RESULTS OF COVERAGE AND PROCESSING CHANGES TO THE 1980 INDIVIDUAL STATISTICS OF INCOME PROGRAM

Peter Sailer, Charles Hicks, Dave Watson, and Dan Trevors
Internal Revenue Service

This paper reports on alternative methods employed to close out the statistical processing of the Tax Year 1980 sample of individual income tax returns (Forms 1040/1040A) used for the Statistics of Income (SOI) program. Particular emphasis will be given to the impact of these alternatives on the early "Advance Data" tabulations that are based on an early cutoff of sample receipts for a given tax year. These early tabulations are required annually, for budgetary and tax policy reasons, by the Treasury Department's Office of Tax Analysis and the Congressional Joint Committee on Taxation and are needed no later than the end of November of the year in which the returns are filed.

Organizationally this paper is divided into three sections: (1) an overview of the evolution of SOI individual income tax return statistical processing system, (2) analysis of results of the early cut-off for 1980 Advance Data production processing, and (3) a discussion of subsequent enhancements and research activities of the Statistics of Income Division for the individual income tax return statistical processing system as well as a few suggestions for areas of possible future improvements.

1. RECENT HISTORY OF SOI PROCESSING

Data processing needs of the Statistics of Income program are accomplished as a byproduct of the IRS responsibility to process the income tax returns filed for tax administration purposes. Accordingly, the processing of a statistical sample from these returns is at best a secondary concern to the Service. To the extent possible, statistical needs are incorporated into the mainstream of revenue processing procedures to minimize disruption of the administrative processing of tax returns and to reap the greatest possible benefits from data already entered into the IRS master file system for administrative purposes.

Currently, the individual income tax return sample is identified from "transaction tapes" prepared by the ten IRS service centers for the Individual Master File (IMF). Prior to Tax Year 1974, all of the data used in the Form 1040 SOI program were manually abstracted or edited onto hardcopy edit sheets from the tax returns as a separate off-line statistical processing operation. For Tax Year 1974, the first attempt was made to use IMF data for SOI. For that year, a limited number of codes and amounts were computer printed onto edit sheets for returns in the sample using the transaction files. These data were subsequently reviewed by statistical clerks for statistical acceptability. At the same time, additional data were needed for statistical purposes, but were not available from the transaction files. Such data were abstracted from the tax return edited as necessary. As confidence and experience were gained in the usability of the IMF data, more SOI data came to be based on this source. During this evolution, enhancements were added to this new system, such as computer checking the validity of the statistical data while the returns were still on hand. Unfortunately, as enhancements were added, the ability to meet interim processing deadlines was strained. In fact, meeting some of these deadlines became impossible because of the processing lags these modifications created. Therefore, for Tax Year 1980 a major revision was made to one of the most critical dates.

This change accelerated the sample cut-off date by two weeks, effectively moving up the ending date for inclusion of sampled returns in the Advance Data report from mid-October to the end of September. In terms of the number of returns involved, accelerating the cut-off date by two weeks apparently excluded 2,200 return records from the sample of 160,133 returns. As it turned out though, the number of returns excluded from the sample because of this early cut-off, as compared to those under the previous years' early cutoffs, were approximately the same. Traditionally, any designated sample return records for which revenue processing was incomplete (at that given point of time) were omitted from the Advance Data file, thus resulting in Advance Data tabulations that were actually based on returns designated for the sample by mid-September or earlier.

The new effort, for Tax Year 1980, was designed to assure that all returns designated through mid-September were processed and shipped on tape for statistical processing (at the IRS Data Center, in Detroit) within a time frame that was approximately two weeks earlier than previous years. Most of the "final" returns shipped to the Data Center at cutoff time for inclusion in Advance Data included only transaction file data. Time did not permit any perfecting of these data, or the manual editing of data not available from the transaction file. Because of the missing data, these records would not pass the computerized validity tests used to further process the records into a form from which tables could be produced. Therefore, portions of the missing data had to be imputed and these imputations were limited to those necessary to enable the record to pass all of the tests. The imputed amounts were determined, if possible, from data present on the record. Otherwise, amounts were estimated on a proportional basis, using data available from returns for which the manual editing had been completed. For example, net

capital gain or loss was the amount on the transaction file and was therefore the only capital gain amount available to complete processing. However, details on long- and short-term capital gains, which normally would have been manually edited from the return, were needed. If the alternative minimum tax had been used by a taxpayer, this tax figure was available from the transaction files, and since the long-term capital gain excluded from adjusted gross income is one of the items used to compute alternative minimum tax, the gain amount could be determined by working backwards from the tax. If alternative minimum tax was not present, or the entire gain amount could not be determined using this method, long-term or short-term gain was determined based on the proportion of each on similar returns for which these amounts had been manually edited.

For the later SOI complete report, a more detailed and sophisticated set of imputation factors was developed to cover the manually-edited items for these returns. Again, characteristics present on similar returns were taken into account in making the imputations. If these proved inadequate, then distributions were made on a proportional basis using as a guide the 1980 "Advance Data," but only for fully processed returns.

The following section describes the results of the early cutoffs for Tax Year 1980. It should be noted that these procedures were not undertaken simply on faith. Rather, a series of tables was produced, using the 1978 Statistics of Income File, in which the dropping of late-filed returns and the weighting of earlier ones was simulated. In a paper written about this simulation [1], James Dumais and Ray Shadid concluded that a cycle 36 (mid-September) cutoff should be adequate for producing basic income and tax estimates, and that an earlier cutoff for the Complete Report was also feasible.

2. ANALYSIS

This section presents an analysis of the effects on Advance Data and on SOI of advancing the closeout date and including imputations and "unedited" transaction file data in the statistics for 1980. The tabulations discussed are shown at the end of this paper.

It is important to note that these two strategies for expediting the data actually involved only a very small proportion of the total sample. When the transaction file data are used in combination with imputations, the manual editing step can be bypassed and the return record processed immediately through the service center directly to the IRS Data Center. If the return comes in very late, it may not be included at all in the file used to prepare the Complete Repoort. However, its absence will be offset by weighting the existing returns already on file. As is shown in figure A, only 0.9 percent of the final estimate for adjusted gross income was based on returns processed in this manner; only 0.2 percent of the final estimate was derived by assigning higher weights to sampled returns to compensate for returns not yet sampled. Not surprisingly, returns "forced" through the system were concentrated in the highest income and in the deficit classes.

For the Advance Data closeout, the

Figure A.--Complete Report: Percent of Total Estimate of Adjusted Gross Income by Estimation Method

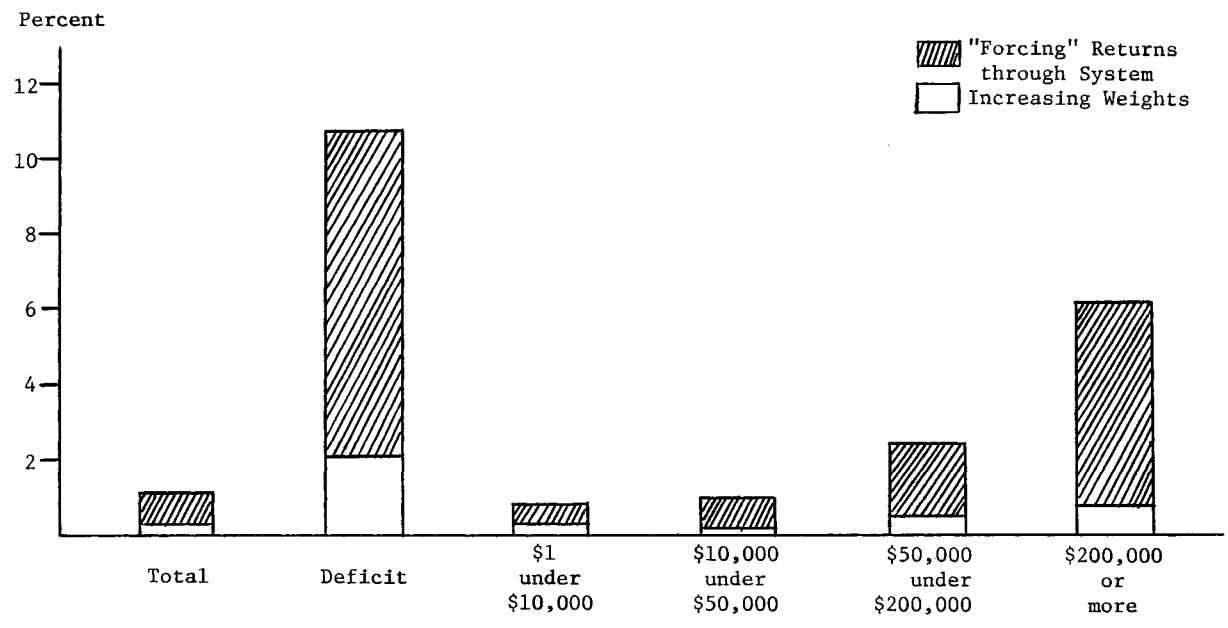
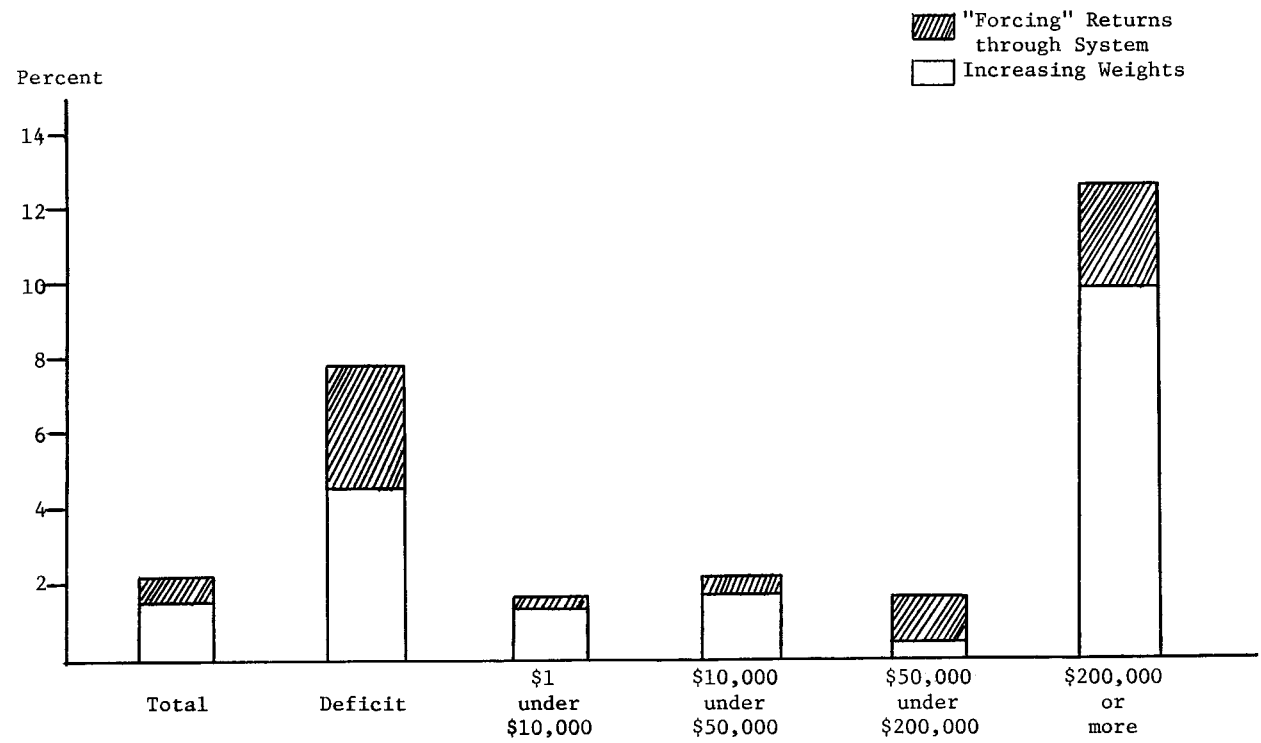


Figure B.--Advance Data: Percent of Total Estimate of Adjusted Gross Income by Estimation Method



proportion of returns forced through the system was slightly lower than that for the later SOI Report--0.6 percent vs. 0.9 percent for the SOI Report. Obviously, as can also be seen from Figure B, the portion of the Advance Data estimates derived by weighting early returns to replace later returns was much higher--1.5 percent overall (0.2 percent for the SOI Report); 9.8 percent for high-income returns (0.6 percent for the SOI Report).

The strategy for expediting publication of data for 1980 consisted simply of moving up the closeout date by two weeks for Advance Data (to mid-September) and by four weeks for the later SOI Report (to the end of November). Since the number of returns processed for a year are so close to completion by mid-September, let alone by late November, it would not appear that cutting off two weeks earlier would have a significant impact on the results. In fact, all of the items in the SOI Complete Report (prepared from the file with the late November closeout) were within a fraction of a percent of the data from the final file, as is shown in Table 1. (The final file was one created after all returns selected for the sample during Calendar Year 1981 had been processed through the system.) As would be expected (given the overlap between the two samples), all of the differences shown in Table 1 are much less than the expected sampling variability at the one standard deviation level of significance.

Table 2 shows a comparison between all of the items customarily shown in Advance Data Reports with the comparable items from the corresponding Complete Reports, for Tax Years 1976 through 1980. For the two most basic data items -- adjusted gross income and total tax liability -- the average differences between the early and Complete Report estimates were 0.1 and 0.2 percent, respectively. For 1980, the differences were 0.5 percent for adjusted gross income, and 0.8 percent for total tax liability, a moderate increase over the previous years, but still, it would appear, perfectly acceptable as early estimates. (It should be noted that the increased differences are due only in part to the earlier cutoff for returns processing. An unexpected surge in the number of returns with very high income filed during December of 1981 also led us to understate the weights for the top income class in producing the 1980 Advance Data estimates).

However, while the results were reasonably encouraging for the basic items, certain rarer items presented a real problem. Table 2 also compares Preliminary and Complete Report data for two items that are relatively rare, that are somewhat complicated for the taxpayer to compute, and that have been subject to frequent tax law changes in recent years: the minimum tax and the alternative minimum tax. Quality of the data for these items for the years with an October 1 Advance Data cutoff was relatively

poor. The mid-September cut-off for 1980 appears to have made matters worse.

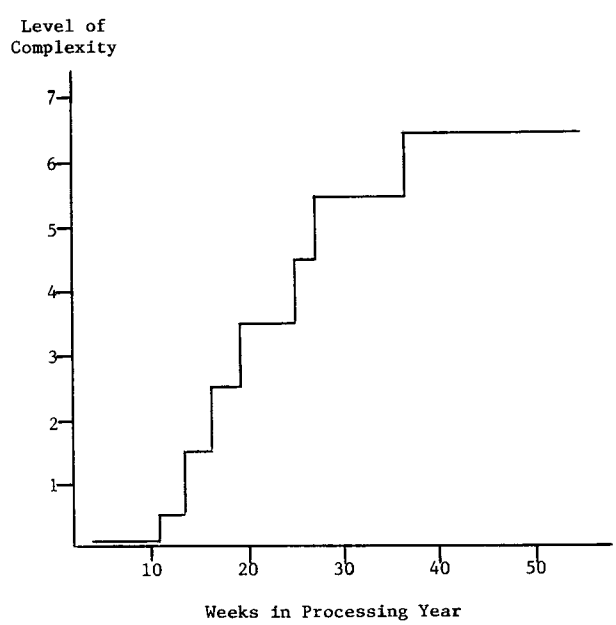
Obviously, even though a very small percentage of returns is filed after mid-September, the characteristics of these returns differ sigificantly from those of the other returns, so that weighting the other returns to compensate for the returns filed later in the year is not the complete answer.

The differences discussed so far are attributable to the fact that late returns tend to be different from early returns. One reason for the difference is that returns filed later in the year tend to be more complex than those filed earlier. This can be seen clearly from Figure C, which shows the level of complexity of returns by the week in which they were processed. The measure used to indicate complexity is the number of schedules attached to the basic Form 1040.

As can be seen from Figure C, the median value of this measure of complexity remains at less than one (schedule) through the end of March (week 12 of the Processing Year); then rises steadily through mid-July (week 27). Between mid-July and mid-September (weeks 27 through 36), it remains fairly constant at between five and six (schedules), and then rises once more to between six and seven. The median value for the year as a whole is 3.5 (schedules).

The first recommendation for future improvement is to return to an October 1 cutoff, and make up the two weeks in other ways -- specifically, by reducing the amount of time it takes to edit returns and by printing out from the transaction files computer-generated codes that tell the editors precisely what other information to look for on the returns, based on the data available from the transaction files, and which data items need to be separately abstracted or edited from the return. Thus, for instance, a code based on the data available from the transaction files could advise the editor to go to the investment credit schedule to obtain additional data, based on the presence of the investment credit on the transaction file. Such a system has been instituted for tax year 1981, as is further explained in the next section.

Figure C.--Median Level of Complexity by Processing Week



The second recommendation is not to weight the very earliest (and simplest) sample returns to compensate for the later, complicated ones. Rather, we should increase the weights primarily on the returns that come in after June 1, which is the time by which most timely-filed returns have been processed. The trade-off here is that, while reducing the bias that is built in by weighting early returns to make up for late ones, we would also increase the sampling variability of the estimates by applying unusually large weights to a small group of late returns. Nonetheless, when looking at an estimate like alternative minimum tax which is off by 28 percent, even though expected sampling variability is only 3 percent, the trade-off appears acceptable.

The second strategy for expediting processing was the so-called "forced" processing of data from the IMF directly into the statistical file. What was sent to the IRS Data Center in these cases was a record that had only transaction file data. Since they comprised a very small proportion of the total sample, the data for these returns had little impact on the quality of the totals presented. However, there are a number of items needed for SOI not available from the transaction files and, for the "forced" returns, these items had to be imputed. Later, a small subsample of these returns was retrieved and fully processed in order to evaluate the quality of the imputations. These differences are summarized below.

Figure D shows that, for items where the IMF transaction data offered a number of clues which improved the imputations, the imputations were quite valid. For instance, in the case of net long-term capital gains, we had the advantage that the missing item had to balance not only to net capital gain in adjusted gross income, but also to the alternative minimum tax. For many returns, there was in effect only one plausible imputation and the method of imputation used proved quite accurate. On the other hand, the imputed credit card interest deduction was based on the proportion of the total interest deduction that came from this source on returns processed earlier in the year. This proportion turned out to be invalid because the returns processed later in the year tended to have much larger interest deductions, of which credit card interest was a much smaller proportion. If an imputation for credit card interest is needed for future SOI programs (it is not part of the 1981 SOI program), the percentage will need to be varied by income class and also possibly by processing week [2]. It is important to note that, even though the imputations for such items as credit card interest and net short-term capital loss were not very accurate, because the imputed

Figure D. Imputed Data as a Percent of Fully Processed Data

Item	Imputed as a Percent of Final Estimate 1/	Potential Distortion of Final Estimate 2/
General sales taxes deducted	83.6	-0.2
Personal property taxes deducted	93.1	-0.1
Credit card interest deducted	207.0	0.3
Union dues deducted	133.2	0.2
Net short-term capital gain	99.6	-0.0
Net short-term capital loss	27.6	-4.7
Net long-term capital gain	99.3	-0.0
Net long-term capital loss	98.4	-0.0
Personal service gross income	97.2	-0.1
Deductions from personal service gross income	157.5	0.7
Personal service net income	96.8	-0.1

1/ Obtained from 103 returns for which data were first imputed, then abstracted and keyed under regular procedures.

2/ Assuming the same processing (non-sampling) error for all the remaining unedited returns as for the sample of 103, this is the percentage by which the published estimate could be off.

amounts represented such a small percentage of the final estimates, the affect on the quality of the final estimates was quite small.

In addition to developing improved imputations for future years, another change instituted which should improve the quality of any future "forced" return estimates is the incorporation of even more data items from the transaction files into the SOI files. Past constraints on the number of transaction file items that could be incorporated into the SOI system have now been eliminated by other changes made to the SOI processing system.

In conclusion, we are reasonably satisfied with the results of the early closeout and forced processing strategies instituted for the 1980 SOI program. Both strategies were applicable to only a small percentage of the total returns and, in terms of cost-benefit, the gain in timeliness of the SOI data was considerable.

3. FUTURE RESEARCH EFFORTS

Evolution of the 1040 Statistics of Income (SOI) system, with its expanded use of transaction file data, diminishes the need for a manually prepared edit sheet for every sample return selected. In fact, unpublished analyses conducted within the Division indicate that most Form 1040A data could be accepted at "face value" from the transaction files without the need for visual inspection at the service centers. For Tax Year 1981 (Filing Year 1982), the structured fixed-design edit sheet has now been eliminated. Because of the physical limitations of printing transaction file data on fixed format edit sheets, only 32 percent of the data items needed for SOI could previously be obtained from the IMF. Elimination of the fixed design edit sheet now allows us to obtain 63 percent of the total SOI items from the IMF. Instead of the edit sheet, data are read out onto ordinary computer printout paper, but only if computer tests or checks determine that there are data inconsistencies or indicate that additional data need to be obtained from the return by the statistical editors. Return records that pass the tests and checks and consequently require no additional statistical editing are processed directly to the tapes to be sent to the Data Center. Figures through June 1982 show that approximately 28 percent of the 1981 returns selected for SOI have been processed directly to the Data Center tape file. Returns processed within this time frame are the "early filers" and consist typically of Form 1040A and the simpler Form 1040 returns. During the last half of the year the more complex and prior-year Form 1040 returns will be processed by the Service and then selected for the SOI sample. Thus, a decrease in the percentage of returns (records) processed directly to the Data Center file can reasonably be expected. Under the new system, statistical editors can now focus their attention directly on those returns requiring review and only to the specific area or areas within the return record in need of scrutiny. In addition, the amount of time spent on batching and controlling has been reduced significantly.

Simplified systems and record design for source data capture permit simplified transcription techniques to be used. IRS employs a unique direct data entry system (DDES) designed for the administrative processing of tax returns that possess generalized parameter driven (GPP) transcription capabilities for any off-line processing. GPP, while effective, is not necessarily efficient. It is however, the medium that must be used to meet SOI transcription needs. Conversion to the equivalent of an unstructured edit sheet and resultant simplification of the processing system has resulted in an approximate 600 percent increase in the transcription rate. Nominal resources are also saved in the areas of paper, printing, and computer time. Eliminated then is the structured edit sheet whose function had evolved from that of an essential source data capture document to one of an intermediate (or lesser) role of data display.

For Tax Year 1981, sample receipts for both Advance Data and SOI report processing will be cut off early. Unlike Tax Year 1980, the Advance Data cutoff has been automated to the extent that data processed through mid-

October, as opposed to early October, can be included without compromising the delivery dates. Plans are to cut off the sample for the SOI Report processing at about the beginning of December. Sample designation will continue, however, through the end of Calendar Year 1982. Any returns thus excluded, but which have data characteristics whose absence from the sample could bias the results, can be introduced into the file during the processing at Detroit. These returns include, but are not restricted to, high income nontaxables, large adjusted gross income or deficit returns, or returns with a large amount for any specified data item.

One further change introduced for the 1981 program is the combined processing, testing, and correction of individual (1040) data and sole proprietorship (Schedules C and F) data [3]. As a result, the previous practice of splitting off these two files, controlling them separately, and then recombining them will no longer be necessary.

Areas for future research include possible telecommunication of data between the service centers and the Data Center as opposed to conventional shipping methods now used. Consideration is also being given to alternative methods of handling prior-year returns, currently included in the sample as "stand-ins" for delinquent current-year returns yet to be filed. If such returns could be eliminated from the sample, a good deal of "exception processing" could be avoided. Finally, feasibility studies for modifying the on-line error resolution (currently planned for tax administration purposes) so that it can be applied and adapted to SOI processing is also under way.

The Statistics of Income Division, like most Federal statistical organizations, is increasingly faced with budgetary constraints. Future budgetary constraints may have to be met by greater use of the IMF data, computerized imputation and correction routines, and elimination of possibly superfluous manual functions. With the increasing trend by taxpayers to file as late in the calendar year as is legally possible, the early cut-off of the sample may prove to be only an interim solution. Long-range strategy indicates that streamlining and standardizing of procedures is the direction in which SOI Division must proceed.

ACKNOWLEDGMENTS

The authors would like to thank their many colleagues in the Internal Revenue Service who helped in the preparation of this report. In particular, thanks to John DiPaolo, Keith Gilmour, and Robert Wilson for their extensive review and many, much-needed suggestions for improvements. Thanks to Sylvia Martin of the IRS Data Center for her help in creating the files needed to produce the analytical material, and to Wendy Alvey, Clementine Brittain, Beth Kilss, and June Walters for their aid in preparing the charts, both for the written report and for the version presented at the meetings. Thanks also to Mary Haigler and Joyce Coleman for typing the many drafts of this paper. Blame for any shortcomings should be attributed in equal measure to each of the authors.

NOTES AND REFERENCES

[1] Dumais, James and Shadid, Raymond, "Individual Statistics of Income: Advancing the Closeout Date," 1981 American Statistical Association Proceedings, Section on Survey Research Methods.

[2] For other possible approaches to imputing missing data, see Hinkins, Susan, "Imputation of Missing Items on Corporate Balance Sheets," 1982 American Statistical Association Proceedings, Section on Survey Research Methods.

[3] Wolfe, Raymond, Methodological Changes in the Statistics of Income Sole Proprietorship Programs--Dominant Business Processing, unpublished working paper available from the Statistics of Income Division, Internal Revenue Service.

[4] Blacksin, Jack and Plowden, Raymond, "Statistics of Income for Individuals: A Historical Perspective," 1981 American Statistical Association Proceedings, Section on Survey Research Methods.

[5] Wilson, Robert A., and DiPaolo, John, "Statistics of Income: An Overview," 1981 American Statistical Association Proceedings, Section on Survey Research Methods.

[6] Internal Revenue Service, Statistics of Income--1980 Individual Income Tax Returns, Washington, D.C. 1982.

[7] Internal Revenue Service, Statistics of Income Bulletin, Winter 1981-82, Washington, D.C. 1982.

Table 1.--Selected Income and Tax Items, Complete Report and Final File, Statistics of Income-1980, Individual Income Tax Returns

(All figures are estimates based on samples--money amounts are in thousands of dollars)

Item	1980		Percent Difference
	Complete Report	Final File	
Total Number of Returns	93,902,469	93,902,441	0.0
Adjusted Gross Income (less deficit)	1,613,731,497	1,613,574,098	0.0
Salaries and Wages	1,349,842,802	1,349,818,631	0.0
Dividends in AGI	38,761,253	38,722,884	0.1
Total Adjustments	28,614,061	28,616,177	0.0
Total Itemized Deductions	218,028,139	218,030,253	0.0
Income Tax before Credits	256,294,315	256,254,609	0.0
Total Tax Credits	7,215,839	7,211,436	0.0
Income Tax after Credits	249,078,475	249,043,173	0.0
Minimum Tax	412,638	413,414	0.1
Alternative Minimum Tax	850,326	854,261	0.5
Total Tax Liability	256,251,076	256,220,651	0.0
Total Taxpayments	271,501,122	271,488,923	0.0
Tax Due at Time of Filing	32,843,576	32,828,329	0.0
Total Overpayment	49,458,344	49,461,323	0.0

Table 2.--Selected Income and Tax Items: Percent Difference between Preliminary and Complete Reports, Average for Tax Years 1976 through 1979, and for Tax Year 1980

Item	Average Percent Difference, Tax Years 1976-1979	Percent Difference, Tax Year 1980
Total Number of Returns	0.2	0.3
Adjusted Gross Income (less deficit)	0.1	0.5
Salaries and Wages	0.2	0.4
Dividends in AGI	0.5	2.0
Total Adjustments	1.1	2.1
Total Itemized Deductions	0.6	1.5
Income Tax before Credits	0.3	0.8
Total Tax Credits	1.8	6.5
Income Tax after Credits	0.2	0.6
Minimum Tax	7.0	21.7
Alternative Minimum Tax	15.7*	28.1
Total Tax Liability	0.2	0.8
Total Taxpayments	0.2	0.6
Tax Due at Time of Filing	1.1	2.1
Total Overpayment	0.6	0.6

*Difference for 1979; alternative minimum tax not in effect for 1976 - 1978.