, and not true data errors, were eliminated.

Empirical Results

Quarterly accuracy rates for each employee who worked on the Estate Tax Study for 2002 were generated using the product review data (see Figure 1). These rates were calculated using the number of returns that had at least one error charged to the first edit divided by the total number of returns that had been selected for quality review. The accuracy rates for all of the employees are not very high. However, these rates are a return level measure; any return with one or more errors is considered incorrect. The Form 706 includes an average of 150 data entry fields, while complex returns can have more than a thousand entries; so, the probability of making just one mistake is very high. In fact, the average number of errors for each return is only 6.3.

Traditionally, supervisors have focused quality improvement efforts on those fields that are in error most frequently. By looking at the occurrence of variables *ex-ante*, using the first edit data, and *ex-post*, using the final corrected data file, it is possible to identify the frequency of original edit errors in the quality review sample. Figure 2 shows the percent changes in frequencies for variables on the file; each diamond represents a different variable. Frequencies change because many variables on the file represent balance sheet items, assets like stocks, bonds, mutual funds, and various types of real estate, which are not necessarily present in each decedent's portfolio. When an asset is incorrectly classified, not only does it change the dollar value of estimate, it also changes the frequency of occurrence of that particular attribute or asset type in the population estimates. This can be particularly problematic if the asset is of special interest to researchers. For example, there has been much discussion in the press about providing estate tax relief to small business owners. Errors that either under- or overcount the number of estates that have small businesses could have an impact on this debate. The percentages shown on the graph represent the aggregate correct frequency in the overall quality review sample, less the aggregate number originally reported, divided by the correct number. Negative percentages indicate cases where an asset was incorrectly included on the first edit. For example, the first employee may have incorrectly classified a balance sheet entry as a publicly traded stock, while the second employee may have correctly classified it as a mutual fund invested in a mix of financial assets. The percent changes in frequencies are generally close to zero, but there are some notable outliers.

Figure 1: Employee Accuracy Rates

	Accuracy Rates			
Employee	Quarter 1	Quarter 2	Quarter 3	Quarter 4
17000	46.3%	23.9%	41.7%	21.7%
17100	25.0%	0.0%	0.0%	0.0%
17200	29.2%	30.8%	31.9%	40.0%
17300	57.1%	100.0%	91.7%	33.3%
17400	52.1%	28.6%	50.0%	37.9%
17500	44.4%	24.1%	54.8%	0.0%
17600	42.2%	51.9%	33.9%	46.2%
17700	41.9%	28.6%	39.3%	34.5%
17800	49.1%	25.0%	58.5%	45.6%
17900	52.3%	34.3%	59.0%	50.0%
17001	23.1%	34.2%	18.6%	44.7%
17002	39.2%	33.3%	36.2%	45.0%
17003	22.9%	20.7%	37.8%	29.1%
17004	34.2%	31.6%	22.0%	72.7%
17005	30.8%	0.0%	0.0%	37.9%
17006	26.5%	27.7%	41.4%	42.9%

Figure 2: Percent Change in Frequencies, Original and Final Edits

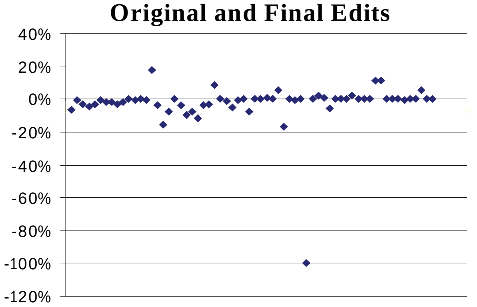
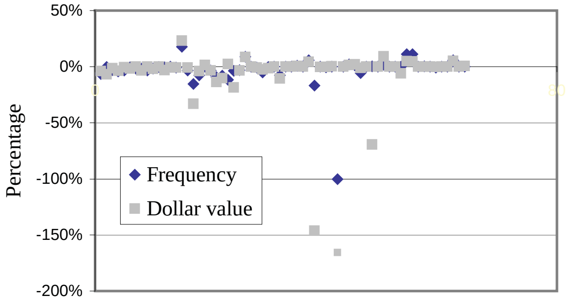


Figure 3 shows percentage changes in dollar amounts between first and second edits overlaid on the frequency differences shown in Figure 2. Each point represents a single variable on the file. While the pattern for the dollar differences is similar to that of the frequencies, with many differences close to zero, the magnitude of the dollar differences is larger for several variables. There are two variables for which the original entries resulted in aggregate dollar values that were overstated by roughly 150 percent. This highlights the potentially large effects on final estimates that can arise from even one large dollar value error, especially for variables that are not widely distributed in the overall population. Thus, it is important to monitor both the size and frequency of data entry errors.

Figure 3: Percent Change in Dollar and Frequency Values, Original and Final Edits



Unweighted error statistics are clearly useful for monitoring data quality and assessing opportunities for operational improvements during a study period. However, since the SOI study of Federal estate tax returns is based on a stratified random sample of the filing population, the effect of data entry error on final population estimates derived from this sample will vary inversely with the selection rate associated with each return. Using appropriate sample weights, it is possible to use the 10-percent QR sample to estimate the effects of data entry errors on population estimates derived from the remaining 90 percent of the returns in the final

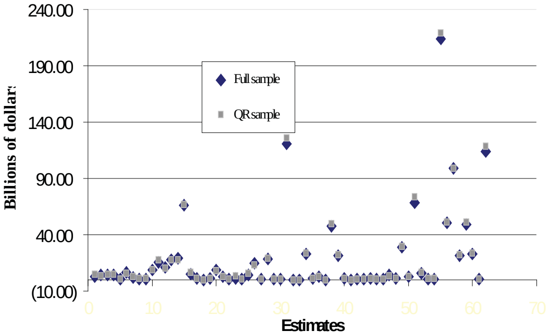
SOI data file that were not subjected to double-entry quality review. Weighted estimates provide a different perspective on the effects of nonsampling error due to the nature of the underlying estate study sample and the fact that the financial characteristics of estate tax decedents vary greatly among age and wealth classes. For example, younger decedents and those with large estates are selected into the estate tax sample with certainty and comprise more than 40 percent of the total sample file. Both groups of decedents are more likely to have had portfolios that are more complex and, thus, more subject to data entry errors than their either less wealthy, or older, cohorts. This is because many older wealth holders convert their portfolios to assets that produce tax-preferred income, usually resulting in returns that contain fewer business arrangements, which are more difficult to classify than market assets. Because the quality review sample is not stratified, weighted estimates will provide a more balanced measure of the overall effects of data entry errors on final estimates. Weighted estimates for the quality review sample were generated by using the design-based weight from the stratified estate study sample (W_s), multiplied by a quality review weight (W_q). The quality review weight itself was developed by first post-stratifying the quality review samples within the original selection strata as indicated below[1]:

$$\textbf{Final Weight} = \quad W_s * W_q$$
$$\textbf{Where } W_s = N_i/n_i$$
$$\textbf{Post-Stratification: } W_q = n_{if}/n_{qif}$$

For some strata, the quality review sample was either zero or too small to create a post-strata cell. For these cases, strata were collapsed across age categories so that estate size classes were preserved.

Figure 4 shows full population dollar value estimates from the quality review data using the post-stratified quality review weight and compares them to population estimates using the full weighted estate study sample. Each pair of data points represents a different variable on the file. The quality review data estimates for each variable are denoted by the gray squares, and the full sample estimates are denoted by the black diamonds. For most variables, the QR sample estimates are larger than the population estimates from the full estate sample, indicating that the QR sample introduces a positive bias. This bias arises because the QR sample is a simple random sample of a stratified sample that favors large dollar value returns. In such cases, ratio raking can often be employed to decrease the bias; however, in this case, the QR sample size was insufficient in the lower gross estate size classes.

Figure 4: Full Sample vs. QR Sample Estimates



While the weighted QR data estimates are somewhat biased due to the design of the sample, they still provide an important indication of the effects of data entry errors on final estate tax sample estimates. Figure 5 shows weighted and unweighted estimates of aggregate differences between original and final values of both frequency and dollar value estimates for selected variables. A negative value means that a variable was over represented in the original, uncorrected data, and a positive value means it was originally underrepresented. Weighted results rank errors differently for some of the variables. For example, errors in classifying noncorporate business assets had a much greater impact on final weighted estimates than would have been evident had the analysis been limited to examining the unweighted QR data. Conversely, the unweighted QR data implied that the effects of errors on estimates of farm real estate were greater than they are in the final, weighted estimates. Clearly, using weighted estimates, along with the unweighted quality review data, provides a more

Figure 5: Differences between First and Final Edits

Data Element	Frequency	Dollar Value
Noncorporate Businesses	**-11.00%** *-5.29%*	**-5.79%** *-3.55%*
Closely held stock	**-3.06%** *-3.42%*	**-1.01%** *-0.71%*
Real estate	**6.70%** *6.82%*	**7.34%** *6.17%*
Farm land	**-0.91%** *-1.95%*	**-1.09%** *-3.66%*
Funeral expenses	**0.25%** *0.09%*	**0.15%** *0.04%*

Values in *italics* are unweighted estimates

[1] The subscript "if" signifies that certain reject returns were removed from the estate study sample prior to post-stratifying.

balanced method of assessing where to focus data quality improvement efforts.

Figure 6 compares the weighted percent differences between original edit estimates and final, corrected estimates with coefficients of variation (C.V.) from the full estate tax study sample in order to relate the sampling and nonsampling variances associated with selected fields. For some estimates, such as the values for noncorporate businesses and publicly traded corporations, the nonsampling error attributable to data entry is much greater than the sampling variance. For others, such as estimates of stock in closely held or untraded corporations and farm land, the sampling error, represented by the C.V., is actually greater than the nonsampling error attributable to data entry errors, indicating that data entry errors are not a significant cause of additional variance in the estimates. Fields for which nonsampling error is relatively large provide opportunities for future data quality improvement efforts.

Figure 6: Data Entry Error vs. Sample Variance

	Frequency		**Money Amount**	
Data Element	**% diff**	**C.V.**	**% diff**	**C.V.**
Non-corporate businesses	-11.00%	4.45%	-5.79%	3.89%
Publicly traded stock	15.02%	.78%	20.00%	1.17%
Closely held stock	-3.06%	3.47%	-1.01%	2.18%
Real estate	6.70%	1.92%	7.34%	2.19%
Farm land	-.91%	4.34%	-1.09%	4.68%
Funeral expenses	.25%	.57%	.15%	1.19%
Spousal trusts	4.25%	2.97%	1.29%	1.58%

Conclusion

There is much to be learned through careful analysis of the data generated by SOI's double-entry quality review systems. The results of these analyses can be used to improve data collection systems and enhance worker training. Information on nonsampling error should also be useful to data users who could use data quality metrics to more accurately interpret economic modeling results and to ultimately build models that are more robust.

This analysis, however, revealed that the database format and the type of data that are collected from the quality review samples make certain types of analysis difficult, if not impossible. While a complete copy of the second edit is saved for all QR returns, the original, uncorrected first edit values are not saved when first edit errors require corrections. Information on discrepancies is kept in all cases, but, because corrections can involve changing any number of related fields, it is difficult to reconstruct exactly the first employee's original entries. If more sophisticated analysis is desired, including the study of secondary errors that arise as a result of a primary data entry error, archiving a complete copy of the first edit, along with associated error reason and discrepancy codes, should be considered.

It is also important that supervisors apply error reason and discrepancy codes consistently. All too often, discrepancies are resolved by several different supervisors. Some, especially those serving in a temporary capacity, may feel a great deal of peer pressure to avoid assigning errors to individual employees, even in cases where the assignment of an error would not directly impact employee performance appraisals, such as when an error is attributable to lack of clarity in editing instructions. This inconsistency makes it difficult to measure the extent to which errors exist and to learn of ways to avoid them in the future.

Related to this problem is that the measure of employee performance currently in place is not adequate. It is simply unfair to use a return level measure of accuracy when the difficulty of the work is so variable across returns. A more balanced measure would relate the number of individual errors an employee makes to the number of fields he or she actually edited, thus giving full consideration to the number of edit decisions that were made on each return.

Finally, there are sample design issues that became apparent from this analysis. The QR sample is biased and could be improved by taking into consideration the underlying structure of the estate tax study sample design. Even this would not provide coverage of variables that are relatively rare, but perhaps important, in policy debates. To address this problem, samples could either be increased or targeted to include more returns with important characteristics, such as those filed for small business owners, or returns that, because of the types of entries made during first edit, are more likely to contain significant problems. Samples could also vary with worker skill levels. One possibility would be to develop a system that sets a weekly QR sample rate for each individual employee based on individual rolling average accuracy rates. Sample rates could be set automatically based on preset performance standards. Automating the process would avoid putting supervisors in the awkward position of having to 'punish' poor performers with additional oversight, making it easier to match feedback and training efforts to performance levels.