James E. Bahnke and Timothy D. Wheeler, Internal Revenue Service

This paper provides an overview of changes in methods used in the Corporation Statistics of Income Program to control corporation income tax returns through the statistical processing system and to test and correct data taken from the returns for the purpose of producing annual statistics as required by law [1].

Organizationally, the paper is divided into four parts. Part one provides an overview of the corporation statistics program. Part two discusses new methods in controlling returns through the IRS Data Center, a processing center located in Detroit, Michigan. Part three discusses the process of data testing in prior years and part four discusses new methods in data testing. Data testing is the computer process for checking internal consistency of the financial data.

BACKGROUND

In 1916, a law was enacted requiring the preparation and publication of annual statistics with respect to the operation of the federal income tax laws. For this and other reasons, an annual sample of approximately 95,000 corporation returns are selected from a population of approximately 2.9 million corporation returns filed each year.

Returns are filed at ten service centers located throughout the United States. The Statistics of Income Program selects corporation returns from the revenue processing system of the Service to perform statistical editing, the manual process of abstracting and transcribing selected corporation data onto documents known as edit sheets [3]. Statistical editing is performed by about 135 editors at the ten service centers and the Data Center. The edit sheets prepared at the service centers are forwarded to the Data Center for further processing. Returns shipped to the Data Center for editing are returned to service centers, for subsequent revenue processing, upon completion of data abstraction. Once the returns have been edited they are transcribed onto tape and data tested. This testing, performed at the Data Center, is composed of computerized checks and balances to examine the validity of the data. Upon completion of data testing and correction, the resulting file (on computer tape) is called the Accepted File. The Accepted File is the basis of the annual statistics of income.

CONTROLLING RETURNS THROUGH THE DATA CENTER PIPELINE

The function within the 1980 program which has been greatly expanded over previous years is that of controlling. For 1980, the Data Center developed a computer system that closely monitors the receipt of returns at the Data Center. This system provides the Statistics of Income Division with the ability to identify returns missing from the sample and take action.

It cannot be assumed that all designated returns will automatically be sent to the Data Center for subsequent manual processing. Even further, it cannot be assumed that a designated return will go through all the manual and computer processes and ultimately find its way to the Accepted File. Designated returns may not be available at the service centers, and thus, never be sent to the Data Center. Also, returns could be misplaced at the Data Center or inadvertently returned to the originating service center before Data Center processing. Returns could be edited, but never transcribed, and again never find their way to the Accepted File. Each return is monitored as it passes through various checkpoints along the manual processing pipeline. When a return is misplaced, it can be recognized almost immediately so that the situation can be remedied with little or no impact on any further processing.

Return data from Internal Revenue Service revenue processing operations are used to create service center transaction tapes, which are the basis for the sample selection. Using a selection program based on the Employer Identification Number of each return, the size of assets and income, and business activity, the returns are selected from the tape file and the identity information is printed on lists which can be used to locate the return. Service center transaction tapes are generated each week (cycle) at all service centers. Weekly tapes containing revenue processing data from each return selected for the sample are merged at the end of each month and shipped to the Data Center. The Data Center begins controlling returns at the point of receipt of the transaction tape. All tapes are loaded into an online data base. Missing cycles can be easily recognized, and service centers are notified accordingly, so that missing cycles can be processed and released to the Data Center as soon as possible.

Returns, or edit sheets for returns edited at service centers, are expected to arrive at the Data Center within 120 days from the date each return is designated for the sample. For returns not received within that period, a list is furnished to each service center so the returns can be located and sent to the Data Center for processing.

Once a return arrives at the Data Center, its Employer Identification Number (EIN) and Document Locator Number (DLN) are key entered onto the control system using an interactive terminal where these points of identification are matched to the online data base file. For matched returns, the date of receipt is posted to the file and marked as received. For any unmatched conditions, the returns are added to the data base file, but flagged as not having a matching transaction file. As returns move

through the manual processing system in the Data Center, each return's progress is entered into the data base. Thus, each return must be marked as complete before it can move into the next phase of processing. In addition to controlling, information can be tabulated from the data base concerning volume of returns received and elapsed turnaround time for returns at each control point as well as the entire pipeline. Finally, information such as asset size, business receipts, and sample code can be obtained with a file inquiry instead of writing and processing a separate application program.

TESTING IN PRIOR YEARS

After editing and transcription every return record must pass the consistency tests before it can become part of the Accepted File. While we do not audit taxpayer data, before any tables can be produced, the information from Form 1120, U.S. Corporation Income Tax Returns and related schedules must be tested for internal consistency. Consistency tests are written to identify errors in reporting by the taxpayer because of the taxpayer's lack of understanding of the tax law, tax form and instructions, or failure to report or misreporting data. In addition, tests identify editing and transcription errors or omissions. To write consistency tests, we need the edit sheet and editing instructions (for program content and field locations), sampling specifications (for sample code tests), table specifications (for table items), and various chart analyses of the previous consistency test programs, including frequency of individual test failure and corrections applied.

Consistency tests are designed to identify returns that:

(1) have impossible conditions.
(2) have items out of balance with totals shown.
(3) have improper relationships between data items.
(4) Have certain characteristics that require review, such as a bank with an extremely large amount of cost of goods sold.

Each consistency test has a number associated with it for identification purposes. There are two main types of consistency tests: error condition tests and information condition tests. Error tests are subsequently broken down into four categories: automatics, which are computer statements which automatically correct error conditions; error conditions which are corrected manually; size error tests which are used primarily to derive various value relationships within the data; and, lastly, sample code tests which are used to point out inconsistencies or problems in the sampling program.

Information tests, on the other hand, are primarily designed to identify corporations whose assets and receipts are not characteristic of their industry. They are also used to compare current and prior-year data and to identify conditions which require further analysis by the professional staff.

IMPROVED TESTING

The Consistency test program was thoroughly reviewed after the 1979 program. Four tax years were examined during this review: 1976 through 1979. During that assessment, test requirements were analyzed to determine the changes which were needed to reduce error rates and the cost of the overall program. Over the four years there was a sharp increase in the number of errors identified by the testing program. In fact, there was a 25% increase in Tax Year 1978 over 1977 alone.

After reviewing the associated consistency test requirements for those years, one problem surfaced; the rigidness that prior programs were built on. Tolerance levels for tests were very low and were never adjusted during the course of the program. Tolerance factors are the amount, plus or minus, that a particular calculation may be out of balance before the computer identifies an error or information condition. Further, the data that the prior-year programs yielded were never assessed for changes or for the purpose of deleting tests altogether. Also, tests needed to be revised because of inflation; the dollar values in the tests no longer applied. Some tests failed to identify true error conditions and needed revisions. Finally, corrections were made without the use of the tax return, which was sent, upon completion of editing, back to revenue processing at the Service Centers. The only returns available during testing were certain large returns, which had been microfilmed.

With previous programs in mind, it was decided that five basic principles would be applied in developing a new set of requirements for the data testing phase of the Tax Year 1980 program. First, previous years' tests would be reviewed to determine if they had an application in the 1980 program and to determine if and what method of correction could be applied to the test. In the past, data were tested without much thought to the correction process at the time the tests were written. Next, it was recognized that proper controls needed to be applied to the overall consistency test requirements. Third, flexibility had to be built into each and every test, yet at the same time a standardized format would also be sought to enhance the level of communication between the Data Center and Statistics of Income Division. Lastly, it was decided that an ongoing review of production data would take place from the beginning of the program until its closeout.

As a result of assessing prior year consistency test requirements, the 1980 specifications resulted in a reduction from approximately 720 tests in 1979 to 508 in 1980. The review of the tests also resulted in an expansion of the use of the automatic test. In previous years when an automatic test was applied, the computer would make the correction with no documentation of the changes made. The

1980 program provided for indicators to be posted on each record so that it would be readily apparent where automatics were applied. Further, an automatic test register, which displays a sample of the records after the test was applied was printed. Thus, National Office personnel were able to get a firsthand look at these records for analytical purposes. For 1981, the system will be improved by printing the record to the automatic register before and after the automatic was applied. This will allow the reviewer to see the exact manipulations performed.

In previous years, returns which were deemed by the National Office or Data Center personnel as not being part of the sample could be readily deleted prior to the closing of the file. It was decided that for 1980, all returns would stay in the file and that erroneous or duplicate returns would be marked for possible exclusion from further steps such as weighting and tabulation. This ensures that all records that were input into the system remained in the system. In previous years records could be deleted, sometimes incorrectly, with no documentation of the deletion. Also, in previous years, records were easily accepted into the Accepted File that still contained error conditions. While this shortcoming was not eliminated for 1980, the steps required to accept a record into the file were tightly controlled.

There are two ways that a record can get into the Accepted File. The first is if there are no error or information conditions on the record. The second is when error or information conditions are present on the record, but it is determined that no correction is necessary. This is called "accept coding" a record. In our new approach a record that is manually accept coded will be displayed on an Accept Code Register.

The Accept Code Register is sent to the National Office for professional personnel to review. Further, personnel accepting the record were required to provide his or her identification numbers which were entered onto the record. In addition, various tests were identified as MUST tests. These were conditions that had to be manually corrected and could not be present if a return was to be accept coded. As a result of these measures, it was relatively certain that major error conditions would be resolved before going to the Accepted File.

Another control which was instituted was that of the sample code register. The sample code register contains records that have an incorrect sampling condition. In previous years, nearly all the sample code tests were automatically corrected by the computer. In many cases, this resulted in the record being sample coded using data currently in the record, not the data at the point of selection. For 1980, records with sample code inconsistencies were entered on a Sample Code Register and sent to the National Office for professional personnel to review.

Another area which was greatly enhanced was that of standardization. The approach which was taken in the writing of consistency test specifications has made it much easier to understand the intent of each consistency test and has ultimately reduced the number of misinterpretations by computer programmers. Also, management review has become faster and more meaningful.

One of the most important control features is test analyses, which display the number of times each consistency test reads out. Additional test analyses were incorporated into the 1980 program. These analyses indicate if a particular test is reading out at a high frequency, or never reading out. Also, reviewing the analysis for a recycle indicates how often a particular test was corrected, or if a test read out more often for a recycle, indicating a serious problem. A recycle is a group of records that is re-consistency tested, because error or information conditions were still present.

Review of the analyses showed that the number of errors decrease approximately 75% from one cycle to another. The approximate 75% decrease in errors between cycles helps to pinpoint problem test areas for the 1980 program; that is, if a test decreases between cycles less than 75%, perhaps some part of the program should be modified. A detailed review of each test would provide extremely important data on "test performance." Certain tests had a higher correction frequency on original cycles than recycles.

Another improvement for the 1980 program was that the first ten test numbers of error conditions identified during the original test cycle of each record are stored on the record. This provides valuable review information. A person correcting a recycle can tell at a glance what conditions were present on the original record. Also, we plan to perform a detailed analysis to see if certain tests always read out together. This could result in the elimination of tests.

Although there have been a greater number of controls added to the 1980 program, increased flexibility has also been achieved. An example is that of tolerance factors. In previous years, tolerances were manually coded on a test-by-test basis into the computer program. For 1980, tolerances are handled through the use of a parameter and can be easily adjusted up or down depending on the data, and changes can be made without the need to alter the program. The program has also been structured to allow for groups of tests to be bypassed; thus, if conditions which originally appeared to be necessary are determined to be of little value after the initial review of production outputs, a series of tests can be bypassed with no impact on the computer program. There has also been an area in the program set aside so that new tests can be added without affecting the main body of the original program. These new tests can be developed apart from the main program, so that regular production can continue without adversely affecting the timely delivery of production outputs to the user.

Although we have used computed amounts in the past, the number of computed amounts derived in the consistency tests has increased. A computed amount represents the sum of several

component amounts and is automatically included on the data record. There were 24 computed amounts in 1980 compared to eight for 1977. A reason for the increased use of computed amounts was the change in 1978 to whole dollar editing rather than thousands of dollars. Since the tax return had also increased in complexity, it was decided to have the computer perform more computations. This eliminated the need for personnel to do the computations themselves, reducing the possibility of introducing additional errors.

In 1980, a new method of testing computed amounts was used. For example, Computed Total Assets, is printed displaying the total of the components of total assets. Then a consistency test reads out if the edited amount for Total Assets does not equal the Computed Total Assets figure. The new computed figure, Computed Asset Imbalance, is then derived by subtracting the edited Total Asset amount from the Computed Total Asset figure. This enables the error resolution clerk to see the exact amount of imbalance without having to do the calculation. This type of computed amount was also used for the liability side of the balance sheet and both the income side and deduction side of the income statement.

The last area which was assessed was that of the amount of review necessary after program implementation. Prior to production, 7,694 live data records were processed through the consistency test program for analysis purposes. These records were processed three times, with changes applied to the actual tests after each cycle. By reviewing the frequency of test readouts at this early stage of processing, potential problems were identified and resolved. After production began, every cycle was assessed thoroughly for the possibility of adjusting tolerance factors, deleting or adding test conditions, and modifying correction instructions.

The editing function was modified slightly for 1980. Previously, after a return was edited, it was immediately shipped back to its originating service center and was not available during consistency test processing. In 1980, it was decided that, for a small number of tests, which checked important areas such as the balance sheet, the return would be used to resolve any error conditions that were uncovered.

An analysis of the 1979 and 1980 original cycle ouputs for 34 tests, which were used in this new editing approach, called "Validation Testing," indicated an average reduction of approximately 76% in error conditions during the subsequent consistency testing. A further evaluation of the validation testing and of the entire error correction procedure is the percent reduction in error readouts between cycles of error resolution batches relating to validation tested and nonvalidation tested records during the subsequent consistency testing. The reduction in error rate between the original and first recycle was 78% for validation tested records and 90% for nonvalidation tested records. The reduction in error rate between the first recycle and second recycle was 89% for validation tested records and 88% for nonvalidation tested records. There was no significiant difference between the error rate for validation and nonvalidation tests. No conclusion has been drawn from these comparisons, except that the 34 tests involved were corrected approximately 85% on each cycle, a rate which we find acceptable.

Even though the 1980 validation effort was small in scale, the results were quite positive and it accomplished its objective of reducing the error resolution cost and improving the quality of the statistics [2].

The overall results of these changes in the 1980 consistency test program are a 10% reduction in the error rate. Even more important is the improvement made in the quality of the data. A program of this size cannot be placed into a production environment without being constantly reviewed. Keeping this in mind, further enhancements beyond 1980 are presently being implemented [3]. As with the 1980 program, cost and timeliness will be major factors in determining future improvements [4]. It is hoped that on an annual basis changes can be made so that the program runs ever more smoothly, effectively, and produces statistics which are meaninoful to users.

ACKNOWLEDGMENTS

The authors wish to thank Dan Rosa and Karen Cys who reviewed the manuscript, and Douglas Brooks, who typed several drafts of this paper.

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

[3] Cys, K., Hinkins, S., and Rehula, V., "Automatic and Manual Edits for Corporation Income Tax Returns," 1982 American Statistical Association Proceedings Section on Survey Research Methods.

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