Comparing Advance and Final Estimates: 1990 SOI Corporate Sample

John Czajka, Mathematica Policy Research, Inc. and
Susan Hinkins, Internal Revenue Service

The Statistics of Income Division (SOI), within the Internal Revenue Service, provides estimates of financial items of interest to economists and policy makers, based on annual samples of tax returns. Because these data are used to analyze the effects of current tax policy, to estimate effects of proposed policy changes, and to measure and analyze the U.S. economy, timeliness is important for the efficacy of these data.

Unfortunately, timeliness is a problem with administrative data based on tax returns. The 1990 corporate tax returns were generally filed between June 1990 and March 1992. Sample selection occurs after the basic tax data have been verified and entered onto an electronic data file. Therefore, the 1990 sample selection was not complete until July 1992. After selection, the return must be retrieved and the detailed information collected and checked. The final 1990 estimates were not complete until January 1993.

The SOI Division has an ultimate goal to provide the users with estimates on demand, or a continuum of data over time. This requires major changes in our sampling perspective and in estimation techniques. In particular, increased use of model-based estimates will be required. As a first step, SOI provides advance data estimates before the sample is complete. Preliminary 1990 estimates were provided May 1,1992.

■ Preliminary Data

The final estimates are design-based, post-stratified estimates of population and subpopulation totals:

$$\sum_h \frac{N_h}{n_h} \sum_1^{n_h} x_{hi}$$

where h indicates the stratum, and N_h and n_h are the population and sample sizes, respectively, for stratum h.

The preliminary 1990 data file was defined as the sample selected by January 1,1992, with a few exceptions as noted later. The corresponding simple weighted estimate for the preliminary file would be

$$\sum_h \frac{\hat{N}_h}{m_h} \sum_1^{m_h} x_{hi}$$

where $m_h \leq n_h$ sample units are present in stratum h in the preliminary file. The population sizes N_h are not known but must be estimated. Using this estimate assumes that, within strata, the preliminary sample is a random subsample of the final sample. It does not require subject matter expertise to doubt this model.

For many variables, the corporate population is very skewed; a relatively few of the largest corporations contain most of the total dollar amounts. Therefore, large corporations are selected into the sample with certainty, and these large corporations now make up almost a third of the total sample. Unfortunately these large corporations are more likely to be in the latter part of the sample, and not in the preliminary sample. This can be seen in Table 1, which shows some properties of the preliminary sample, selected by Jan. 1,1992. This represents 90% of the final 1990 sample. Large corporations are defined here as having more than $50 million in total assets.

Table 1. The Advance Sample Represents:

	Overall	For Banks
Total N	94%	97%
# of Large Corps.	88%	98%
Total Amounts ($)		
Interest income	85%	84%
Total assets	86%	90%
Long term capital gains	85%	90%
Loss	80%	77%

Approximately 700 of the very largest corporations are designated as critical cases. Since these largest cor-

porations are so crucial to the estimates and can be extremely variable from year to year, some current information is needed in the preliminary estimates (Hinkins & Mulrow). Therefore, if the return is not selected in the preliminary sample, extra measures are used to obtain the information. This includes sending a short questionnaire directly to the corporation, requesting approximately 20 items from the tax return.

There are other mechanisms affecting the properties of preliminary data. Corporations using extensions on the time to file will be more likely to be missed in the advance sample. Table 1 indicates that (large) corporations with a loss are more likely to file later in the filing period and be missed in the preliminary sample. Looking at the second column of Table 1, there is also an apparent industry effect. Banks are less likely to be "late" and large banks do not appear to be more likely to be late than small banks, at least with this definition of large. Again, however, banks with large losses are seriously underrepresented in the preliminary sample.

Because the "late" returns are not like the "advance" returns, the properties of the late returns need to be modeled. The first models used have been simple ratio adjustments based on prior year results. Ratio adjustments were used for only the 29 items considered most important by our primary user of the preliminary estimates -- the Bureau of Economic Analysis. Since there appears to be an industry effect, these ratios are calculated by industry groupings. The simple weighted 1990 advance data estimate is adjusted by multiplying each weighted industry estimate by the corresponding ratio of the final 1989 estimate to a simulated preliminary 1989 estimate. This model assumes, for example, that if the simple weighted advance estimate of loss for banks underestimated the final amount in 1989, then the 1990 weighted advance estimate will also underestimate the 1990 final, by approximately the same ratio. In order to incorporate some indication of the variability of this ratio over time, a second ratio estimator was calculated, using the average of the 1989 and the 1988 ratios.

■ Comparing Advance Estimates to Final

In this section, the 1990 advance estimates are compared to the final estimates for the 29 variables that were ratio adjusted. The relative error is measured as the difference between the advance estimate and the final estimate, as a percentage of the final. The final estimate is always the post-stratified estimate calculated from the final file.

The item Orphan Drug Credit is relatively rare. The relative error of the simple weighted advance estimate is 6%, but when rounded to millions, as published, it is 0%. When rounded to millions, the ratio adjusted advance estimate also has a relative error of 0%.

Similarly the item Nonconventional Fuel Source Tax Credit is relatively rare, but more volatile. The relative error in the simple weighted advance estimate is only 3%. The ratio adjusted estimate using 1989 information is drastically worse, due primarily to the variability of the ratio for one of the influential industries. The ratios of the final weighted estimate to the advance weighted estimate over the three years, for this one industry, were:

1988	1989	1990
.90	2.00	.98.

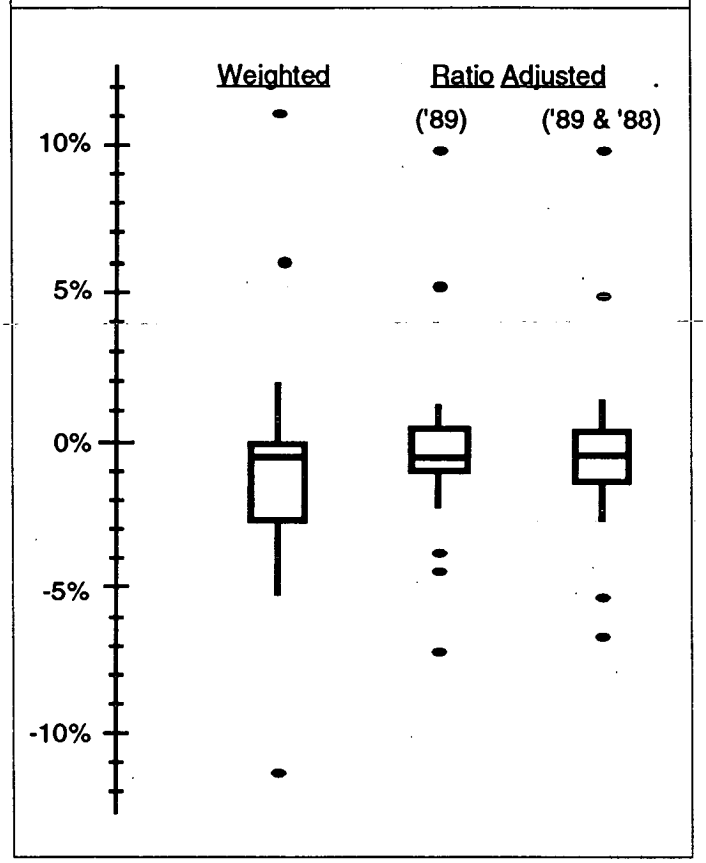
There is large variation in this item. Using the 1989 ratio in this case is dramatically worse than using the 1988 ratio or making no adjustment. Of course, we would not have known this when we applied the ratio adjustment to the 1990 advance estimate.

Figure 1 shows box and whisker plots of the relative errors in the 1990 advance data estimates for the remaining 27 variables. The first box plot shows the simple weighted estimates, before the ratio adjustment. These estimates generally underestimate the final totals.

The second plot shows the errors for the ratio adjusted estimates, using the 1989 ratios. The ratio adjustment noticeably improves the advance estimates overall. There are still some large errors in estimates of totals, however. The third plot shows the relative errors using the average of the 1988 and 1989 ratios. This adjustment is also an improvement over the simple weighted estimate.

Table 2 summarizes the relative errors for the six outliers in Figure 1. The simple weighted estimates for two of these items were essentially unchanged by using the 1989 ratio adjustments. Two other items, Possessions Tax Credit and Loss, were greatly improved by using the ratio adjustment. The advance estimates of the items Minimum Tax Credit and Short Term Capital Gains were

Figure 1. Distribution of Relative Errors in 1990 Advance Estimates of Totals



noticeably worse using the 1989 ratio adjustments. Using the average of the 1988 and 1989 ratios, the relative error in Short Term Capital Gains dropped to -2.5%. For this variable, the 1988 ratios were better predictors of the 1990 results.

The first five variables, outliers in the 1989 ratio adjusted errors, will be looked at in more detail in the following section.

■ Components of Error

Two types of errors in the preliminary estimates are evaluated: 1)the error due to the estimation of the missing corporations (modeling error), and 2)nonsampling errors, due to errors in the advance data.

In the first year of the new system, there were some problems in the preliminary estimation that should not occur in the future. First, the 1990 advance sample was smaller than expected. The preliminary data should have included all corporations selected by the end of 1991, plus the critical cases. Because of computer and processing problems, the sampled records were not entered and cleaned as rapidly as expected. Therefore the 1990 preliminary sample included only 80%, rather than 90%, of the final sample. While we do not expect the filing patterns to be exactly the same over the years, this smaller than expected preliminary sample could have reduced the effectiveness of the ratio adjustments.

The nonsampling errors were also larger than expected. Several procedures were not in place or were not error free at the time of the preliminary estimation. For example, this resulted in 62 records being left in the preliminary file that were rejected in the final file. Industrial classifications were not completely checked, and so some records were not correctly classified in the preliminary file.

Noticeable nonsampling errors were caused by missing items on the critical cases. For some critical cases, the only available 1990 data came from the short questionnaire, requesting only 20 items. The missing items were not imputed but were left as zero. Also, corporations did not always respond to the questionnaire; these missing records were estimated by substituting the prior year data. While there were only 23 critical corporations with missing or partial data, they made an impact on the advance estimates of totals.

The two major components of error are measured as follows. An original advance data estimate, Ao, either as a simple weighted estimate or with a ratio adjustment, is the estimate as calculated in May 1992, including all the processing errors. The cleaned advance data estimate, Ac, uses the correct, final amounts for these advance records, the final industry classifications, etc. The 62 rejected records are not included in Ac. The final estimate, F, uses the complete final sample, with post-stratified final weights. The nonsampling (relative) error in the advance estimate can be measured by:

$$(Ao - Ac) / F$$

The modeling error can be measured by

$$(Ac - F) / F$$

Table 2. Relative Errors in the Advance Estimates - Outliers

Item	Weighted Estimate Total	1989 Ratio Adjusted Estimate		
		Total Error	Modeling Error	Nonsampling Error
Cash and property distribution	11.1%	9.8%	- 2.4%	12.2%
Depletion	-4.7%	-4.4%	-2.3%	-2.1% (-1.9)
Loss	-11.3%	-3.8%	-3.8%	0.0%
Minimum tax credit	1.7%	5.7%	6.0%	-0.3%
Short term capital gains	-5.2%	-7.2%	-6.2%	-1.0% (-1.0)
Possessions tax credit	6.0%	-1.2%	-1.3%	0.1%

These two components add up to the total relative error in the original advance data estimate: (Ao - F) / F. The last two columns in Table 2 show these components of error for the ratio adjusted advance estimates.

Nonsampling Error

There were 65,702 corporations that appeared in both the advance file and the final file. For 65,507 of these, there was no new information gathered between the advance and the final file.

There were 195 corporations that had different information available in the advance file than in the final file. The most important of these were the 23 critical cases that were unavailable in time for the advance file. Twenty-one of these had partial 1990 information available primarily from the corporation's response to the short questionnaire. Two critical cases had no current year information; they were estimated by their 1989 records. The complete 1990 information was available for all but 2 of these records in the final file. The advance data records for the remaining 172 corporations were generally replaced by more current or more complete information in the final file (a full year return replacing a part year return, for example).

The last column in Table 2 shows the nonsampling error in the ratio adjusted advance estimates for the 5 outliers. The percentages in parenthesis show the contribution due to incomplete information on the 23 critical cases.

Three out of the five outliers had noticeable non-sampling error. The most significant, Cash and Property Distribution, is a special case. This nonsampling error was not due to the new process of creating an advance file, nor to incomplete critical cases; instead it reflects revisions that have always occurred during the final review. This process needs to be reconsidered.

For the variable Depletion, the nonsampling error contributed almost half of the total error. And for both Depletion and Short Term Capital Gains (STCG), the nonsampling error was primarily due to the incomplete critical cases. The error due to missing information on critical cases in the item Depletion (-1.9%) was essentially due to only 4 records. The remaining nonsampling error (-0.2%) was due to errors in the industry classification. The nonsampling error in STCG (-1.0%) was almost entirely due to missing information on 8 critical cases.

If these incomplete critical cases noticeably affected the estimates of totals, they obviously had significant effects on their associated industry estimates. For example, the modeling error for the advance estimate of STCG in Insurance companies was only 0.1%, but the nonsampling error due to the incomplete critical cases was -4.5%. Similarly for the item Depletion, in three industrial categories, one incomplete critical case in each industry resulted in relative errors of -8%, -10%, and -13% in the total industry estimates. These are truly critical cases.

Incorrect industrial classifications in the advance file did not appear to significantly affect the advance estimates of totals. But such errors can certainly affect the industry estimates. For one major industry, the relative error in the advance estimate of Depletion was all nonsampling error: -6.6% due to incomplete critical cases and -9.6% due to incorrect industrial classification. (The model error was only 0.4%.) The class most affected by incorrect industrial classification was "Holding Companies," and the most extreme error was an over-estimate of 396%.

Model Error

There were 16,856 "late" records that appeared in the final file but not the advance file. The model error represents how well (or how poorly) these missing records were estimated. For a very simple model, the ratio adjustments did fairly well. Generally the ratios move the weighted advance estimate in the right direction, but not always the right distance.

The outliers shown in Table 2 indicate that there can be considerable variation in the ratios over time. The ratio adjustments for the variable Loss move the advance estimates in the correct direction, just not far enough. The two largest modeling errors, in Minimum Tax Credit and in STCG, are due to directional changes in the ratios for influential industries. For Minimum Tax Credit, the ratio of the final to the advance weighted estimate for one major industry changed from 1.08 in 1989 to 0.99 in 1990, resulting in over-estimating the 1990 final amount. For Short Term Capital Gains, an influential ratio, changed from .95 in 1989 to 1.22 in 1990.

■ Conclusions

The ratio adjustments, especially in the absence of nonsampling error, significantly improve the advance estimates, and give some indication of the possible variability in these estimates. However, the ratio adjustment was considered a temporary measure. It is unwieldy to compute and store a separate ratio for each variable. Ratio adjustments were made for only 29 variables and there are almost 100 primary variables of interest. We hope to improve on the preliminary estimation by adjusting the advance weights using estimates of the propensity to be in the preliminary sample, based on prior year results. The fact that the simple ratio adjustments perform well is an indication that better models may be feasible.

However, we found that prior to improving the modeling component of the advance estimate, the nonsampling error must be reduced. In particular, emphasis has been given to imputing the missing items on critical cases and correcting industrial classification in the advance data. The latter should have been corrected for the 1991 advance data; we can soon check this. The imputation of the incomplete critical cases is in progress and should be in place for the 1992 file.

■ Bibliography

Czajka, J., Hirabayashi, S., Little, R., and Rubin, D. (1992), "Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics," *Journal of Business and Economic Statistics*, vol. 10.

Hinkins, S. and Mulrow, J. (1992), "Preliminary Estimates from the 1990 SOI Corporate Sample," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Rosenbaum, P. and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, pp. 41-55. ■