AUTOMATIC AND MANUAL EDITS FOR CORPORATION INCOME TAX RETURNS

Karen Cys, Susan Hinkins, and Victor Rehula, Internal Revenue Service

The purpose of this paper is to outline a major change in the method used in the Corporation Statistics of Income Program to transfer raw data from corporation income tax returns to magnetic tapes for the purpose of producing annual statistics required by tax law. The statistics are used by the Department of the Treasury and Congress to analyze existing and proposed tax laws and by others, both inside and outside the government, to analyze economic and financial data.

Organizationally, the paper is divided into three parts. Part one provides an historic overview of the corporate statistics program and describes the manual process of abstracting and transcribing selected corporate data onto documents known as edit sheets. The transcribing of the data using complex and specialized sets of instructions for the different types of income tax returns is known as statistical editing. Part two discusses some recent improvements in the statistical editing procedures, a system of automatic and computer-assisted editing, which will provide more complete statistical information at a reduced cost. Part three provides a brief look at our plans for the future.

BACKGROUND

Since 1916, raw data have been abstracted from the nation's corporation income tax returns in order to comply with the newly enacted tax law. This tax law required an annual publication of tax return data [1]. Since those early years, very little basic change in the method of abstracting has occurred. Currently, we are still picking up data from the income tax return and entering it on edit sheets with pencil in hand. We have made some progress though. For 1916 we edited each of 341,253 returns that were filed by the nation's corporations. Beginning with 1951, a probability sample was used as a basis for data tabulated. Today, however, while the number of corporation returns filed has grown to 2.9 million, we are to edit only a sample totalling approximately 95,000 returns. Also, beginning in 1981 tax year, the abstracting of the data was changed from a total manual operation by large groups of editors using adding machines to a partial computer operation.

Although the number of returns has been reduced from those early years because of sampling, the total workload has increased enormously. Due to the greater financial detail needed by the Treasury Department's tax analysts, legislators, and other users of our data, we are required to edit more information from each return. Of course, the tax legislation over the years has added much more detail to the return as well.

For the 1981 Statistics of Income (SOI) program, we are picking up 395 different money amounts and some 85 codes used to classify, indicate content, or identify processes. In contrast, for 1916, only 4 money amounts and a single code, the industry code, were used. The editing process is also complicated due to the increase in the number of forms and schedules. In 1916, there was a single return form for corporations and no attached schedules. Currently, there are six return forms for each of the different classifications of corporations ranging from the basic 1120 return form usable by most corporations to Form 1120F for foreign corporations doing business in the United States. Also, there are now 11 schedules or forms from which we extract data. Schedule D, on which is reported capital gains and losses, is one example and the more recent Form 6793, the Safe Harbor Lease Information Return, is another.

For 1916, the statistics for corporations reported only four money amounts from the return: gross income, total deductions, net income or deficit, and tax. There were four tables, each showing number of returns and the above amounts. The classifications for the tables were by industries, or states and corporations showing net income and corporations showing no net income.

During the early years, statistical editing for Statistics of Income (SOI) purposes was done at the National Office in Washington, D.C. During the early 1960's, the editing of the returns for SOI purposes was transferred from the National Office to the service centers. As the computer age dawned and flourished, some of the editing of the smaller asset size returns was transferred beginning in 1968 from the service centers to the newly established IRS Data Center in Detroit, Michigan. Today, the burden of editing the corporation returns is held by about 135 editors at the ten service centers and the Data Center.

We have defined our SOI year to include not only returns of corporations with calendar year accounting periods, but returns reporting accounting periods as early as July of the preceding year and those reporting periods ending as late as June following the calendar year (a span of 23 months) [2]. Since corporations, like other taxpayers, are allowed extensions to their normal filing time, the editors find that editing returns for a single SOI year covers a span of 14 to 15 months. This long period of time serves to complicate the business of editing since the editors are working on returns from several different SOI years during the same time period. The main cause of this complication is due to the different effects of tax law for different years.

Another editing complication arises because there is no legal requirement for the corporation to fulfill its tax return filing requirements by filling in, line for line, the U.S. tax return form. Due to the complexity of tax law and the large differences between companies' industries in organizational and financial matters, the development of a standard tax return form acceptable to all concerned may not be possible. It is our experience that many corporate taxpayers, if not most, will report many of the details of

their financial operations on their own schedules in their own format. Although the return form itself conforms to generally accepted accounting practices, conversion of the taxpayer's own forms and its own terminology to the proper "tax return" concept is often very difficult, even for the most experienced and astute editor.

Terminology plays a critical role in the complexity of the editing process. There is no single accepted method of accounting used throughout the country but rather there are several acceptable "guidelines", many of which are unique to geographic locations and industries. Terms peculiar to petroleum refining operations such as "delay rentals," for example, can be found more frequently, as expected, in the returns filed in the Southwest than those from other parts of the country.

To assure that the editing process is done with a maximum of accuracy and consistency from editor to editor and region to region, the Statistics of Income Division prepares editing instructions for each SOI year. These instructions, which for 1980 consisted of 250 pages, provided details not only for editing normal and rather straightforward terms such as "total assets" or "total deductions" but also included instructions for the exceptions and non-standard situations that might be encountered. Whenever an unfamiliar or uncommon term was encountered on several returns for a year, it was included in the instructions. For example, if the item "commercial drafts or paper", was reported in the category "other assets" on the taxpayers return, the instructions would require that it be edited as part of "Trade Notes and Accounts Receivable" since our investigation has revealed that it is more closely related to this item than to "Other Assets." Complete instructions covering every possible term or variation of terms or other unusual conditions, of course, is not possible, so a great deal of latitude has been allowed for personal judgement of the editor in the interpretation of instructions and terminology. This has led to different interpretations across the country which were not documented.

Another complication arises since the same data items might be edited differently depending upon the industry of the reporting company. For example, the amount included under "certificates of participation" has been edited differently depending upon the industry of the reporting company. For example, the amount included under "certificates of participation" has been edited as "Other Current Liabilities" for all banks (SOI industry codes 6030 through 6090) and certain other credit agencies (SOI industry codes 6120 and 6199). For all the other industries, when this term occurred it has been edited as "Other Liabilities."

Once the returns have been edited and the data transcribed into the computer system the data are tested for errors and inconsistencies. Errors and inconsistencies can arise from mistakes either in editing, transcription or may in fact be uncorrected taxpayer reporting errors. The correction process, however, has never been entirely satisfactory since recourse to the return was limited. After SOI editing occurred, the returns were sent back to the normal revenue processing center. They were not generally available for statistical purposes except for a small sample of returns and edit sheets which were selected as part of a quality review program [3].

In order to deal with some of these basic problems inherent in the system, new techniques were implemented for tax year 1980. Immediately after a return had been edited, it was transcribed, entered into the computer, and subjected to math or validity checks. Errors were corrected on site while the return was still available for statistical use. For 1980, 30 tests were applied to each record. Some of these basic tests included out of balance checks for asset items, liability items, dividend items, receipt items and deduction items [4].

PLANNED CHANGES

While these changes helped to improve the program, it had become evident that a substantial change in the overall processing approach would be needed to keep pace with the increase in demand for larger samples, more timely publications, and reduced financial resources. Beginning with the 1981 tax year, we are implementing a two-phase program to develop a more effective and efficient editing operation. This program consists of (1) simplified initial manual editing with (2) automatic or computer-assisted supplementary editing.

Under the new system, the editing process has been broken down into basic steps. As in prior years, large returns (these are generally defined to include returns reporting assets of $250 million or more) and their accompanying tax forms and schedules are edited on site in the service centers on a single six page edit sheet that includes over 400 codes and items. In order to make the editing an easier task, the codes and items on this edit sheet have been arranged to reflect the sequence of the return form and that of the various other forms and schedules. Previously, the edit sheet had been sequenced more to suit the needs of the statistical analysts in the National Office who designed the edit sheet rather than the editors in the field.

The editing of the returns for the small corporations has been drastically simplified. These returns, including easily edited attachments, are edited at the Data Center on a four page edit sheet that has also been arranged to reflect the basic return form sequence. Data from the more difficult to edit attachments such as Forms 4562 (Depreciation), 3468 (Investment Credit), and 3468-B (Business Energy Investment Credit), as well as all data from taxpayers' own schedules and spread sheets, and certain data unique to Form 1120L and 1120-DISC returns are excluded from the four page edit sheet and edited at the second phase. The editors at the Data Center merely enter a code for the existence of these forms or for any "missing" data from the basic tax return form which may be presented in the taxpayers' own schedules. For instance, if the editors find that the taxpayer has inserted the phrase "See Statement 1" on the

basic tax return form instead of a money amount, then the editor will simply enter an appropriate code indicating the general location of the missing data (whether in the income statement, balance sheet, tax credits, etc.). These codes enable the editor to edit the return package quickly. In prior years, there was much time spent leafing through the entire return package for the indicated data and shifting back and forth, to and from the basic tax return form.

Also, under this new approach, the editors in the first phase no longer examine the taxpayers' schedules for summary or catch-all items such as "other income," "other deductions," "other assets," etc. and allocate any identifiable amounts to specific income, deduction, asset or liability fields on the edit sheet. This process is delayed until the second phase of editing.

In addition, the editing of delinquent or prior year returns has been eliminated. Prior year returns that are filed during the current tax year often present special problems for the editors since many of the data items are either not present on the older tax form or are present but are displayed differently. In prior years, the rationale for including delinquent returns was that they would provide estimates of the types of current year returns that were not filed in time to be included in the sample. However, not only are these late returns more expensive to process, but because of inflation and tax law changes, they may no longer be adequate estimates of the current year's late returns [5].

As a result of these changes, and the desire to streamline every aspect of the initial editing process, we have made extensive changes to the editing instructions. For the large returns, the editing instructions are still about 250 pages but now include dictionaries for the income statement and balance sheet items. These dictionaries which present the income, deduction and balance sheet terms in alphabetical order are very useful when it comes to allocating amounts from taxpayers' own running schedules or spread sheets.

The instructions for the small returns have been reduced to about 90 pages. The instruction for each data element is limited to the edit sheet field number, name of the data field, and the physical location of the item on the tax form or schedule (including the form or schedule number, page, and line number).

These editing changes were field tested in December of 1981, prior to the start of the editing of the 1981 tax returns, using the old editing method as a controlled comparison [6]. Two groups of 8 to 10 randomly selected editors each edited a representative sample of 80 returns. Half the editors in each group edited the 80 returns using the old, current instructions and half edited the same returns using the new simplified instructions. The editing time was recorded for each return. The results of the test data show a 40% decrease in the average editing time using the new procedures. Present editing rates for the 1981 SOI year, are over two returns per hour compared with less than one return per hour for the 1980 SOI year.

Once the edit sheet data have been entered into the computer at the Data Center, the large returns (or records as they are now called) are subjected to 70 tests for consistency while the smaller records undergo over 350 different tests. About half of the 350 tests include automatic corrections. Records that fail the tests with automatic correction provisions will be corrected by the computer and will be considered correct records by the computer program.

It is this consistency testing and the process of automatic and computer-assisted editing of the smaller records that is the key to the efficiency of this new system. The expansion from 30 tests for the smaller returns in 1980 to over 350 will actually enable us to reduce the manual editing effort for these returns. Perhaps the best example of this occurs with industry coding. Previously the editor used the taxpayer supplied "Principal Business Activity" (PBA) code, together with the business activity description and the editor's own determination of the major source of the company's receipts to determine the SOI industry code. Under the new system, the prior year SOI code is automatically assigned by the computer for both the large and small returns if the 1981 edited PBA code matches the PBA code of the previous year. If the prior year return is not in the file or if the PBA codes differ, the record is flagged and printed out so that an editor can manually edit the code. However, for certain small returns (those with total assets under $500,000), the PBA code is automatically transferred to the SOI industry code even if the prior year return is missing from the current SOI file. The PBA code, however, must be a valid SOI code for the automatic transfer to take place. As part of the testing for this new system, over 9,000 returns were subjected to the test. Table 1 shows that less than 30% of the returns read out for manual industry coding. If this ratio holds true, then we can expect about 69,000 returns to be automatically coded for 1981. Because of this reduction of manual coding, we anticipate not only an improvement in the quality of our industry data but also substantially lower processing costs.

Other automatic editing operations include the transfer of negative amounts reported by the taxpayer in otherwise positive fields, into the appropriate negative field. An example of this situation is the transfer of a negative income amount such as negative "other interest" into the appropriate deduction field, "interest paid." Because the entire operation involves four steps, (1) deleting the negative amount, (2) subtracting it from the old total field, (3) subtracting it from the appropriate deduction field, and (4) subtracting it from the appropriate total deduction field, the automatic changes not only are less expensive to perform than the old manual method but also are more efficient since all chance of human error in addition or subtraction has been eliminated. Table 2 shows that out of the 9,263 returns, 876 invalid negative entries on the income statement and balance sheet were automatically transferred to the correct field.

In addition to the savings anticipated from automatic industry coding and automatic transfer of negative amounts, savings are also expected from the automatic merger during consistency testing of two edit sheets for selected types of returns. Prior to 1981, two edit sheets were prepared for Mutual Savings Banks with life insurance departments. One edit sheet was prepared for the Savings Bank parent which filed on Form 1120 and the other for the life insurance department which filed on Form 1120L. In order to present valid data for mutual savings banks in our statistics, it was necessary to manually merge the 1120L return, data item for data item, with the parent. Although the number of these types of returns was relatively small, error was manually introduced as a result of the manual mergers. Starting with 1981 however, due to a change in tax law, there will be additional returns that require the combination of edit sheets. Insurance companies can now file as part of consolidated returns, i.e., Form 1120 parent with a Form 1120L subsidiary.

The non-automatic consistency tests were greatly expanded to assist the manual editing function. Records that have (1) failed the industry code comparison test, or (2) failed the "non-automatic" balance or validity checks, or (3) in the case of the smaller returns, coded for additional editing will be printed out in hard copy for manual processing.

Some of the computer-assisted tests include the manual editing of "missing" data (those line items on the return form where the taxpayer entered "See attached statement"). Although this editing is delayed until consistency test processing, the delay enables us to gather some information on taxpayer reporting characteristics.

Other editing during this second phase includes the Forms 4562, 3468, and 3468-B which were coded during phase one for later editing. Our original intent for the delayed editing of these forms was to edit these schedules on a sample basis since they occur frequently and are very time consuming and difficult to work. However, the weighting problem associated with subsampling a sample eventually precluded this approach (at least for the time being). We still included the delayed editing of these schedules in the system, since we think that editing these schedules continuously, one after the other, will result in the positive benefits of efficiency and accuracy of assembly line production.

Another improvement resulting from changes in the consistency testing program involves both the manual and automatic editing of taxpayers' summary or catch-all schedules, i.e., other income, other deductions, other assets, etc. During the initial manual editing phase, only the "other" amounts shown on the tax return were edited. The editors did not examine the taxpayers' own schedules and allocate the amounts to specific fields. During consistency testing, if the ratio of the "other" amounts to the "total" amounts (total income, total deductions, total assets, etc.) exceed certain predetermined proportions, then the return will be printed out for manual editing (Table 3). The editors will then examine these "other" schedules and allocate specific amounts to a maximum of four fields. The original amounts in the fields are stored as are the four allocated "other" schedule fields, providing us with documentation of the changes made to taxpayer entries. The computer then automatically redistributes the amounts, making the necessary computations. In addition, a subsample (8% to 10%) of those schedules where the ratio of "other" to "total" is less than the predetermined proportions will also be printed out for manual editing during consistency testing (Table 3). The rate of "other" schedules imputed ranges from a high of 72% for "other income" to a low of 49% for "other deductions/cost of goods" (Table 3)[7].

CONCLUSION AND AREAS FOR FUTURE STUDY

Although many changes have been designed for the 1981 SOI and some are only now being implemented, modifications and improvements are already underway. In some cases, our original plans have proved to be too ambitious and had to be postponed to later years. The important thing, we think, is that we recognize that our editing system must keep pace with program requirements and resource availability. These innovations for 1981 will undoubtedly be improved upon for 1982 SOI.

Plans are currently underway to implement a data base system for accessing return data directly through the use of on-line computer terminals. One aspect of this system is a control operation that will enable us to correct editing and transcription errors in selected identification entries. This early data correction process provides us with a means of controlling the sample by monitoring the returns and accompanying documents as they flow through the different phases of the processing system.

ACKNOWLEDGMENTS

The authors wish to thank Nathan Shaifer, who edited the manuscript, and Douglas Brooks, who typed the several drafts of this paper. We would also like to thank Ruby Alford and Lillie Norman, of the Detroit Data Center, for their help in testing the simplified editing instructions.

NOTES AND REFERENCES

[1] Powell, W., and Stubbs, J., "Using Business Master File Data For Statistics of Income," 1981 American Statistical Association Proceedings, Section on Survey Research Methods.

[2] Crum, W. L., "Fiscal-Year Reporting for Corporate Income Tax," National Bureau of Economic Research, Inc. 1956.

[3] Schwartz, O., "More Quality for the Money in Statistics of Income," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[4] Bahnke, J., and Wheeler, T., "Corporate Statistics of Income: Data Testing,"

1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[5] Some of the estimation techniques used to compensate for these incomplete data are described in the following papers: Dumais, J., and Shadid, R., "Individual Statistics of Income: Advancing the Closeout Date," American Statistical Association 1981 Proceedings, Section on Survey Research Methods and Harte, J., "Post-Stratification Approaches in the Statistics of Income Program," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[6] Cys, K., and Hinkins, S., "Editing Experiment, Corporation Income Tax Returns, Forms 1120S," Statistics Division, Internal Revenue Service.

[7] Procedures for imputing data for incomplete or missing balance sheets are discussed in Hinkins, S., "Imputation of Missing Items on Corporate Balance Sheets," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

Table 1.—1981 CORPORATION VALIDATION ERROR ANALYSIS

Test	Description	Number of Times Failed
1	Invalid SOI Industry Code	2461
2-28	Invalid Code	864
29	Problem Code Present	1391
30-31	1120M to be Edited	14
32-37	Invalid Codes for 1120S	6
38-47	Invalid Codes and Amounts for 1120L or 1120M	12
48	Invalid Amount on Rejects	11
47-56	Invalid Amounts or Elements	113
57	Print Out Other Income Schedule	1667
58	Print Out Other Deduction and Cost of Goods Sold Schedules	3983
59	Print Out Other Current Assets and Other Assets Schedules	2501
60	Print Out Other Current Liabilities and Other Liabilities Schedules	2408
61-68	Data from Supplemental Schedule Missing	408
69-70	Data from Form 3468 Missing	6551
71-72	Data from Form 3468-B Missing	112
73-74	Data from Form 4562 Missing	8109
75-76	Print Out 1120-DISC Validation Edit Register	1
77-92	Balance Sheet Inconsistencies	2672
93-104	Income Statement Inconsistencies	5848
105-108	Relationship of Balance Sheet for Finance Industry	542
109-126	Schedule D Items—Inconsistencies and Relationship	201
127-132	Relationship of Tax to Other Amounts	910
133-135	Relationship on Form 4626	122
136-137	Relationship on Form Form 6249	12
138-140	Relationship on Form 6765	5
141-142	Employer Identification Number—Relationship	20
143-189	Miscellaneous Tests	7

Total Records Processed.....9263
Total Records with Errors.....9126

Table 2.—1981 CORPORATION VALIDATION AUTOMATIC ANALYSIS

Test	Description	Number of Times Failed	Test	Description	Number of Times Failed
1-54	Move Invalid Negative Amounts	876	121-122	Delete Negative and Insert Zero in Field	1956
5-64	Move Other Income Amounts from "Other" Schedules	475 1/	123-124	Indicators for Consolidated 1120L	3
65-82	Move Cost of Goods Sold and Other Deduction Amounts from "Other" Schedules	1768 1/	125-134	Correction of Codes	156
83-101	Move Other Assets and Other Current Assets Amounts from "Other" Schedules	992 1/	135-136	Correction of Amounts	0
102-114	Move Other Liabilities and Other Current Liabilities Amounts from "Other" Schedules	102 1/	137-153	Miscellaneous Checks on Corrections to "Other" Schedules	4168 1/
115-120	Change Invalid Negatives to Absolute Values	2087	Total Records Processed.....9263 Total Records with Automatics.....3736		

1/ These automatic tests are applied to subsequent cycles only.

Table 3.—PERCENT OF OTHER SCHEDULES BEING MANUALLY EDITED/IMPUTED DURING VALIDATION

Form 1120 Schedule	Percent Manually Edited		Percent Imputed
	$\frac{\text{"Other"}}{\text{Total}} >$ Predetermined Percentage	$\frac{\text{"Other"}}{\text{Total}} <$ Predetermined Percentage	
Other Income	18	10	72
Other Deductions and Cost of Goods Sold	43	8	49
Other Current Assets and Other Assets	27	9	64
Other Current Liabilities and Other Liabilities	26	8	66