USING BUSINESS MASTER FILE DATA FOR STATISTICS OF INCOME PURPOSES

William T. Powell and Joel R. Stubbs, Internal Revenue Service

This paper describes the results of the most recent of several small-scale pilot studies that compared Business Master File (BMF) data with those from the Statistics of Income (SOI) file for corporation income tax returns.

Organizationally, the paper is divided into 4 sections. Section 1 provides the historical background of both the corporation Statistics of Income (SOI) program and the Business Master File (BMF) revenue processing system. In section 2, the industrial classification is described for returns processed for BMF and SOI purposes. Comparability of BMF-SOI financial information is discussed in section 3. The last section includes the conclusions and areas for future study.

1. HISTORICAL BACKGROUND

The Corporation Statistics of Income (SOI) program, as well as the other major SOI programs [1], began with the passage of the Revenue Act of 1916 which called for the publication of annual "facts deemed pertinent and valuable" with respect to the operation of the income tax laws. Statistics of Income for 1916, which contained data on individuals and corporations, was approved by the Secretary of the Treasury on June 1, 1918, and thus became the first report to fulfill those requirements of the Revenue Act of 1916. From the beginning, and up to the present, the primary users of corporation SOI data have been the tax policymakers and the revenue estimators of the Treasury Department. Congress has also made extensive use of the data. Since that first publication, there has been a large increase in the business activity of the country. Figure 1 shows the growth in corporation returns as estimated by SOI, from 1916 to the most recent year for which data are available.

In the early years (1916-1922), items and classifications were very limited, restricted chiefly to the State where the return was filed, industrial activity of the corporation, and totals for a few amounts such as gross income, deductions, and tax. As time went on, size classifications, primarily by assets and by business receipts were introduced. Additional items were also added, particularly balance sheet data and detail on how various return totals were computed.

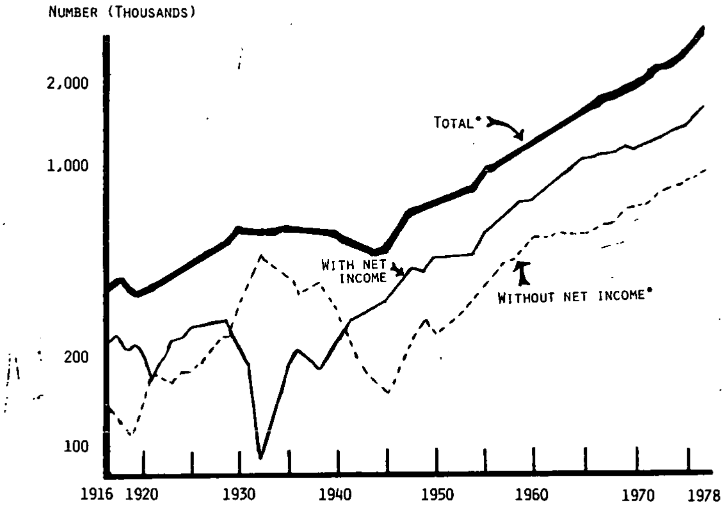
Up until the 1930's, corporation SOI publications gave some detail by income class and State, but only summary information for types of income, deductions, and so forth. Because of the limited detailed information, SOI data could not be compared to salary and wages data collected as part of the various economic censuses [8], nor could it be compared to data contained in other source materials on income and taxes. Beginning in 1933, the Department of Commerce requested special tabulations to be used for the first estimates of National Income [3]. Since that time, corporation SOI information has been a primary source of data used in these estimates to produce the Gross National Product.

As the tax law has become more complex, requirements for more detailed data have grown correspondingly. Currently, the basic corporation SOI program presents estimates on approximately 250 data items classified by asset size and by industrial activity. This corresponds to the approximately five items classified by State and by industrial activity that appeared in the early publications.

BMF System

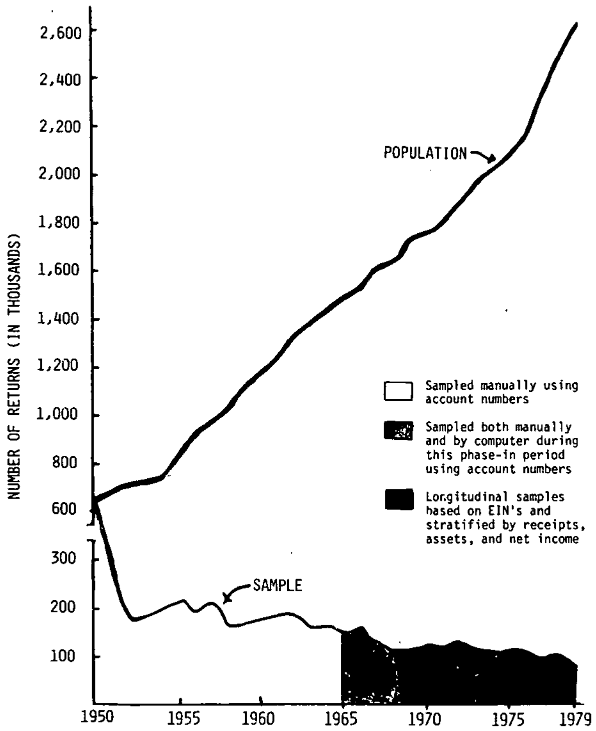
In early 1959, the Treasury Department gave approval for the computerization of the revenue

FIGURE 1.--NUMBER OF CORPORATION RETURNS AS ESTIMATED FROM STATISTICS OF INCOME, 1916-1978



*Includes inactive corporations.

Figure 2.--Corporation Population and SOI Sample, Tax Years 1950-1979



processing system. Prior to this, the IRS had no centralized accounting system. Document processing was limited to the return itself, and even this processing terminated with the preparation of balance due or refund data. There were three main features to the automated systems:

1. A Master File of all taxpaying entities - individuals and businesses,
2. a permanent identifying number for each taxpayer, utilizing numbering schemes developed earlier by the Social Security Administration (Employer Identification Numbers (EIN's) for businesses and Social Security Numbers (SSN's) for individuals), and
3. centralized processing through use of the service centers and the National Computer Center.

The Business Master File (BMF) system, containing data on all business taxpayers, went nationwide on January 1, 1965, and the Individual Master File (IMF) has been operational since January 1, 1967.

With centralized processing and, more importantly, with a permanent taxpayer identifying number, the SOI samples could be designated by using the Master File systems. Automated sample designation procedures were phased in with the implementation of the BMF and IMF systems. By the late 1960's samples for both the corporation and individual SOI programs were being designated by computer using Master File taxpayer identification numbers. As an integral part of the designation programs, a computerized control system has been included so the statisticians who designed and implemented the sampling schemes can see that they are functioning correctly. Through the more sophisticated techniques available with the computer, improved designs have resulted, allowing for smaller samples to be employed. For example, as figure 2 indicates, estimates in the 1951 corporation report were based on a manually selected, stratified, systematic sample of approximately 285 thousand corporation returns from a population of 687 thousand [11]. For Tax Year 1980 estimates of corporation income tax return data will be based on approximately 90 thousand returns from a total of over 2.7 million expected to be filed.

Revenue processing needs differ in several respects from the objectives of SOI. Some items required for SOI are not even part of the revenue processing system. In many cases, BMF data items are defined differently than similar items edited for SOI, or are "perfected" to varying degrees depending on the bearing they have on tax liability. For these reasons, data entry for revenue processing and for the corporation SOI program have so far been separate operations. In the remainder of this paper we will be looking at the possibility of identifying and possibly combining some of these separate operations.

2. BMF-SOI INDUSTRY CODING

One of the most important classifiers of data used in the corporation Statistics of Income program is the industrial activity of the corporation. The initial industrial or business activity code is supplied by the taxpayer. The instructions for the Form 1120, U.S. Corporation Income Tax Return, ask the taxpayer to enter a four digit business activity code on page one of the return. This code is based on the industrial activity accounting for the largest percentage of business receipts of the taxpaying entity.

Due to the importance of the industry code as a classifier in the corporation SOI program, specially trained editors are used to assign SOI industry codes to those returns designated for the SOI sample. This is done because of the potential inaccuracy of the taxpayer supplied industry code entered on the BMF.

A detailed study was carried out in 1975 that compared the BMF industry code with the specially edited SOI industry code for the same returns. This study found that there was approximately 68 percent agreement at the minor industry (four digit) level, 75 percent agreement at the major industry (two digit) level, and 86 percent at the industrial division level [7]. At the minor industry level the agreement ranged from a low of 2 percent in some of the "not allocable" industries to a high of 99 percent for the life insurance industry. Table 1 presents data on the percent agreement for these codes at the industrial division level.

Despite the BMF-SOI differences, we are now researching two uses of the BMF industry codes that should reduce costs and improve overall data quality.

One approach being examined is the use of longitudinal data, where a library of BMF and SOI industry codes would be established. If the taxpayer did not change his BMF industry code from one year to the next, we would accept the prior year SOI code and not independently code for the current year. If the BMF code changed from the prior year to the current year, we would edit for a new SOI industry code for the current year. Because of their importance to the statistics, all "giant" returns (corporations having, in general, assets of $250 million or more) would have their industry code checked every year. As an additional step, a quality assurance procedure would be established where a sample of the smaller returns would be SOI industry coded every year. Aside from providing an estimate of the validity of the codes rolled over from the prior year, this step would ensure that every return would eventually be recoded over a period of time. If, for example, we chose a 20 percent sample for quality assurance purposes, every return in the total SOI sample would be industry coded every five years. This is true because the SOI sample is designed in such a way that returns, once selected for the sample, tend to stay in year after year.

The other approach being examined is the implementation of post-stratification based on BMF industry codes. Post-stratification has been proposed as an economical method of reducing sampling variability. The implementation of post-stratification has been delayed by problems experienced in attempting to obtain BMF population counts by industry code. Because of their potential use in post-stratification, it is important to have BMF industry codes that are comparable to SOI industry codes [17]. It is also important that these codes agree to the fullest extent possible because population counts by industry can be only obtained using the BMF industry codes.

We are currently examining the implications of the disagreement between SOI and BMF industry codes on post-stratification. We have also now obtained population counts by BMF industry code by SOI sample code. These data are currently being tabulated and are to be used on a trial basis in the Tax Year 1979 program.

3. BMF-SOI FINANCIAL INFORMATION

Since 1970, several small-scale pilot studies have been made [e.g., 8, 9] that compared BMF financial data with that from the SOI file for corporation income tax returns. Two of these studies will be reported on here. Both examined actual BMF and SOI money amounts for items that were abstracted from the same line on the Form 1120 return. The first of these included about 200 returns for very large corporations that were filed in the early 1970's (mainly 1971 and 1972). We also have examined a small sample of 50 returns filed this spring for Tax Year 1980.

While the returns used in the earlier (1971-72) comparison had total assets of $250 million or more, the returns in the 1980 comparison had total assets that ranged from $20 thousand to over $183 million. Returns with total assets of $250 million or more were used in the earlier comparison because the impact of editing can be measured better by comparing the larger returns. (Larger returns tend to be more complex and, thus, are more subject to extensive editing by specially trained SOI editing personnel.) The 50 returns in the 1980 comparison were varied in size to get an indication of the impact that editing had on smaller returns.

The percentage difference between the money amounts for the comparable data items in the two studies is shown below.

Percent Difference	1980	1971-72
Under 2%	59.1%	20.5%
2 to 10%	13.6%	45.4%
10 to 20%	11.4%	11.4%
20% or more	15.9%	22.7%

Conceptual Differences

For all of the data items with discrepancies of 20 percent or more, major conceptual differences existed between BMF and SOI approaches at least for certain industries. As an example, in the financial industries for personal credit institutions, all commissions were moved during SOI processing from "other income" to "net receipts." For savings and loan associations, "dividends paid to members" were taken for SOI from "other income" and moved to "interest paid."

The discrepancy for "net depreciation" was caused by moving depreciation for SOI from the "cost of goods sold" schedule--Schedule A. This schedule was also the source for the discrepancy for "employee benefit plans." "Amortization for agreement not to compete" is included in "depreciation" for SOI purposes. When identified in "other current assets", "loans and discounts" are moved into "accounts receivable" for all financial industries.

For "other income", the large BMF-SOI discrepancy was partly due to the failure in revenue processing to search the attached schedules and partly due to data items being moved to "net receipts" for various industries due to conceptual differences in the definition of the data item. The discrepancy for "other deductions", like "other income," was due partly to the failure in the BMF to search the attached schedules for data when primary schedules or lines were blank, and partly due to data items being moved for SOI to the various deduction items depending on the conceptual differences in definition for the var ous industries. An example of the failure in the BMF to search the attached schedule was seen on several returns that had only "net receipts," "cost of goods sold ," "other income," and "other deductions." The items of income and deductions were not distributed at all. A search of the schedules during SOI processing found all of the items of income and deductions.

In the 1971 and 1972 comparison, "cost of goods sold", "gross rents" and "repairs" had large discrepancies that were due to conceptual differences in the definition of the data items. The "cost of goods sold" discrepancy (the amount for SOI was only 78 precent of the BMF total) resulted from it and "net receipts" being deleted from the finance industry for commodity brokers and dealers. The difference between the two amounts was entered into "net ordinary gains." If an amount could be found on the return for commissions, t was moved into "net receipts;" the SOI field for "cost of goods sold" on these returns was always blank.

"Gross rents" for the office and computing machines industry were treated as "net receipts," for SOI purposes, while "salaries and wages" were moved to "cost of goods sold." "Repairs" in the transportation and public utilities industries were treated as "cost of goods sold", in some cases, for SOI purposes. "Amortization" for SOI was twice as large as for BMF also due to conceptual differences in definition that caused data to be moved from "other deductions".

Table 4 presents a percentage distribution of the agreement between items from the two sources. Agreement is shown two ways; the first method does not consider blanks for both items as being in agreement; the second takes blanks into account. The percentage for these two methods varied. For "contributions" agreement was 35 percent under the first method and 99 percent under the second method. The agreement for amortization was 13 percent under the first method and 88 percent under the second method.

4. CONCLUSIONS AND AREAS FOR FUTURE STUDY

Since the data available were so limited, the results from the two small-scale interim studies (they are the first phase of a longer range plan) are inconclusive and preliminary at best. The results, however, may be indicative of those that will come from the forthcoming second phase of the study (that is, a large-scale BMF-SOI match, by computer, of 1979 data from the two sources). This second phase should be completed next spring . We expect that it will provide a more accurate indication of the percent of difference between the comparable data items from the two sources. Additionally, more of the data items that are subject to adjustment due to conceptual differences should be pin-pointed for the additional industries that will be included in this phase of the study.

There are a number of other research studies underway. For example, in the short-run, we plan to try to reduce the number of data items that cannot be used directly from the BMF, for SOI purposes, because of conceptual differences. This will be accomplished by meeting with our major users, the Office of Tax Analysis (OTA) and the Bureau of Economic Analysis (BEA), and determining which data items, if any, can be redefined in a way that will eliminate some of the existing conceptual differences.

For the longer-run, we are seeking other alternatives for using BMF data for SOI purposes. For industry coding, we are considering using data from other government agencies, such as the Bureau of the Census[5]. We are also contemplating adopting the "two-tier data system" approach outlined by Alan Freiden[6]. Freiden recommends that SOI employ the BMF as the "first tier" of its database supplemented by additional data from the return, taken essentially as reported, plus data from outside the tax system. (It is conjectural that the cost and quality of data capture could be much improved by this approach.) To deal with conceptual differences, a "second tier" of data would be built on the first by modeling the relationships between what is provided by the taxpayer and what is needed by policymakers.

Perhaps, we will be able to say more about our research next year.

ACKNOWLEDGEMENTS

The authors gratefully thank Ralph Bristol of the Office of Tax Analysis, U.S. Department of the Treasury, who reviewed the manuscript, and George Hoover of the Internal Revenue Service's Returns Processing and Accounting Division who provided technical advice. We also extend our thanks to Internal Revenue Service employees Terry M. Smith, Detria Liles, Bunice Willis, and Romaine P. Ferguson, who compiled and tabulated the data, Ruth Wise who typed the tables, David Jordan and Ray Samuelson who provided technical and editorial support, and Douglas Brooks and Bob Cripe who typed the text. All are members of the Statistics Division.

NOTES AND REFERENCES

NOTE: Additional materials, not directly referenced in this paper, which bear on this effort are [2,4,10,12 -16].

[1] Blacksin, J , and Plowden, R., "Statistics of Income: A Historical Perspective," American Statistical Association 1981 Proceedings, Section on Survey Research Methods.

[2] Cys, K., "1978 1120 SOI Giant Returns-Tax Return Entry Versus SOI Edit Sheet Entry," memorandum dated August 6, 1981, Statistics Division, Internal Revenue Service, pp. 1-5.

[3] Duncan, J., and Shelton, W., Revolution in United States Government Statistics 1926-1976. U.S. Government Printing Office, Washington, DC 1978.

[4] Elliott, D., "Timely Corporation Statistics", memorandum dated July 1, 1977, Austin Service Center, Internal Revenue Service, pp. 1-4.

[5] Farrell, M., and Sullivan, J , "The Industrial Directory - Status and Direction," American Statistical Association 1981 Proceedings, Section on Survey Research Methods.

[6] Freiden, A., "Data Development for Statistics of Income," unpublished Statistics Division staff paper,

Internal Revenue Service, 1981. (An extract is included in the Appendix.)

[7] Hobbs, J.R., "A Comparison Study of Business Master File Industry Codes to Statistics of Income Industry Codes, 1968 and 1972," memorandum dated August 21, 1975, Statistics Division, Internal Revenue Service, pp. 4-7.

[8] Data are still not available through SOI on wages and salaries but we are currently involved in a study to link these data to income tax return data. Related work is being carried on by the Small Business Administration. See Kirchoff, B., and Hirschberg, D., "Small Business Data Base: Progress and Potential," Americal Statistical Association 1981 Proceedings, Section on Survey Research Methods

[9] Powell, W.T., "Analysis of Data From the Business Master File (BMF)," memorandum dated May 13, 1971, Statistics Division, Internal Revenue Service, pp. 1-11.

[10] Powell, W.T., "Comparison of Statistics of Income Data with Business Master File Data," memorandum dated September 13, 1972, Statistics Division, Internal Revenue Service, pp. 1-12.

[11] Powell, W.T., "Special BMF-SOI Report," unpublished Statistics Division Staff paper, 1980.

[12] Schwartz, O., "SOI Adjustments to 'Other Income' in 1980 SOI Tax Returns," memorandum dated August 23, 1981, Statistics Division, Internal Revenue Service, pp. 1-5.

[13] Internal Revenue Service, Department of the Treasury, Complete Report, Statistics of Income--1976, Corporation Income Tax Returns. U.S. Government Printing Office, Washington DC 1981.

[14] Internal Revenue Service, Department of the Treasury, Historical Summary, Statistics of Income--1916-1965. U.S. Government Printing Office 1968. Corporation income tax returns were not sampled until 1951. Individual income tax returns were sampled for SOI purposes from the beginning of the program. In the early years the reports were based on all individual income tax returns with $5,000 or more in "net income" and on 8 percent of small returns.

[15] Internal Revenue Service, Department of the Treasury, Source Book, Statistics of Income--1976 Corporation Income Tax Returns. U.S. Government Printing Office, Washington, DC 1980.

[16] Office of Federal Statistical Policy and Standards, "Statistical Policy working Paper 6, Report on Statistical Uses of Administrative Records," U.S. Government Printing Office, Washington, DC 1980.

[17] Westat, Inc, "Results of a Study to Improve Sampling Efficiency of Statistics of Corporation Income," contract report Westat Inc., dated January 15, 1974, Westat Inc., Rockville, MD. Details of this study can be made available upon request.

BASIC TABLES

Table 1.--BMF-SOI Agreement by SOI Industrial Division

SOI Industrial Division	Percent BMF Agreement
Agriculture, forestry, fishing..........	79.0
Mining................................	88.2
Contract construction..................	89.2
Manufacturing.........................	88.2
Transportation, communication, electric, gas, and sanitary services...	75.7
Wholesale and retail trade	87.7
Finance Insurance and Real Estate......	84.7
Services..............................	91.7

Table 2.--BMF-SOI Comparison: Number of Items Needed for 1980 SOI Basic (Form 1120) Program

Return Form and Schedule	Number of Items			Description of Item
	BMF Transaction Tapes	SOI Program	Needed	
Form 1120:				
Gross income.....	10	11	1	Income statement items
Deductions.......	17	18	1	
Tax credit.......	7	12	5	Tax payments and credits
Schedule A.......	-	1	1	Cost of goods sold schedule
Schedule C.......	-	10	10	Dividends schedule
Schedule I.......	-	4	4	Special deductions
Schedule J.......	13	15	2	Tax computation schedule
Question G(2)....	-	1	1	Entertainment, gifts, etc.
Assets...........	6	20	14	Balance sheet items
Liabilities......	4	11	7	
Schedule M-1.....	-	3	3	Net income per books/tax exempt interest
Schedule M-2.....	-	3	3	Distributions
Subtotal.....	57	109	52	Sum of Form 1120 items
Form 1120 attachments:				
Schedule D.......	3	7	4	Capital gains and losses
Form 4562........	1	6	5	Depreciation
Form 1118........	-	9	9	Foreign tax credit computation
Form 4874........	-	13	13	Credit for work incentive (WIN) program expenses
Form 4626........	-	22	22	Minimum tax computation
Form 5884........	-	39	39	Job credit
Form 3468........	-	12	12	Investment credit computation
Form 3468B.......	1	24	23	Computation of business energy investment credit
Form 4136T.......	2	9	7	Credit or refund of federal tax or gasoline, diesel, or used in qualified taxicabs
Subtotal.....	7	141	134	Sum of Form 1120 attachment items
Total........	64	250	186	Total, Form 1120 SOI items

Table 3.--Comparison of BMF Transaction Tapes and SOI Data

[Money amounts are in thousands of dollars]

Description of Item	Recent (1980) Study				Earlier (1971-72) Study			
	SOI Transcript Edit Sheets	BMF Transaction Tapes[1]	Absolute Difference	SOI as a Percent of BMF	SOI Transcript Edit Sheets	BMF Transaction Tapes	Absolute Difference	SOI as a Percent of BMF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of returns	50	50	-	-	194	194	-	-
Net receipts	895,733	893,495	2,238	100	135,697,345	158,880,218	23,182,873	85
Cost of goods sold	601,298	637,633	36,335	94	79,935,885	102,463,119	22,527,234	78
Total dividends received	3,990	3,993	3	100	2,760,648	2,423,486	337,162	114
Interest on U.S. Government obligations	10,712	10,712	-	100	467,713	466,200	1,513	100
Other interest	76,733	79,903	3,170	96	5,608,109	5,448,611	159,498	103
Gross rents	705	781	76	90	876,446	4,654,345	3,777,899	19
Royalties	24	24	-	100	869,073	900,174	31,101	97
Net capital gains (net short/long-term)	4,351	4,351	-	100	646,657	668,724	22,067	97
Other income	10,210	15,404	5,194	66	1,507,874	8,505,274	6,997,400	18
Total income	397,188	359,766	37,422	110	69,048,592	78,611,944	9,563,352	88
Compensation of officers	14,161	14,139	22	100	479,309	486,815	7,506	98
Salaries and wages	45,505	47,946	2,441	95	8,233,662	11,126,128	2,892,466	74
Repairs	1,650	1,650	-	100	677,999	4,975,543	4,297,544	14
Bad debts	4,420	4,420	-	100	652,471	673,368	20,897	97
Rents paid	7,439	7,417	22	100	2,126,640	2,257,307	130,667	94
Taxes paid	16,443	13,956	2,487	118	5,038,377	4,662,840	4,159,003	108
Interest paid	93,139	74,308	18,831	125	6,786,928	6,398,010	388,918	106
Contributions	1,052	1,052	-	100	68,165	69,771	1,606	98
Surtax exemption	-	-	-	-	1,589	1,612	23	99
Amortization	4	29	25	14	106,633	48,749	57,884	219
Depreciation	40,104	9,249	3,085	434	7,922,046	7,749,245	172,801	102
Depletion	293	283	10	104	1,691,509	1,704,820	13,311	99
Advertising	10,767	10,572	195	102	2,067,787	2,094,139	26,352	99
Pension, profit-sharing, stock bonus, and annuity plans	7,819	7,828	9	100	2,065,177	2,031,198	33,979	102
Other employee benefit plans	6,706	4,403	2,303	152	1,123,700	1,114,408	9,292	101
Other deductions	74,379	89,631	15,252	83	12,682,988	18,549,426	5,866,438	68
Total deductions	321,296	286,450	34,846	112	56,406,413	66,039,723	9,633,310	85
Net operating loss deduction	1,065	1,062	3	100	47,454	11,311	36,143	420
Special deductions (computed)	-	-	-	-	711,259	730,044	18,785	97
Income subject to tax	70,809	-	70,809	-	12,696,825	13,631,884	935,059	93
7004/7005 credit	1,806	1,806	-	100	629,650	649,327	19,677	97
Net estimated tax payments	20,743	20,743	-	100	2,985,314	2,983,545	1,769	100
Credit from regulated investment companies	-	-	-	-	19	189	170	10
Tax due/overpayment	1,001	988	13	101	166,576	163,508	3,068	102
U.S. tax on gas and lubricating oil	-	-	-	-	1,322	1,320	2	100
Foreign tax credit	-	-	-	-	2,683,778	2,304,224	379,554	116
Investment credit	3,148	3,146	2	100	325,690	322,827	2,863	101
Personal holding company tax	-	-	-	-	-	20	20	-
Tax from recomputing investment credit	68	68	-	100	12,529	12,924	395	97
WIN credit	-	-	-	-	16	15	1	107
Long-term gain	4,351	4,351	-	100	593,109	583,710	9,399	102
Minimum tax	29	-	-	-	19,584	19,494	90	100
Income tax	35,439	34,574	865	103	6,395,720	5,991,092	404,628	107
Net ordinary (gain/loss)	8,045	8,143	98	99	-71,765	-67,409	4,356	106

[1] Because BMF transaction tapes data were not available, these data are simulated based on the revenue processing instructions for transcribing data to the transaction tapes.

Table 4.--Frequency Distribution of Agreement Between BMF and SOI Items, 1971-72 Tax Year Sample

Description of Item	Number of Returns				Percent			
	Both Equal	Both Blank	Total (1)+(2)	Dif-ferent	Both Equal	Both Blank	Total (5)+(6)	Dif-ferent
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Net receipts	75	84	159	35	39	43	82	18
Cost of goods sold	42	103	145	49	22	53	75	25
Total dividends received	120	64	184	10	62	33	95	5
Interest on U.S. Government obligations	92	101	193	1	47	52	99	1
Other interest	153	14	167	27	79	7	86	14
Gross rents	112	66	178	16	58	34	92	8
Royalties	50	142	192	2	26	73	99	1
Net capital gains (net short/long-term)	101	84	185	9	52	43	95	5
Other income	60	34	94	100	31	18	49	51
Total income	82	7	89	105	42	4	46	54
Compensation of officers	147	42	189	5	76	22	98	2
Salaries and wages	129	37	166	28	66	19	85	15
Repairs	98	79	177	17	51	41	92	8
Bad debts	137	52	189	5	71	27	98	2
Rents paid	140	45	185	9	72	23	95	5
Taxes paid	137	14	151	43	71	7	78	22
Interest paid	144	35	179	15	74	18	92	8
Contributions	67	125	192	2	35	64	99	1
Surtax exemption	6	-	-	-	3	-	-	-
Amortization	25	146	171	23	13	75	88	12
Depreciation	104	40	144	50	54	20	74	26
Depletion	28	161	189	5	14	83	97	3
Advertising	128	53	181	13	66	27	93	7
Pension, profit-sharing, stock bonus, and annuity plans	141	46	187	7	73	24	97	3
Other employee benefit plans	95	86	181	13	49	44	93	7
Other deductions	59	20	79	115	30	10	40	60
Total deductions	77	22	99	95	40	11	51	49
Net operating loss deduction	3	177	180	14	2	91	93	7
Special deductions (computed)	50	95	145	49	26	49	75	25
Income subject to tax	1	73	74	120	5	38	38	62
7004/7005 credit	82	107	189	5	42	55	97	3
Net estimated tax payments	-	185	185	9	-	95	95	5
Credit from regulated investment companies	67	127	194	-	35	65	100	-
Tax due/overpayment	-	-	190	4	-	-	-	-
U.S. tax on gas and lubricating oil	-	-	-	-	-	-	-	-
Foreign tax credit	-	-	192	2	-	-	99	1
Investment credit	-	-	191	3	-	-	98	2
Personal holding company tax	-	-	-	-	-	-	-	-
Tax from recomputing investment credit	-	-	-	-	-	-	-	-
WIN credit	-	-	-	-	-	-	-	-
Long-term gain	79	63	142	52	41	32	73	27
Minimum tax	-	-	-	-	-	-	-	-
Income tax	87	90	177	17	45	46	91	9
Net ordinary (gain/loss)	83	101	187	10	43	52	95	5

NOTE: Total population is 194 returns.

Table 5.--BMF-SOI Agreement on Major Industry, Tax Year 1972

Major SOI Industry	Percent Agreement
ALL INDUSTRIES	74.9
AGRICULTURE, FORESTRY, AND FISHING	78.3
MINING	87.7
Metal mining	92.7
Coal mining	94.4
Crude petroleum and natural gas	86.1
Nonmetallic minerals (except fuels) mining	85.4
CONTRACT CONSTRUCTION	89.2
MANUFACTURING	72.8
Food and kindred products	75.0
Tobacco manufactures	79.0
Textile mill products	79.0
Apparel and other fabricated textile products	87.3
Lumber and wood products, except furniture	83.0
Furniture and fixtures	70.1
Paper and allied products	78.9
Printing and publishing	86.4
Chemicals and allied products	79.8
Petroleum refining and related industries	57.9
Rubber and miscellaneous plastics products	63.3
Leather and leather products	63.8
Stone, clay, and glass products	75.0
Primary metal industries	62.7
Fabricated metal products, except machinery	75.2
Machinery, except electrical	50.8
Electrical equipment and supplies	67.2
Motor vehicles and equipment	80.1
Transportation equipment, except motor vehicles	53.7
Scientific instruments, photographic equipment, watches and clocks	63.2
Miscellaneous manufactured products	62.1
TRANSPORTATION, COMMUNICATION, ELECTRIC, GAS, AND SANITARY SERVICES	70.6
Transportation	68.8
Communication	86.3
Electric, gas, and sanitary services	67.9
WHOLESALE AND RETAIL TRADE	75.4
Wholesale trade:	
Groceries and related products	54.1
Machinery, equipment, and supplies	36.6
Miscellaneous wholesale trade	75.8
Retail trade:	
Building materials, hardware, and farm equipment	79.7
General merchandise stores	69.3
Food stores	85.1
Automotive dealers and service stations	82.8
Apparel and accessory stores	83.6
Furniture, home furnishings, and equipment stores	67.7
Eating and dining places	88.2
Miscellaneous retail stores	73.7
Wholesale and retail trade not allocable	3.7
FINANCE, INSURANCE, AND REAL ESTATE	75.8
Banking	91.3
Credit agencies other than banks	57.9
Security and commodity brokers, and dealers	68.5
Holding and other investment companies	31.2
Insurance carriers	81.0
Insurance agents, brokers, and service	82.7
Real estate	86.6
SERVICES	71.6
Hotels and other lodging places	83.9
Personal services	90.0
Business services	42.2
Automobile services and miscellaneous repair services	59.7
Amusement and recreation services	81.0
Other services	88.1

Appendix

DATA DEVELOPMENT FOR STATISTICS OF INCOME

Abstracted from a paper by Alan Freiden

There are two interrelated problems that have made it difficult for the Statistics of Income (SOI) program to perform its mission. First, the data development system is too inflexible to respond quickly as new requirements for analytical data arise. This is due, in part, to the ad hoc character of the various stages of the data development process. The present SOI approach is elaborate and unwieldy having grown up over many years. In a sense there is no real system design; there is merely a 65 year "history."

The other major problem facing Statistics of Income is that it is extremely difficult to relate raw tax return data to the conceptual measures that are ultimately of interest. Traditionally, the SOI program has avoided performing a pure data capture function choosing instead to apply a complex set of consistency tests, corrections, and imputations whose effects on the conceptual meaning of the final data products are very difficult to analyze. It is not clear, for example, whether the errors that are being corrected are real (that is, they are made by the taxpayers), conceptual (because the data do not conform to a particular analytical need), or random (data transcription errors, for example). Unfortunately, until recently, error resolution has taken place without access to the original tax documents. The result, then, is that the SOI program is producing data whose conceptual meaning has not always been clearly and explicitly defined and whose analytic content could not be changed quickly even if there were such a precise definition.

A TWO-TIER DATA SYSTEM

The first job of a government statistical agency is to preserve and organize the basic information available from various documents, surveys, or administrative records. One role for SOI, therefore, is to preserve the facts and figures about the operation of the Tax System. In order to give content to this idea we might begin by treating tax return data as an abstract notion that is divided into three distinct and logically separate ideas. These categories are measured (or raw) data, conceptual data, and the physical representations of the first two forms. There are significant advantages to be gained from making these distinctions but that is something extremely difficult to do in everyday practice. This section will define these categories of data and will describe how an alternative data system could be designed to support them.

The first category of data contains the basic information that is available for SOI but has not been collected expressly in support of its purposes. These data are from sources such as income tax returns, corporate financial reports, reports to government agencies for regulatory purposes, and surveys (of both businesses and individuals). This information may be written documents, computer tapes, or other physical forms. Of special interest are computerized data files that are produced for administrative purposes. Examples of such data are the IRS Master files for individual and business returns.

This basic information is the raw material for the production of the conceptual and analytical information that is actually of interest. Conceptual data items implement theoretical notions that may not appear directly in measureable economic transactions or in explicit financial reports. Therefore, it is necessary to construct them from the data that is available. Such constructions may range from the simple rearrangement or combination of measured data items to purely speculative or even counterfactual constructions that are of an "as if" nature. All forecasts, for example, are of this type. An example of how information from the same sources may be used to produce measurements of very different theoretical concepts are the many definitions of corporate profits: the tax concept, the national income accounting concept, and the book measure. All of these variables rest on the same corporate accounting information but they implement different analytical constructs.

The final category of data is its physical representation. Most conceptual data is published either in machine readable form or as ordinary printed documents. Most measured data is available in the same forms. The goal of the SOI program, therefore, should be to record and organize as much of the measured data as possible, store it in machine readable form, and then construct (or aid in the construction of) alternative sets of data from different conceptual definitions that are of use in research and the policy process.

The Data Development Process

The first of the three components of the data development process is data capture. Here, administrative data (Master Files) as well as supplementary items from the tax returns are encoded into machine readable form. However, there may be information from another tax year or even from outside the tax system altogether that is thought to be of interest at a later stage. This information should also be captured. The encoding of data items should be as mechanical (which is not synonymous with simple) as possible at this stage. Whether the basic data is encoded by humans or machines, the only modifications that should be made to the basic data involve the resolution of individual data items and the treatment of special cases. The goal at this stage should be to keep the amount of non-basic or a priori information that might be introduced by discretionary decision makers to an absolute minumum.

This means that the encoded data will be full of "errors." Since the definition of an error may rest on the analytical purposes to which the basic data will later be applied, it is better to leave the "errors" in so that they may be adjusted in very different ways at the next stage of the data development process.

Once the raw data have been encoded they must be stored into a database. The database will be most useful if it includes not only the raw data themselves but also descriptive information as well as basic data from other sources. The descriptive information (data about the data) should include a data dictionary of the stored items, a user's guide to the use of the database, and a narrative history of the data development process. This database is, in one sense, a final product since it is the representation of the tax system that will be seen by most analysts. Its quality as a representation of the tax system will be testable. That is, can the original administrative documents and tax returns be reproduced using this database? If so, then it is an accurate model of the system. If not, then noise has entered the system and the quality of the basic data capture process needs to be improved.

The final stage in the data development process is the mapping of the basic database into a number of alternative conceptual-analytical databases. These mappings are separate computer programs that calculate conceptual measures from the basic data items. These mappings, which are implementations of theoretical economic or accounting models and may be written by outside users, will be very different. However, it is likely that only a concise analytical core will differ from application to application so it will be the role of the Division to provide the generalized modules that will be part of all conceptual mapping programs. Examples of such "software tools" are:

1) A formal language processor for the definition of consistency tests and error resolution calculations;
2) Statistical tools for computing imputations;
3) Optimized procedures for data storage and retrieval;
4) Generalized report generation routines.

If successful, this approach will lead to the formation of a community of users who would be encouraged to share their skills, tools, and special data sources with each other.

Results

The system that results from following the approach outlined above would have two distinct and separable tiers. The lower tier would interface directly with the IRS tax return processing system and would capture as much basic information as possible. At this level, the goal should be to limit the amount of a priori information that is embodied in the basic observations (that is, the data developers should make as few corrections as possible) while expanding the amount of outside information that is made part of the stored data. Such outside information, which will be used by the computer programs performing the mappings onto various conceptual definitions, is contained , say, in corporation financial statements and previous year basic tax return (and conceptual) data. The second tier, then, is the set of conceptual databases that are produced by subject matter experts manipulating the raw information available on the first tier.

The benefits from viewing the data development process this way and splitting it into its three separate stages would be many. First, some basic data would be available to users more quickly. In fact, preliminary versions of the basic database could be made available even before being completed. As long as users knew that important observations were missing (the largest corporations, for example), they would still proceed with analyses. In addition, they would know that the logical structure of the information in the database (if not its content) is stable. This means that the only possible changes to the database would be the addition of new or updated observations as they became available. Next, the mapping of the basic data into alternative conceptual forms would be extremely flexible. This is clearly useful but it also provides a mechanism for investigating the impact of applying a priori information and discretionary data corrections to the basic data. Finally, since the production of the SOI publication would itself be a major second tier project using the basic database, there would be a large number of resident users who could critique the development of the earlier stages of the data development process.