Special Studies in Federal Tax Statistics

2003

Compiled and Edited by
James Dalton and Beth Kilss*
Statistics of Income Division
Internal Revenue Service

Selected Papers Given in 2003
at the Annual Meetings of the
American Statistical Association

*Prepared under the direction of Tom Petska, Director, Statistics of Income

Special Studies in Federal Tax Statistics 2003

CONTENTS

1

Are the Rich Getting Richer and the Poor Getting Poorer?

Welniak
Strudler ◆ Petska ◆ Petska
Williams
Toder

Measuring Household Income Inequality Using the CPS

Edward J. Welniak, U.S. Census Bureau

Disclaimer: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

▶ Introduction

This paper examines the use of the Current Population Survey (CPS) to measure income inequality. It begins with a brief overview of the CPS followed by a presentation of how three income inequality measures track over time using CPS household data. It goes on to examine topcoding issues associated with CPS income data, how CPS topcoding affects the measurement of income inequality, and concludes with a discussion of CPS income data quality issues.

▶ An Overview of the CPS

The CPS is a national random household sample survey conducted monthly by the Census Bureau for the Bureau of Labor Statistics. The monthly sample size for the CPS is about 78,000 households.[1] The survey has been conducted for more than 50 years.

The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutional population. Respondents are interviewed to obtain information about the employment status of each member of the household 15 years of age and older.

Estimates obtained from the CPS include employment, unemployment, earnings, hours of work, and other indicators. They are available by a variety of demographic characteristics including age, sex, race, ethnicity, marital status, and educational attainment. They are also available by occupation, industry, and class of worker. Supplemental questions to produce estimates on a variety of topics, including school enrollment, income, previous work experience, health, employee benefits, and work schedules, are also often added to the regular CPS questionnaire.

One of the CPS's most widely used supplements is the Annual Social and Economic Supplement (ASEC).[2] The ASEC is the source of annual income, official poverty, and health coverage statistics for the U. S. The ASEC has been used to compile annual income summary measures for families and people since 1947 and for households since 1967. Households have become a more comprehensive unit of analysis over time due to changing living patterns (a smaller percentage of people currently live in family situations than 50 years ago). Household income data is constructed from income information collected about the civilian, noninstitutionalized population 15 years old and over.[3] Households exclude people living in group quarters.

Income collected in the CPS ASEC is defined as money income received on a regular basis, before deductions for taxes and other expenses, and does not include lump-sum payments or capital gains. It includes wages and salary, self-employment (net after expenses), unemployment compensation, worker's compensation, Social Security, Supplemental Security Income, cash public assistance, veterans' payments, survivor benefits, pension or retirement income, interest, dividends, rents, royalties, estates, trusts, educational assistance, alimony, child support, assistance from outside the household, and other miscellaneous money income received on a regular basis.

The income data collected in the CPS ASEC have become more detailed over time. In 1967, data were collected on eight sources of income. The 1967 sources included wages and salaries, which were one of the two original income sources asked in 1947, two sources of self-employment income (farm and nonfarm, which were added in 1950), and five additional sources added in 1967:

Social Security; interest, dividends, estate, trust, or rent; public assistance or welfare; unemployment compensation, worker's compensation, government employee pensions, or veterans payments; and private pensions, annuities, alimony, royalties, or regular contributions from people not living in the household. The number of income sources continued to expand until 1979 when the CPS ASEC allowed for the identification of over 50 income sources while recording up to 27 income values. The income sources have remained unchanged since 1979 (see Welniak for a complete discussion of the evolution of the CPS ASEC questionnaire and processing system).

In addition to an increasing number of income sources collected in the CPS ASEC, the values recorded for these sources also increased. In 1967, the format of the CPS questionnaire allowed for the recording of amounts up to $9,999 for each of the eight income sources. In 1970, the format of the questionnaire changed allowing the recording limits to increase to $99,999 for six of the eight income sources (wages and salaries; farm self-employment; nonfarm self-employment; interest, dividends, estate, trust, or rent; unemployment compensation, worker's compensation, government employee pensions, or veterans payments; and private pensions, annuities, alimony, royalties, or regular contributions from people not living in the household). In 1979, the questionnaire allowed the recording of up to $99,999 for 23 income sources.[4] In 1985, the limit for recording earnings from longest job increased to $299,999. The final recording limit increase occurred in 1993 when each of the four earned income sources allowed the recording of amounts to $9,999,999.

▶ Measuring Household Income Inequality

Several measures of income inequality are available for analysis. Two important properties an inequality measure should possess are scale invariance and the principle of transfers. An inequality measure is said to be scale invariant if the measure does not change when a constant is added to all income values in the distribution. An inequality measure possesses the principle of transfers if the measure rises (falls) when income is transferred from the poorer household to a richer one (or vice versa).

This paper examines the changes in three measures of household income inequality that possess these qualities: the Gini Coefficient, the Mean Logarithmic Deviation of Income (MLD), and the Atkinson Index.

The Gini index is a measure of income concentration derived from the Lorenz Curve. The Lorenz Curve is obtained by plotting the cumulative percent of units on the X-axis against the cumulative percent of aggregate income accounted for by these units on the Y-axis. A diagonal line from 0 percent to 100 percent would represent the Lorenz Curve if all units had exactly the same income. Lorenz Curves plotted from actual data typically fall below the diagonal. The Gini index is the proportion of the total area below the diagonal that is between the diagonal and the Lorenz Curve. Thus, the Gini index ranges from 0 (perfect equality) to 1 (perfect inequality).

The Atkinson measure of inequality takes a current income distribution and translates it into a social welfare function. The measure is expressed as a ratio of the current welfare function to a welfare function of equally distributed income. The Atkinson measure incorporates a parameter, e, which allows the user to quantify an aversion to inequality. The greater the e value, the more aversion there is to inequality. The value of e ranges between 0 and 1, with 1 indicating maximum inequality aversion with emphasis on the lower end of the income distribution.

The MLD measures the average ratio of the log of the population mean to each observation. The MLD belongs to the Generalized Entropy family. It can be used to measure both within and between group income inequality.

▶ Historical Perspective on Household Income Inequality

Each of the inequality measures displayed in Table 1 was derived from the Census Bureau's internal data file. They show an increase in income inequality between 1967 and 2001, to varying degrees: the Gini index increased 17 percent, the MLD 36 percent, and the Atkinson increased between 28 percent (e=0.75) and 38 percent (e=0.25).[5] Between 1967 and 1980, the Gini index was relatively unchanged. The 1980 MLD and Atkinson measures were at or slightly below their 1967

levels. Each of these measures was at or near its all-time lows by 1974 and was beginning to show signs of increasing. In 1974, the Gini was already above its all-time low set in 1968. By 1982, all of these measures were at or above their 1967 levels and were increasing.

Most of these measures showed growth in income inequality through the late 1980's. By 1989, the Gini and Atkinson measures were measuring income inequality at levels comparable to their all-time highs. The Gini was 8 percent higher than in 1967; the Atkinson with its aversion parameter set to be more sensitive to changes in the upper end of the income distribution (e=0.25) was 13 percent higher; the Atkinson with a midlevel inequality aversion parameter (e=0.5) was 10 percent higher; the Atkinson with an inequality aversion parameter more sensitive to changes in the low end of the income distribution (e=0.75) was 9 percent higher; and the MLD was 7 percent higher.[6]

There appeared to be little change in income inequality between 1989 and 1991.[7] Each of the measures showed growth in inequality between 1991 and 1993, though it is hard to quantify the growth because of survey methodology changes that took place in 1993. In 1994, the CPS ASEC introduced computer-assisted personal interviewing and increased the recording levels for earnings to $1 million as well as increasing the recoding levels for other income sources. Ryscavage (1995) found that as much as one-half of the growth in inequality between 1992 and 1993 may have been the result of these methodological changes. Since 1993, each of the measures has shown periods of fluctuation, culminating in an increase in income inequality by 2001.

▶ Income Topcoding and Inequality Measurement

This section will examine the impact that income recording limits had on the measurement of income inequality. Discussion will focus on the changes to the CPS ASEC questionnaire in 1970, 1979, 1985, and 1993 and also the topcoding limits place on the public-use file.

As discussed earlier, the CPS ASEC has undergone several changes with regard to changing income questions and income recording and processing limits. In 1970, income-recording limits increased to $99,999. This change affected 12,505 people in 12,101 households (33 percent). Ignoring the processing change, each of the income inequality measures showed a slight increase between 1969 and 1970. However, had income recording and processing limits remained at their 1969 and earlier levels, each of the 1970 inequality measures would have been considerably lower (see Table 2). The Gini index would have been 15 percent lower, the MLD 19 percent lower, and the Atkinson between 21 percent and 28 percent lower (28 percent when e=0.25, 25 percent when e=0.5, and 21 percent when e=0.75).

The next change occurred in 1979, affecting 82 people in 81 households (0.1 percent). It had virtually no effect on measured income inequality.

The 1985 change affected 385 people in 380 households (0.6 percent). Between 1984 and 1985, ignoring the processing change, each of the income inequality measures showed a slight increase. However, had income limits remained at their 1984 levels, none of the income inequality measures would have shown any change between 1984 and 1985.

The most dramatic increase in income inequality occurred between 1992 and 1993. Only part of the increase, however, can be attributed to income limits (see Ryscavage). Increased income limits affected 170 people in 167 households (0.3 percent) and caused increases in each of the income inequality measures. The Gini increased 2 percent, the MLD increased 4 percent, and the Atkinson increased between 4 percent and 8 percent (8 percent when e=0.25, 6 percent when e=0.5, and 4 percent when e=0.75).

Public access to microdata requires the Census Bureau to limit some information to ensure the privacy and confidentiality of respondents. Topcoding income is one of the privacy measures used. For some years, the public-use topcodes and internal processing limits on the CPS ASEC were the same. Table 2 shows measures of income inequality derived from the CPS ASEC public-use data along with measures derived from internal Census Bureau data (Old/ New Processing Limits) for selected years. Public-use data show that, as with internal data, all income inequality measures have increased over the 1967-2001 period, but each of the public-use derived

measures showed more growth than the internal measures. The public-use Gini grew by 19 percent, compared to 17 percent using internal data; the MLD grew by 40 percent, compared to 36 percent; and the Atkinson grew by between 34 percent and 45 percent, compared to between 28 percent and 38 percent for internal data.[8] The larger growth in income inequality using public-use data is the result of: 1) topcoded income in 1967 which reduced measured income inequality and 2) increased high income through the plugging of mean topcoded values beginning in 1996 (1997 CPS ASEC).

▶ Income Inequality Without Reporting Limits

In actuality, there are two restrictions that limit the reporting of high-income values on the CPS: a data collection limit and a processing limit. The questionnaire limits the reporting of income by restricting the number of digits available for recording an amount during data collection. This limit was set by physical restriction of a paper questionnaire. In 1993, this physical restriction virtually disappeared with the advent of computer-assisted data collection. A data processing limit is applied to minimize the possible impact of recording (keying) errors, help maintain respondent confidentiality, and prevent volatility and distortion of annual statistics. It also compromises the survey's coverage of the income distribution and may understate income inequality. Prior to 1993, income recording and processing limits were the same.

Table 3 shows the current questionnaire and processing limits and the number of people who exceeded the processing limits for selected income sources on the 2000 CPS. There were no cases that reported income in excess of the data capture limit.

Allowing unrestricted income reporting increased aggregate household income by about 0.1 percent and affected income inequality measures to varying degrees. The Gini index was the measure least affected by allowing unrestricted income reporting, showing an increase of 1.1 percent (see Table 4). The MLD was slightly more affected, increasing 1.9 percent. Unrestricted income reporting had the most effect on the three Atkinson measures. As would be expected, the measure with the highest sensitivity to changes in the upper end of the income distribution (e=0.25) increased 5.4 percent, while the measure most sensitive to changes in the lower end of the distribution (e=0.75) increased only 2.2 percent.

▶ High-Income Sample Turnover and Its Impact on Income Inequality Measures

One major concern with allowing the unrestricted reporting for high-income cases is sample turnover and the impact the loss or gain of very high-income sample cases could have on interpreting annual changes in income inequality. For example, an examination of high income reporting on the 1999 CPS ASEC (1998 income) and the 2000 CPS ASEC (1999 income) showed that sample turnover accounted for the loss of four high income households, with one of the those households having a maximum $9,999,999 in earnings reported. Between 1998 and 1999, there was virtually no change in any of the income inequality measures.[9]

▶ Comparison of CPS Income Data With Administrative Sources

Any income inequality measure is only as good as the data used to construct it. One way to gauge the quality of the CPS ASEC income is by comparing it to independent sources. This section uses National Income and Product Account (NIPA) summaries and matched Internal Revenue Service individual tax return information as benchmarks for evaluating CPS ASEC income data (see Roemer for a discussion of how to reconcile the NIPA and CPS ASEC income definitions).

The most recent comparison of CPS and NIPA data uses 1996 income data. Table 5 shows that CPS aggregate income in 1996 was at 93 percent of NIPA benchmarks. The quality of CPS data varied widely from 53 percent for self-employment income to 102 percent for wages and salaries. Since 1990, most of the income groupings (earnings, property, and transfers) have shown a general trend toward slightly improved CPS data quality. Pensions, however, registered a 12-percentage point decline.

Earnings are a major component of income. In 2001, over $5.3 trillion (82 percent) of the total $6.4 trillion collected in the CPS ASEC were from earnings; 77 per-

cent were from wages and salaries alone. A recent study matched 28,213 1996 IRS tax units to fully reported 1997 CPS ASEC records. Table 6 shows how well the CPS ASEC-reported wage data corresponded with tax data by tax wage interval. Approximately equal proportions of CPS wage earners reported amounts above tax amounts as did earners reporting amounts below. The total reporting discrepancy amounted to $210 million, or 23 percent of the $913 million reported by these CPS households. Roemer's work with these matched data (2001) found that the CPS ASEC netted excess aggregate wages in all of the income intervals except the highest, $150,000 and over.

▶ Conclusions

Each of the inequality measures examined using internal CPS ASEC data painted a similar picture of changing household income inequality over the 1967-2001 period. Overall, income inequality rose between 17 percent and 38 percent, depending on the measure.

The methodological changes that occurred in the 1971 and the 1994 CPS ASECs had a noticeable impact on inequality measurement. With nearly one-third of the households on the 1971 CPS ASEC having restricted incomes due to income reporting limits, income inequality may have been understated by between 15 percent to 28 percent in prior years. A much smaller percentage of households (0.3 percent) were affected by the introduction of higher income recording limits in the 1994 CPS ASEC, resulting in a possible understatement of income inequality of between 2 percent and 8 percent.

The CPS ASEC has been criticized for its inability to accurately measure income inequality because it fails to collect high-income values. A review of income inequality measures using unrestricted income values reported on the March 2000 CPS showed that processing limits only modestly affected estimates of income inequality. Removing the processing limits would increase measured income inequality by between 1 percent and 5 percent.

Restricted income information on the public-use version of the CPS ASEC causes a further reduction of measured income inequality in years prior to 1996. The plugging of mean values for topcoded respondents beginning with the 1997 public-use CPS ASEC brought public measurement of income inequality more in line with internal measurement. The net result, however, is an overstatement of income inequality growth over the 1967-2001 period.

A review of independent benchmarks showed that the quality of the CPS ASEC income data seemed reasonable. Overall, aggregate CPS ASEC income was at 93 percent of NIPA totals. A comparison to tax returns showed that the CPS ASEC had more reported wages than on tax returns in all but the highest income categories

▶ References

Technical Paper 17 (1967), "Trends in the Income of Families and Persons in the United States: 1947-1964," U.S. Government Printing Office, Washington, DC.

Current Population Report, Series P60-204 (2000), "The changing Shape of the Nation's Income Distribution: 1947-1998." Government Printing Office, Washington, DC.

Jones, Arthur Jr, "Measuring Household Income Inequality, 1967-1997," presented at the Joint Statistical Meetings for the American Statistical Association, Baltimore, Maryland, August 1999.

Roemer, Marc (2000), "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates 1990-1996," U.S. Census Bureau.

Roemer, Marc, "An Evaluation of High Income Reporting on the March Current Population Survey (CPS)," U. S. Census Bureau, Internal Memorandum, March 2001.

Ryscavage, Paul, "A Surge in Growing Income Inequality?," *Monthly Labor Review,* August 1995, pp. 51-61.

Welniak, Edward, "Effects of the March Current Population Survey's New Processing System on

Estimates of Income and Poverty," 1990 ASA proceedings.

▶ Footnotes

[1] The CPS sample size increased in 2001 from approximately 50,000 households to 78,000 to improve estimates for the State Children's Health Insurance Program.

[2] The ASEC was formerly known as the CPS March Income Supplement.

[3] People 14 years old and over prior to 1989.

[4] The income limits were $9,999 for Social Security; $5,999 for Supplemental Security Income; $19,999 for public assistance; and $29,999 for veterans' payments.

[5] The growth rates in income inequality between 1967 and 2001 for the MLD and Atkinson (e=.25 and e=.5) were not statistically different from one another.

[6] The growth rates from 1967 to 1989 for the Gini and Atkinson (e=.25) were statistically different from one another, as were the growth rates for the MLD and Atkinson (e=.25).

[7] Between 1989 and 1991, the Atkinson Measure with e=.25 declined 2.5 percent.

[8] There was no difference between the MLD growth rate and the growth rates for the Atkinson e=0.25 and e=0.5.

[9] The MLD showed a significant decline of 2.7 percent.

Table 1. Measures of Household Income Inequality: 1967 to 2001

Year	Inequality Measures				
			Atkinson		
	Gini Index	MLD	e=0.25	e=0.50	e=0.75
1967	0.399	0.380	0.071	0.143	0.220
1968	0.388	0.356	0.067	0.135	0.208
1969	0.391	0.357	0.067	0.135	0.209
1970	0.394	0.370	0.068	0.138	0.214
1971	0.396	0.370	0.068	0.138	0.214
1972	0.401	0.370	0.070	0.140	0.216
1973	0.397	0.355	0.068	0.136	0.210
1974	0.395	0.352	0.067	0.134	0.207
1975	0.397	0.361	0.067	0.136	0.210
1976	0.398	0.361	0.068	0.137	0.211
1977	0.402	0.364	0.069	0.139	0.213
1978	0.402	0.363	0.069	0.139	0.213
1979	0.404	0.369	0.070	0.141	0.216
1980	0.403	0.375	0.069	0.140	0.216
1981	0.406	0.387	0.070	0.141	0.220
1982	0.412	0.401	0.072	0.146	0.226
1983	0.414	0.397	0.072	0.147	0.226
1984	0.415	0.391	0.073	0.147	0.225
1985	0.419	0.403	0.075	0.151	0.231
1986	0.425	0.416	0.077	0.155	0.237
1987	0.426	0.414	0.077	0.155	0.238
1988	0.427	0.401	0.078	0.155	0.236
1989	0.431	0.406	0.080	0.158	0.239
1990	0.428	0.402	0.078	0.156	0.236
1991	0.428	0.411	0.078	0.156	0.237
1992	0.434	0.416	0.080	0.160	0.242
1993	0.454	0.467	0.092	0.178	0.266
1994	0.456	0.471	0.092	0.180	0.268
1995	0.450	0.452	0.090	0.175	0.261
1996	0.455	0.464	0.093	0.179	0.266
1997	0.459	0.484	0.094	0.183	0.272
1998	0.456	0.488	0.093	0.181	0.271
1999	0.457	0.475	0.092	0.180	0.268
2000	0.462	0.490	0.096	0.185	0.275
2001	0.466	0.515	0.098	0.189	0.282

Source: U.S. Census Bureau, Current Population Survey, selected ASEC Supplements.

Table 2. Impact of Income Limits on Household Inequality Measures

Year	Gini Index			MLD		
	Public Use	Old Processing Limit	New Processing Limit	Public Use	Old Processing Limit	New Processing Limit
1967	0.390	NA	0.399	0.363	NA	0.380
1970	0.394	0.334	0.394	0.363	0.299	0.370
1979	0.394	0.404	0.404	0.342	0.369	0.369
1985	0.414	0.414	0.419	0.380	0.396	0.403
1993	0.425	0.444	0.454	0.424	0.451	0.467
2001	0.464	NA	0.466	0.510	NA	0.515

Year	Atkinson								
	e=0.25			e=0.50			e=0.75		
	Public Use	Old Processing Limit	New Processing Limit	Public Use	Old Processing Limit	New Processing Limit	Public Use	Old Processing Limit	New Processing Limit
1967	0.065	NA	0.071	0.133	NA	0.143	0.208	NA	0.220
1970	0.065	0.049	0.068	0.133	0.104	0.138	0.208	0.169	0.214
1979	0.065	0.070	0.070	0.133	0.140	0.141	0.206	0.216	0.216
1985	0.072	0.072	0.075	0.146	0.147	0.151	0.225	0.226	0.231
1993	0.076	0.085	0.092	0.154	0.168	0.178	0.238	0.255	0.266
2001	0.094	NA	0.098	0.184	NA	0.189	0.278	NA	0.282

Source: U.S. Census Bureau, Current Population Survey, selected ASEC Supplements.

Table 3. High Income Reporting, by Income Source: 1999
(Limits in dollars)

Income Source	Questionnaire Limit	Processing Limit	Number of cases with reported values exceeding the processing limit	Number of cases with imputed values exceeding the processing limit
Earnings	9,999,999	1,099,999	26	7
Interest	9,999,999	99,999	19	54
Dividends	9,999,999	100,000	23	21
Rent	9,999,999	99,999	26	14
Retirement	999,999	99,999	26	NA

Source: Roemer 2001.
NA not available.

Table 4. Household Income Inequalty Measures by Presence of Income Reporting Limits: 1999

Inequality Measure	With processing limits	Without processing limits	Percent change
Gini index	0.457	0.462	1.1
MLD	0.475	0.484	1.9
Atkinson:			
e=0.25	0.092	0.097	5.4
e=0.50	0.180	0.186	3.3
e=0.75	0.268	0.274	2.2

Source: Roemer (2001)

Table 5. March CPS as a Percent of National Income and Product Account Benchmarks: 1990 to 1996

Income Source	1990	1991	1992	1993	1994	1995	1996
Wages and Salary	95.9	96.4	95.6	99.7	101.9	101.4	101.9
Self-Employment	68.5	65.3	58.6	58.9	54.8	48.5	52.6
Earnings	**93.0**	**93.0**	**91.3**	**94.8**	**96.4**	**95.1**	**96.1**
Interest	67.1	68.3	67.6	79.7	72.3	83.9	83.8
Dividends	40.9	45.7	49.2	54.3	54.6	62.6	59.4
Rent and Royalties	85.0	74.1	69.8	65.2	64.8	58.7	58.6
Property	**62.8**	**63.3**	**63.2**	**69.8**	**65.7**	**72.9**	**70.9**
Social Security and Railroad Retirement	90.6	88.6	87.1	87.8	92.3	92.0	91.7
Supplemental Security Income	78.9	84.6	75.5	84.2	78.0	77.1	84.2
Family Assistance	74.4	74.4	72.2	76.4	73.1	70.5	67.7
Other Cash Welfare	85.6	77.5	81.6	101.3	105.2	95.8	80.5
Unemployment Compensation	79.9	82.5	72.8	77.6	90.0	91.3	81.6
Worker's Compensation	89.5	89.1	82.5	77.0	77.7	69.3	62.7
Veterans' Payments	73.9	82.9	77.7	85.5	84.7	94.9	89.6
Transfers	**87.6**	**86.8**	**83.6**	**85.6**	**89.5**	**89.2**	**88.3**
Pensions	**88.9**	**85.5**	**83.1**	**83.6**	**83.1**	**78.2**	**76.6**
Total	**89.3**	**89.4**	**88.0**	**91.7**	**92.9**	**92.2**	**92.6**

Source: Roemer 2000.

Table 6. Comparison of Fully Reported CPS Wages and Matched IRS Tax Return Wages: 1996 (Includes both filers if joint return)

Tax Return Wage Range	Number of Tax Units	CPS below Tax Return (%)	CPS above Tax Return (%)	Total Discrepancy (thousands of dollars)	Share of Discrepancy (%)
Total	28,213	49.7	50.3	210,055	100.0
Zero	476	0.0	100.0	12,952	6.2
1 to 2,499	2,160	62.0	38.0	4,524	2.2
2,500 to 4,999	1,991	58.0	42.0	6,799	3.2
5,000 to 9,999	3,030	54.8	45.2	13,411	6.4
10,000 to 14,999	2,807	52.6	47.4	14,341	6.8
15,000 to 19,999	2,488	52.1	47.9	10,938	5.2
20,000 to 29,999	4,237	47.6	52.4	22,623	10.8
30,000 to 39,999	3,112	47.6	52.4	21,032	10.0
40,000 to 49,999	2,394	43.9	56.1	16,834	8.0
50,000 to 59,999	1,733	44.0	56.0	14,603	7.0
60,000 to 74,999	1,730	44.9	55.1	17,176	8.2
75,000 to 99,999	1,189	43.1	56.9	15,258	7.3
100,000 to 149,999	589	49.6	50.4	13,795	6.6
150,000 and over	277	69.7	30.3	25,769	12.3

Source: Roemer 2001.

An Analysis of the Distribution of Individual Income and Taxes, 1979-2001

Michael Strudler and Tom Petska, Internal Revenue Service, and Ryan Petska, Ernst and Young LLP

Different approaches have been used to measure the distribution of individual income over time. Survey data have been compiled with comprehensive enumeration, but underreporting of incomes, inadequate coverage at the highest income levels, and omission of a key income type jeopardize the validity of results. Administrative records, such as income tax returns, may be less susceptible to underreporting of income but exclude certain nontaxable income types and can be inconsistent in periods when the tax law has been changed. Record linkage studies have capitalized on the advantages of both approaches, but are costly and severely restricted by the laws governing interagency data sharing.

This paper is the fifth in a series examining trends in the distribution of individual incomes and tax burdens based on a consistent and comprehensive measure of income derived from individual income tax returns.[1,2,3,4] In the previous papers, we demonstrated that the shares of income accounted for by the highest income-size classes clearly have increased over time, and we also demonstrated the superiority of our comprehensive and consistent income measure, the 1979 Retrospective Income Concept, particularly in periods of tax reform. In this paper, we continue the analysis of individual income and tax distributions, adding for 3 years (1979, 1989, and 1999) Social Security and Medicare taxes to this analysis. The paper has three sections. In the first section, we briefly summarize this measure of individual income derived as a "retrospective concept" from individual income tax returns. In the second section, we present the results of our analysis of time series data. We conclude with an examination of Gini coefficients computed from these data.

▶ Derivation of the Retrospective Income Concept

The tax laws of the 1980's and 1990's made significant changes to both the tax rates and definitions of taxable income. The tax reforms of 1981 and 1986 significantly lowered individual income tax rates, and the latter also substantially broadened the income tax base. The tax law changes effective for 1991 and 1993 initiated rising individual income tax rates and further modifications to the definition of taxable income.[1,2,3,4] Law changes effective for 1997 substantially lowered the maximum tax rate on capital gains. The newest law changes have lowered marginal rates starting with 2001 and will again lower the maximum tax rate on long-term capital gains, as well as decrease the maximum rates for most dividends. With all of these changes, the questions that arise are what has happened to the distribution of individual income, the shares of taxes paid, and average taxes by the various income-size classes?

In order to analyze changes in income and taxes over time, consistent definitions of income and taxes must be used. However, the Internal Revenue Code has been substantially changed in the last 23 years—both the concept of taxable income and the tax rate schedules have been significantly altered. The most commonly used income concept available from Federal income tax returns, Adjusted Gross Income (AGI), has changed over time making it difficult to use AGI for inter-temporal comparisons of income. For this reason, an income definition that would be both comprehensive and consistent over time was developed.[5, 6, 7, 8] The 1979 Retrospective Income Concept was designed to include the same income and deduction items from items available on Federal individual income tax returns. Tax Years 1979 through 1986 were used as base years to identify the income and deduction items, and the concept was subsequently applied to later years, including the same components common to all years.

The calculation of the 1979 Retrospective Income Concept includes several items partially excluded from AGI for the base years, the largest of which was capital gains.[1,2,3,4] The full amounts of all capital gains, as well as all dividends and unemployment compensation, were included in the income calculation. Total pensions, annuities, IRA distributions, and rollovers were added, includ-

ing nontaxable portions that were excluded from AGI. Social Security benefits were omitted because they were not reported on tax returns until 1984. Also, any depreciation in excess of straight-line depreciation, which was subtracted in computing AGI, was added back. For this study, retrospective income was computed for all individual income tax returns in the annual Statistics of Income (SOI) sample files for the period 1979 through 2001. Loss returns were excluded, and the tax returns were tabulated into income-size classes based on the size of retrospective income and ranked from highest to lowest. Percentile thresholds were estimated or interpolated for income-size classes ranging from the top 0.1 percent to the bottom 20 percent.[9,10,11] For each size class, the number of returns and the amounts of retrospective income and taxes paid were compiled. From these data, income and tax shares and average taxes were computed for each size class for all years.

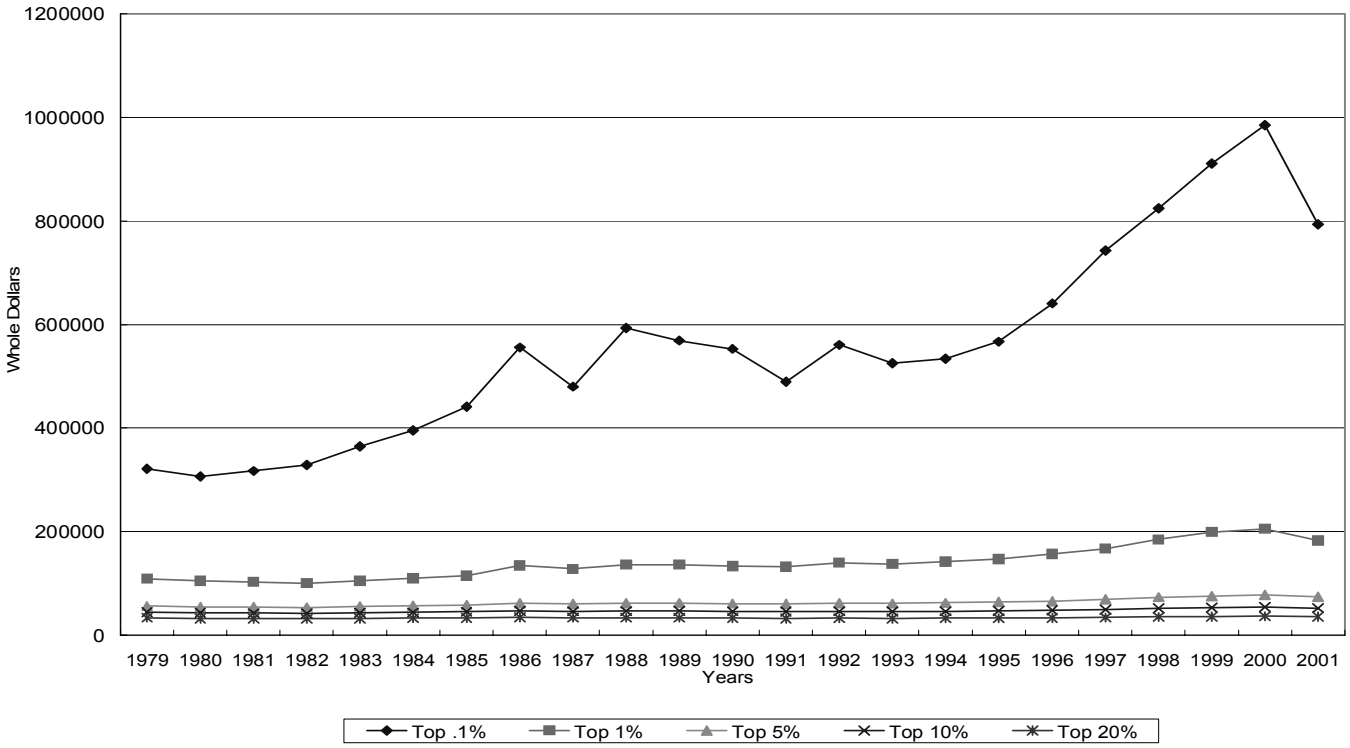
▶ The Distribution of Income and Taxes

With this data base, we sought to answer the following questions—have the distribution of individual incomes (i.e., income shares), the distribution of taxes (i.e., tax shares), and the average effective tax rates (i.e., tax burdens) changed over time? As a first look at the data, we examined the income thresholds of the bottom (or entry level) of each income-size class, and a clear pattern emerged. While all of the income thresholds have increased over time, the largest increases in absolute terms, and on a percentage basis, were with the highest income-size classes.

For example, while $233,539 were needed to enter the top 0.1 percent for 1979, $1,405,770 were needed for entry into this class for 2001. This represents a more than 500-percent increase. Also, while $79,679 of retrospective income were needed to enter the top 1-percent size class for 1979, $323,861 were needed for entry into this size class for 2001, an increase of 306 percent. For the top 20 percent, the threshold increased by 159 percent, and, for the bottom 20 percent, the increase was only 124 percent. Since much of these increases are attributable to inflation, we computed constant dollar thresholds, using the Consumer Price Index.[12]

What is most striking about these data are the changes between 1979 and 2001 for the various income-size percentile thresholds (see Figure A). For example, the threshold for the top 0.1 percent grew (using a 1982-

Figure A. Constant Dollar Income Thresholds, 1979-2001 (1982-84=100)



1984 base) from $321,679 for 1979 to $793,772 for 2001, an increase of 147 percent. Similarly, the threshold for the taxpayers in the 1-percent group rose from $109,751 for 1979 to $182,869 for 2001, an increase of over 66 percent. However, the thresholds for each lower percentile class show smaller increases in the period; the top 20-percentile threshold increased only 6.1 percent, and the 40-percent and all lower thresholds all declined.

Income Shares

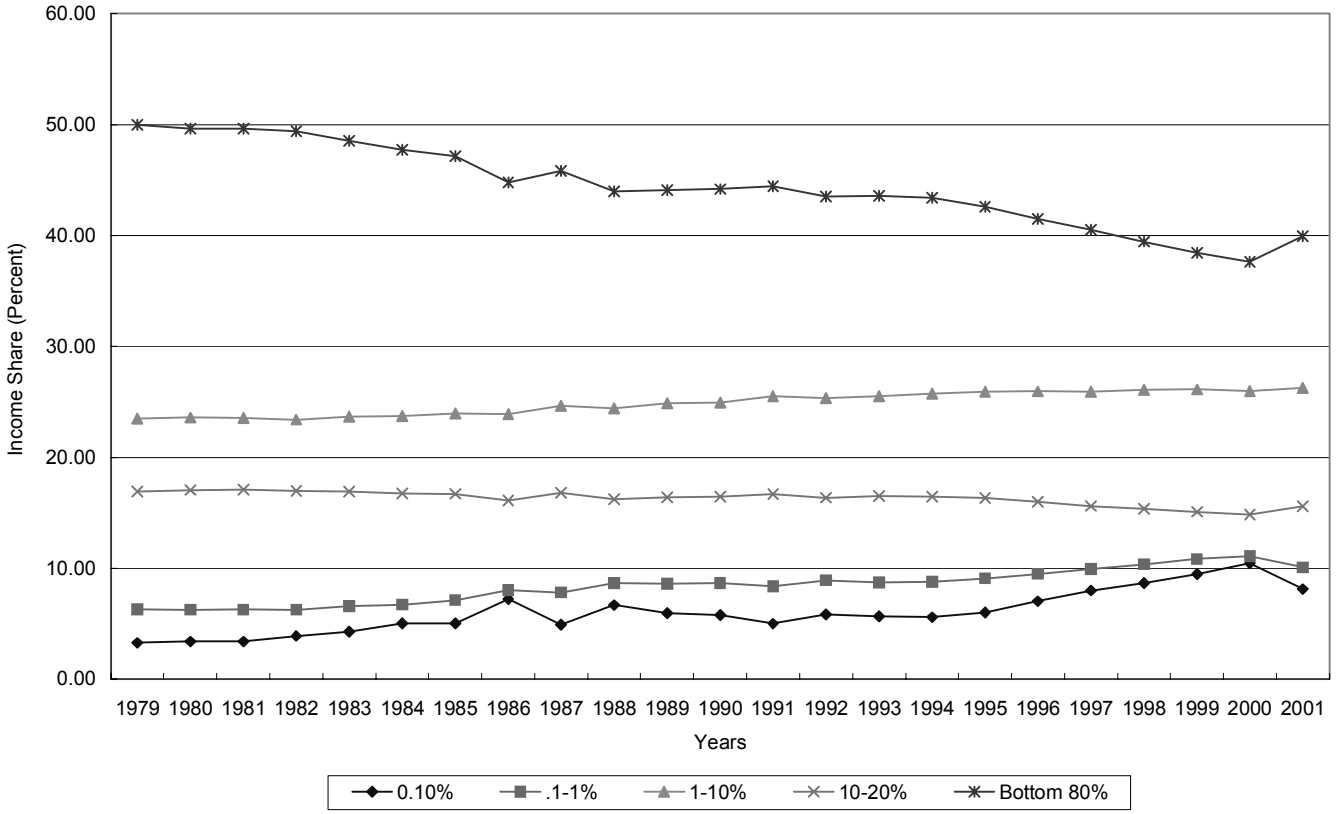
The share of income accounted for by the top 1 percent of the income distribution has climbed steadily from a low of 9.58 percent (3.28 for the top 0.1 percent) for 1979 to 18.22 percent (8.13 for the top 0.1 percent) for 2001. While this increase is quite steady, there were some significantly large jumps, particularly for 1986, due to a surge in capital gain realizations after the passage, but before implementation, of the Tax Reform Act of 1986 (TRA). The top 1-percent share also increased for 1996 through 2000, when sales of capital assets also grew considerably each year. Notable declines in the top 1-percent share occurred in the recession years of 1981, 1990-1991, and 2001.

This pattern of an increasing share of total income is mirrored in the 1-to-5 percent class but to a considerably lesser degree. For this group, the income share increased from 12.60 percent to 15.12 percent in this period. The 5-to-10 percent class's share of income held fairly steady over this period, going from 10.89 percent for 1979 to 11.12 percent for 2001. The shares of the lower percentile-size classes, from the 10-to-20-percent classes to the four lowest quintiles, show declines in shares of total income over the 23-year period (see Figure B).

Tax Shares—Income Tax

The share of income taxes accounted for by the top 1 percent also climbed steadily in this period, from initially at 19.75 percent (7.38 for the top 0.1 percent) for

Figure B. Income Shares by Income Percentile Size-Classes, 1979-2001



1979, then declined to a low of 17.42 percent (6.28 for the top 0.1 percent) for 1981, before rising to 36.30 percent (18.70 for the top 0.1 percent) for 2000 (Figure C).

The corresponding percentages for 2000 for the 1-percent and 0.1-percent groups are 37.68 percent and 19.44 percent, respectively, accounting for the 2000 tax rebate, which is discussed below. For the recession year of 2001 with its large decline in net gains from the sale of capital assets, these shares declined to 32.88 percent for the top 1-percent and 15.78 for the top 0.1-percent group. As with incomes, there were some years with unusually large increases though a common feature for these years was double-digit growth in net capital gains.[7,8]

The 1-to-5 percent size class exhibited relatively modest change in its share of taxes, increasing from 17.53 percent to 19.62 percent in the period. The 5-to-10 percent class, and all lower income-size classes, had declining shares of total tax.

Average Tax Rates—Income Tax

What is most striking about these data is that the levels of the average tax burdens increase with income size in most years (the only exceptions being 1986 for just the two highest groups). The progressive nature of the individual income tax system is clearly demonstrated.

Despite the fact that the overall average tax rate remained virtually the same for 1979 and 2001, the average rate for all but the very lowest size class actually declined.[13] While this at first appears to be inconsistent, it is clear how this did in fact occur—over time, an increasing proportion of income has shifted to the upper levels of the distribution where it is taxed at higher rates (see Figure D).

As for the tax share data, accounting for the 2000 rebate had a significant effect, lowering the overall av-

Figure C. Tax Shares by Income Percentile Size-Classes, 1979-2001

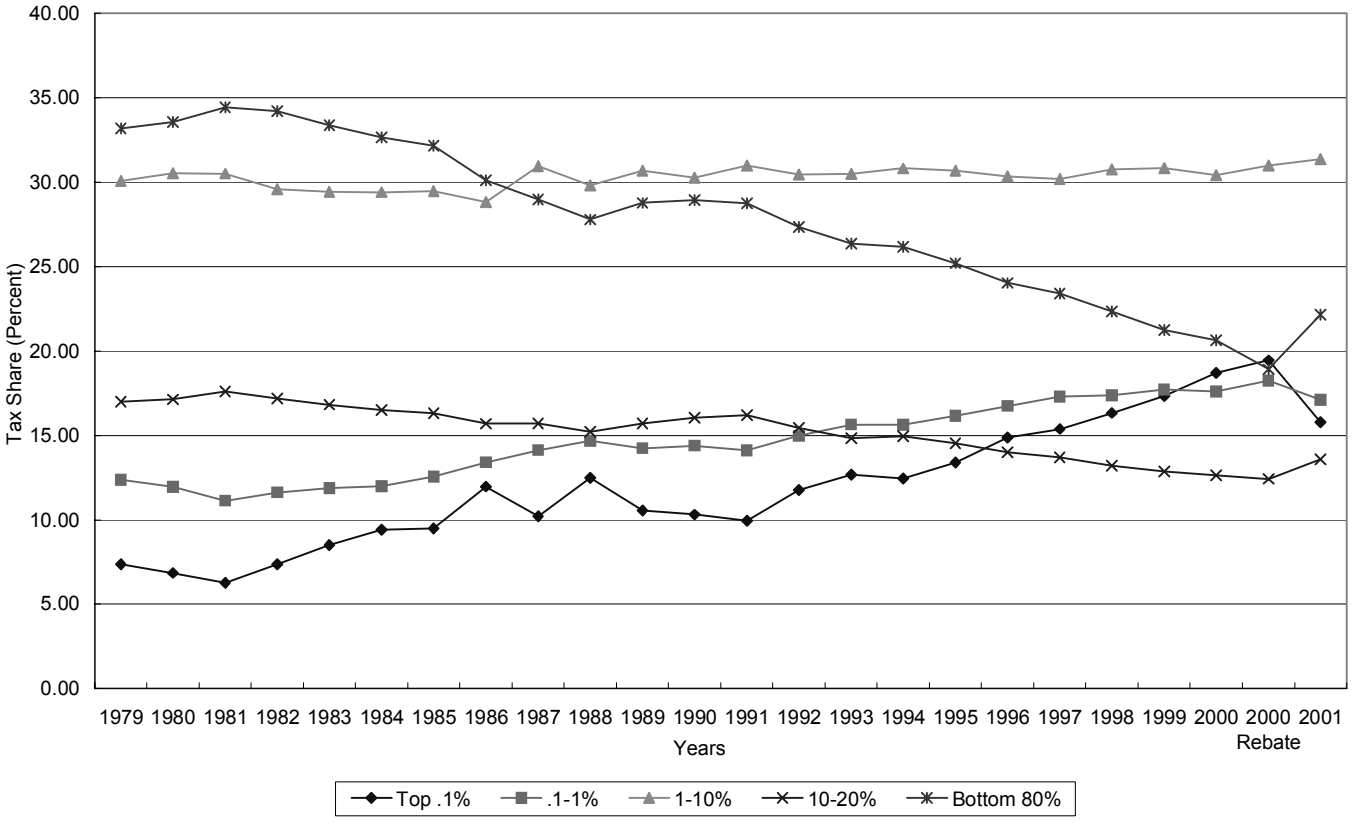
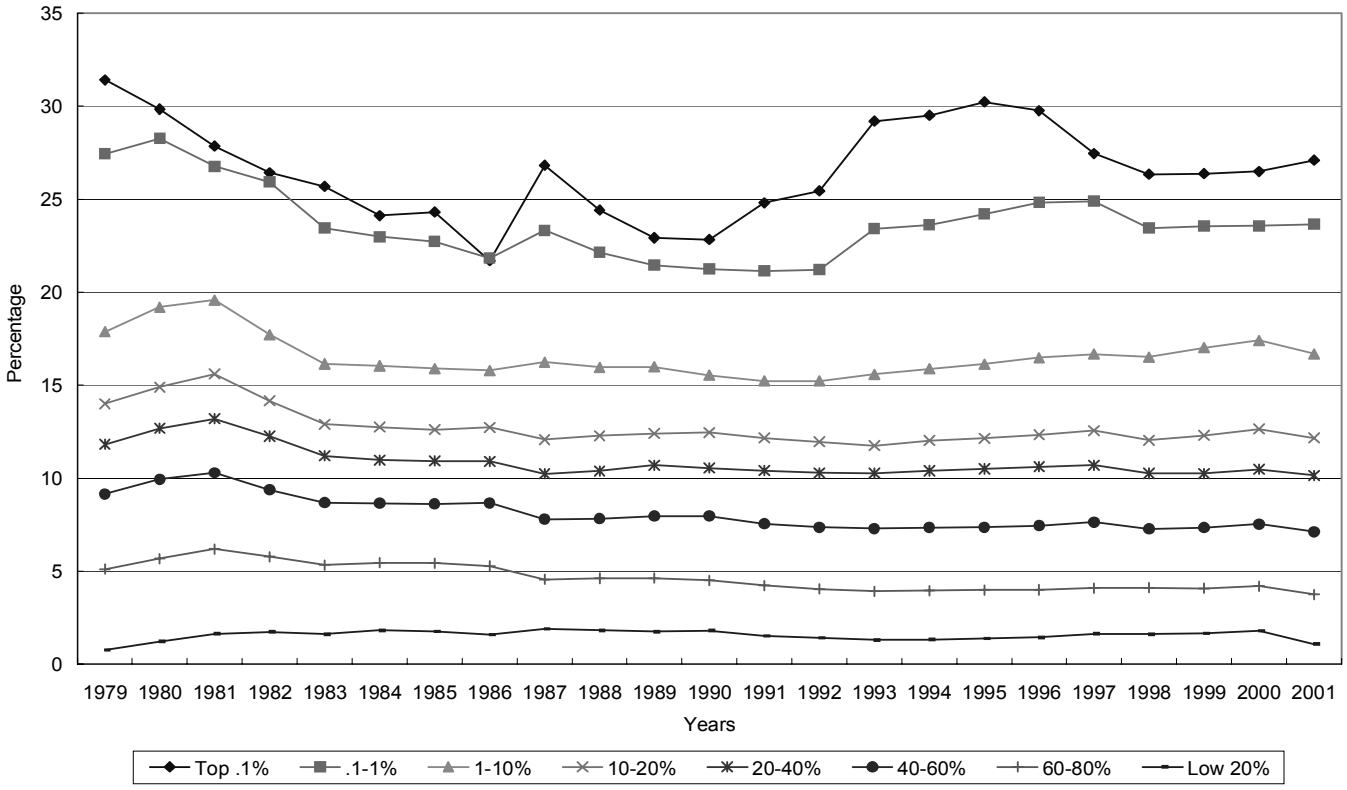


Figure D. Average Tax Rates by Size-Classes, 1979-2001



erage tax rate from 14.85 percent to 14.28 percent. A combination of lower marginal tax rates, larger child tax credits, and recession caused this rate to decrease to 13.96 percent for 2001.

In examining the average tax data by income size, four distinct periods emerge. First, the average tax rates were generally climbing up to the implementation of the Economic Recovery Tax Act (ERTA) effective for 1982. This was an inflationary period and prior to indexing of personal exemptions, the standard deduction, and tax brackets, which caused many taxpayers to face higher tax rates. (Indexing became a permanent part of the tax law for Tax Year 1985.[6]) Also, this period marked the recovery from the recession in the early 1980's.

Similarly, average taxes also climbed in the period after 1992, the period affected by the Omnibus Budget and Reconciliation Act (OBRA). This was not surprising for the highest income-size classes, ones affected by the OBRA-initiated 39.6-percent top marginal tax rate, but the average tax rate increases are also evident in the smaller income-size classes for most years in the 1993 to 1996 period as well.

For the majority of intervening years (i.e., 1982 through 1992), average tax rates generally declined by small amounts for most income-size classes, although the period surrounding the implementation of the 1986 Tax Reform Act (TRA) gave rise to small increases in some classes. Despite the substantial base broadening and rate lowering initiated by TRA, for most income-size classes, the changes to average rates were fairly small. However, it should be kept in mind that individuals can and do move between income-size classes.

The rates for the top 0.1 percent clearly show the effects of the 1986 capital gain realizations, in anticipation of the ending of the 60-percent long-term gain exclusion, which began in 1987. The average tax rate for this income-size class dropped for 1986, but it rose sharply for 1987, before dropping again for each of the next 3 years.

To assess what happened, it is important to look at the underlying data. The substantial increase in capital gain realizations for 1986 swelled the aggregate income and tax amounts for upper income classes and also raised the income thresholds of these top classes. However, since much of the increase in income for these size classes was from net long-term capital gains, which had a maximum effective tax rate of 20 percent, it is not surprising that the average tax rate for these top size classes declined.

Last, are those years affected by the Taxpayer Relief Act of 1997 (1997 through 2001), where the top rate on long-term capital gains was reduced significantly from 28 percent to 20 percent. For 1997, the first year under this law, when the lower rates were only partially in effect, the average tax rate fell for the top 0.1-percent group of taxpayers but increased for all other groups. However, for 1998, the first full year under lower capital gain rates, all groups up to and including the 40-to-60 percent class had reduced average tax rates (while the lowest two quintiles had virtually the same average tax rates). For all groups (except for the 20-40 and the 60-to-80 percent groups in 1999), the average rates returned to increasing for both 1999 and 2000.

The Economic Growth and Tax Relief Reconciliation Act of 2001 (EGTRRA) further reduced marginal tax rates over several years. One of these reductions was an introduction of a 10-percent bracket on the first $6,000 ($12,000 if married filing a joint return) of taxable income. In an attempt to fuel a recovery from recession, this reduction was introduced retroactively in the form of a rebate based on Tax Year 2000 filings. Therefore, we simulated the rebate on the Tax Year 2000 Individual File to see its effects on average tax rates. When the rebate is taken into account, the average rates for 2000 decreased for all groups, except for the top 0.1 and the 1-to-5 percent, reversing the pre-rebate increases. Tax Year 2001 was a mixture of increases and decreases in average tax rates by income group. Most groups paid higher average taxes; however, the 1-to-5 and 5-to-10 percent groups paid lower average taxes along with the bottom 20-percent group.

Tax Shares—Income Plus Social Security Tax

For individual taxpayers, Social Security taxes compose a fairly large portion (about 37 percent for 1999) of the Federal tax burden.[14] To broaden our analysis, we merged data from W-2's with individual income tax records for the years 1979, 1989, and 1999. Total social security taxes included self-employment taxes and taxes on tips reported on tax returns and two times the social security taxes (representing both the taxpayers' and the employers' shares) reported on

W-2's. The employers' share of this tax was added into retrospective income, as well. To further help our analysis, the U.S. Treasury Department's Office of Tax Analysis (OTA) model was used to simulate the effect of the two new tax laws (EGTRRA) and the Jobs and Growth Tax Relief Reconciliation Act of 2003 (JGTRRA), on the 1999 data.[15]

Even including Social Security taxes, the shares of the higher income groups increased (the top 0.1-percent group's share more than doubled from 5.06 percent for 1979 to 11.05 percent for 1999), while the shares of the lower income groups (each group from the 10-to-20 percent group and lower) declined (see Figure E). However, when we simulated all of the provisions of EGTRRA/JGTRRA on 1999 data, tax shares for the top two groups (the 0.1- and the 0.1-to-1 percent groups) declined from 1999 levels, while all other groups increased. Still, for these two groups and the 1-to-5 per-

Figure E. Tax Shares (Including Social Security Taxes) by Percentile Size-Classes, 1979-2001

Year	Top 0.1%	0.1-1%	1-5%	5-10%	10-20%	Top 20%	20-40%	40-60%	60-80%	Low 20%
1979	5.06	8.97	14.69	11.87	17.70	58.28	22.97	12.42	5.12	1.22
1989	6.29	9.43	15.42	12.51	17.63	61.29	21.94	11.18	4.44	1.15
1999	11.05	12.27	16.84	12.03	15.98	68.17	18.83	9.28	3.09	0.63
1999 JGTRRA	9.52	11.31	17.75	12.50	16.39	67.47	19.22	9.54	3.11	0.65

cent, the tax shares were still higher than 1989 levels. Interestingly, the 1-to-5 percent group is the only group whose share increased from 1989 to 1999 (from 15.42 percent to 16.84 percent) and then increased again (to 17.85 percent) under new tax law provisions. This is most likely due to the effect of the alternative minimum tax (AMT) offsetting lower marginal and capital gain rates for this group of taxpayers.

Average Tax Rates Including Social Security Taxes

Unlike the tax shares data, average taxes, including Social Security taxes, vary considerably over time from average income taxes. Including Social Security taxes for 1979, the overall tax system (like the income tax system) was progressive, with each higher income class paying a higher percentage average tax than the classes preceding them (see Figure F). However, this is not entirely true for any of the other years that we merged income tax with W-2 data. For 1989, the system was progressive up to the 5-to-10 percent income class. Above this level, each successively higher income class paid a lower rate than the ones below them, falling to 23.33 percent for the top 0.1-percent income group. In fact, for 1989, the top 0.1-percent group faced a lower rate than all groups from the 10-to-20 percent income group and higher. The highest rate for that year was paid by those individuals in the 5-to-10 percent income group at 25.09 percent, 1.76 percentage points higher than those in the 0.1-percent group.

In contrast, the 5-to-10 percent group paid an average tax of 22.59 percent in 1979, some 9.33 percentage points lower than those in the 0.1-percent group. A large reason for this increase in rate for the 5-to-10 percent group was the increase in Social Security taxes. For 1979, wage earners and their employers paid a combined rate of 8.1 percent in Social Security taxes on earnings up to $22,900. By 1989, this had increased to 13.02 percent on earned income up to $48,000. For 1999, this had further increased to 15.3 percent on earned income up to $72,600. Furthermore, for 1999, for any earned income above the $72,600 maximum, the employee and employer continued to pay Medicare taxes at a combined rate of 2.9 percent.

Despite this rise in Social Security taxes, 1999 combined average taxes returned to a mostly progressive system. The only exception to this progressive tax structure was the 5-to-10 percent income group, who paid higher average rates (26.18 percent) than the 1-to-5 percent income group (25.97 percent). However, the 0.1-to-1 percent and the 0.1-percent income groups paid the highest average taxes at 26.70 percent and 27.51 percent.

When we simulated the provisions of the two new tax laws (EGTRRA and JGTRRA) on 1999 data (without allowing for the sunset provisions), the overall tax system returns to a system looking more like 1989 than 1999. Under the simulation, average tax rates continue to increase until the 1-to-5 percent income class (who pay the highest average tax at 25.76 percent). From there, average taxes fall to 23.34 percent for the 0.1-to-1 percent income group and decline further to 22.57 percent for the 0.1-percent income group. Both of these groups would pay a lower average tax than individuals in the 10-to-20 percent income class. The highest income group winds up paying an average tax that is less than all of the groups above the 20-to-40 percent class. Under the new laws, the 0.1-percent group would pay average taxes that are 3.19 percentage points less than the 1-to-5 percent income group, 2.91 percentage points less than the 5-to-10 percent income group, and 1.24 percentage points less than the individuals in the 10-to-20 percent group. In fact, under the provisions of

Figure F. Average Tax Rates (Including Social Security Taxes) by Percentile Classes, 1979-2001

Year	Total	< 0.1%	0.1 - 1%	1-5%	5-10%	10-20%	20-40%	40-60%	60-80%	Low 20%
1979	20.71	31.92	29.50	24.14	22.59	21.63	19.89	17.35	12.65	8.72
1989	22.24	23.33	24.22	24.84	25.09	23.90	22.37	19.29	13.93	11.47
1999	23.59	27.51	26.70	25.97	26.18	24.96	23.22	19.70	11.83	7.29
1999 JGTRRA	21.90	22.57	23.34	25.76	25.48	23.81	21.58	18.25	10.94	6.97

EGTRRA/JGTRRA, the individuals in the 0.1-percent group wind up paying less than one percentage point (0.99) more than the 20-to-40 percent income group. In contrast, the highest income group paid average combined taxes of 12.03 percentage points higher than the 20-to-40 percent income group in 1979 and 4.29 percentage points higher than this group under existing 1999 laws.

▶ Analysis of Gini Coefficients

To further analyze the data, we estimated Lorenz curves and computed Gini coefficients for all years. The Lorenz curve is a cumulative aggregation of income from lowest to highest, expressed on a percentage basis. To construct the Lorenz curves, we reordered the percentile classes from lowest to highest and used the income thresholds as "plotting points" to fit a series of regression equations for each income-size interval in the 23 years, both before- and after-taxes.

Once the Lorenz curves were estimated for all years, Gini coefficients were calculated for all 23 years for before- and after-tax and are presented in Figure G. The Gini coefficient, which is a measure of the degree of inequality, generally increased throughout the 23-year period signifying rising levels of inequality for both the pre- and post-tax distributions. This result was not unexpected since it parallels the rising shares of income accruing to the highest income-size classes. Over this period, the before-tax Gini coefficient value increased from 0.469 for 1979 to 0.588 (25.4 percent) for 2000, while the after-tax Gini value increased from 0.439 to 0.558 for a slightly higher percentage increase (25.5 percent). The recession in 2001 actually decreased the levels of inequality to 0.564 (pre-tax) and 0.534 (after-tax).

So, what has been the effect of the Federal tax system on the size and change over time of the Gini coefficient values? One way to answer this question is to compare the before- and after-tax Gini values.[16] Looking at this comparison, two conclusions are clear. First, Federal income taxation decreases the Gini coefficients for all years. This is not surprising in that the tax rate structure is progressive, with average rates rising with higher incomes—so, after-tax income is more evenly distributed than before-tax income. A second question is whether the relationship between the before-tax and after-tax Gini coefficient values has changed over time. From G, the after-tax series closely parallels the before-tax series, with reductions in the value of the Gini coefficient ranging from 0.024 to 0.032. The largest differences, which denote the largest redistributive effect of the Federal tax system, have generally been in the periods of relatively high marginal tax rates, particularly 1979-81 and for 1993 and later years. In fact, simulating the tax rebate for Tax Year 2000 results in the largest difference (0.032) over all the years. If this were the only change in marginal rates of the new tax law (EGTRRA), the results would be to increase the redistributive effects of Federal taxes. However, for Tax Year 2001 and beyond, the marginal rates of higher income classes will also be reduced over time until the highest rate will be reduced from its current value of 39.6 percent to 35 percent for 2003. The effects of the new tax laws (EGTRRA / JGTRRA) can be seen in Figure H. This

Figure G. Gini Coefficients for Retrospective Income, Before and After Taxes, 1979–2001

Year	Gini Before Tax	Gini After Tax	Difference	Percent Difference
1979	0.469	0.439	0.030	6.3
1980	0.471	0.441	0.031	6.5
1981	0.471	0.442	0.029	6.2
1982	0.474	0.447	0.027	5.7
1983	0.482	0.458	0.025	5.1
1984	0.490	0.466	0.024	4.9
1985	0.496	0.471	0.024	4.9
1986	0.520	0.496	0.024	4.6
1987	0.511	0.485	0.026	5.1
1988	0.530	0.505	0.026	4.8
1989	0.528	0.504	0.024	4.6
1990	0.527	0.503	0.024	4.5
1991	0.523	0.499	0.024	4.6
1992	0.532	0.507	0.025	4.7
1993	0.531	0.503	0.028	5.2
1994	0.532	0.503	0.028	5.3
1995	0.540	0.510	0.029	5.4
1996	0.551	0.521	0.030	5.5
1997	0.560	0.530	0.030	5.4
1998	0.570	0.541	0.029	5.1
1999	0.580	0.550	0.030	5.2
2000	0.588	0.558	0.031	5.2
2000 Rebate	0.588	0.557	0.032	5.4
2001	0.564	0.534	0.030	5.4

figure illustrates Gini values before and after taxes when including social security taxes with income taxes. The new law decreases the difference between before- and after-tax Gini values for 1999 from 0.025 to 0.022.

To investigate further, the percentage differences between before- and after-tax Gini values were computed and are shown as the fourth column in Figure G. These percentage changes in the Gini coefficient values, a "redistributive effect," show a decline ranging from 4.5 percent to 6.5 percent. As for the differences, the largest percentage changes are for the earliest years, a period when the marginal tax rates were high. The largest percentage reduction was for 1980, but the size of the reduction generally declined until 1986, fluctuated at relatively low levels between 1986 and 1992, and then increased from 1993 to 1996. However, coinciding with the capital gain tax reduction for 1997, the percentage change again declined for 1997 and 1998. Nevertheless, it increased for 1999, 2000, and 2001 (although the 2001 percentage increased slightly if the rebate is included with the 2000 data).

Figure H shows the Gini coefficients for before and after tax (including Social Security taxes) for 1979, 1989, 1999, and 1999 incorporating the new tax laws. The differences between before and after tax are much smaller than for the income tax, ranging from 0.018 for 1989 to 0.025 for 1979. This results in percentage differences of 3.4 percent to 5.4 percent. In all years, except 1999, the after-tax Gini coefficients are somewhat higher than those that result from simply including income taxes.

So, what does this all mean? First, the high marginal tax rates prior to 1982 appear to have had a significant redistributive effect. But, beginning with the tax rate reductions for 1982, this redistributive effect began to decline up to the period immediately prior to TRA 1986. Although TRA became effective for 1987, a surge in late 1986 capital gain realizations (to take advantage of the 60-percent long-term capital gain exclusion) effectively lowered the average tax rate for the highest income groups, thereby lessening the redistributive effect.

For the post-TRA period, the redistributive effect was relatively low, and it did not begin to increase until the initiation of the 39.6-percent tax bracket for 1993. But since 1997, with continuation of the 39.6-percent rate but with a lowering of the maximum tax rate on capital gains, the redistributive effect again declined. It appears that the new tax laws will continue this trend.

▶ Notes

1 Petska, Tom; Strudler, Mike; and Petska, Ryan (2003), New Estimates of Individual Income and Taxes, *2002 Proceedings of the 95th Annual Conference on Taxation, National Tax Association*.

2 Petska, Tom; Strudler, Mike; and Petska, Ryan (2000), Further Examination of the Distribution of Income and Taxes Using a Consistent and Comprehensive Measure of Income, *1999 Proceedings of the American Statistical Association, Social Statistics Section*.

3 Petska, Tom and Strudler, Mike, The Distribution of Individual Income and Taxes: A New Look at an Old Issue, presented at the annual meetings of

Figure H. Gini Coefficients for Retrospective Income (Including Social Security Taxes), Before and After Taxes, 1979-2001

Year	Gini Before Tax Including Social Security Taxes	Gini After Tax Including Social Security Taxes	Difference	Percent Difference
1979	0.469	0.444	0.025	5.354
1989	0.529	0.511	0.018	3.415
1999	0.574	0.549	0.025	4.340
1999 JGTRRA	0.574	0.553	0.022	3.790

the American Economic Association, New York, NY, January 1999, and published *in Turning Administrative Systems Into Information Systems: 1998-1999*.

[4] Petska, Tom and Strudler, Mike (1999), Income, Taxes, and Tax Progressivity: An Examination of Recent Trends in the Distribution of Individual Income and Taxes, *1998 Proceedings of the American Statistical Association, Social Statistics Section*.

[5] Nelson, Susan (1987), Family Economic Income and Other Income Concepts Used in Analyzing Tax Reform, *Compendium of Tax Research*, 1986, Office of Tax Analysis, U.S. Department of the Treasury.

[6] Hostetter, Susan (1988), Measuring Income for Developing and Reviewing Individual Tax Law Changes: Exploration of Alternative Concepts, *1987 Proceedings of the American Statistical Association, Survey Research Methods Section*.

[7] Internal Revenue Service, *Statistics of Income—Individual Income Tax Returns*, Publication 1304, (selected years).

[8] Parisi, Michael and Campbell, Dave, Individual Income Tax Rates and Tax Shares, 1999, *Statistics of Income (SOI) Bulletin,* Winter 2001-2002, Volume 21, Number 3.

[9] For the years 1979 through 1992, the percentile threshold size classes were estimated by osculatory interpolation as described in Oh and Oh and Scheuren.[10, 11] In this procedure, the data were tabulated into size classes, and the percentile thresholds were interpolated. For 1993 through 2000, the SOI individual tax return data files were sorted from highest to lowest, and the percentile thresholds were determined by cumulating records from the top down.

[10] Oh, H. Lock (1978), Osculatory Interpolations with a Monotonicity Constraint*, 1977 Proceedings of the American Statistical Association, Statistical Computing Section*.

[11] Oh, H. Lock and Scheuren, Fritz (1988), Osculatory Interpolations Revisited*, 1987 Proceedings of the American Statistical Association, Statistical Computing Section*.

[12] The CPI-U from the U.S. Department of Labor, *Monthly Labor Review*, was used for deflation of the income thresholds.

[13] Taxes, taxes paid, tax liabilities, tax shares, and average or effective tax rates are based on income tax, defined as income tax after credits plus alternative minimum tax (AMT) less the nonrefundable portion of the earned income credit (for 2000 and 2001, AMT was included in income tax after credits). However, for Figure F, tax includes Social Security and Medicare taxes less all of the earned income credit and refundable child credit.

[14] Internal Revenue Service, *Data Book 1999,* Publication 55B. Total Individual Income Taxes collected from withholding and additional taxes paid with tax forms filed were $1,102.2 billion, while total Social Security taxes were $587.5 billion.

[15] Actually, the OTA model was computed on 1998 individual income tax data and programmed to take all aspects of JGTRRA into account under the assumption that all of the sunset provisions will remain in place. After the results were calculated, the data were increased to 1999 levels. Therefore, income is exactly the same as the rest of the 1999 data, and only the taxes paid differs.

[16] A comparison of the before- and after-tax Gini coefficients does not exclusively measure the effects of the tax system in that the tax laws can also affect before-tax income. For example, capital gain realizations have been shown to be sensitive to the tax rates.

The Distribution of Household Income: Two Decades of Change[1]

Roberton Williams, Congressional Budget Office

Average household income grew by more than 40 percent in real terms between 1979 and 2000, climbing from $52,300 to $74,200 (in 2000 dollars). The rate of income growth varied sharply across the income distribution, however. Average real income of the lowest quintile—or fifth of the distribution—increased just 7 percent over the 21-year period, compared with a 70-percent gain for the top quintile. Growth was even faster at the very top of the distribution: real income for the top 5 percent of households more than doubled, and that of the top 1 percent nearly tripled. Income growth also varied across types of households: incomes of households with children increased at about the same rate as those of all households, incomes of nonelderly childless households grew more slowly, and incomes of the elderly climbed nearly half again as fast as the overall average.

This paper examines trends in household income between 1979 and 2000, utilizing a measure that includes both cash and in-kind income. To look at changes in the distribution of income, the paper divides households into quintiles and further subdivides the top quintile into four parts—the 80th-90th percentiles, the 90th-95th percentiles, the 95th-99th percentiles, and the top 1 percent.[2] Analysis includes both pretax and post-tax income to assess the effects of taxes on the distribution of income. The paper also separates households into three types based on the presence of children and the age of the household head. The paper is purely descriptive; it makes no attempt to examine why incomes changed as they did or to analyze patterns of change among components of income.

▶ Measuring Income

The principal income measure used in this paper is pretax comprehensive household income. That measure counts both cash and in-kind income, including:

- all cash income (both taxable and tax-exempt);
- taxes paid by businesses (the employer share of payroll taxes is imputed to workers, and corporate income taxes are imputed to owners of capital);
- employee contributions to 401(k) retirement plans; and
- the value of income received in kind from various sources (including employer-paid health insurance premiums, Medicare, Medicaid, food stamps, housing and energy assistance, and school breakfasts and lunches).

Income thus includes more sources than people often consider in assessing their well-being. Furthermore, income is counted when it is reported, generally for tax purposes. Thus, capital gains are included in income when they are realized, even though they may have accrued over many years and might thus be more appropriately counted on an accrual basis (that is, the increase or decrease in value in a given year would count as income in that year).[3] As a result, households that realize large gains in a given year may show higher up the income distribution than they would in an average year, and the distribution based on that measure of income would appear to be more unequal than if based on an accrued income measure.

A second measure—after-tax comprehensive household income—is comparable to the pretax measure described above but subtracts the major Federal taxes paid by the household: individual and corporate income taxes, payroll taxes, and excise taxes.[4] Comparing the two measures thus shows the impact of major Federal taxes on the distribution of income.

The unit of observation is the household—groups of people sharing the same living quarters, regardless of their relationships. Households thus include single people living by themselves, nuclear families with no nonfamily

members living with them, unmarried couples living together, and groups of unrelated people sharing a house or apartment, among many other possibilities. In most cases, members of households share all costs out of their common income, and the household is the appropriate unit over which to measure income. In some cases, the only sharing of income is for housing costs, and household members cover their own expenses for everything else; for those groups, using household as the unit of measurement misstates members' well-being.[5]

Placement of households in the income distribution is based not on total household income but rather on "adjusted household income." That measure takes account of the greater needs of larger households by dividing total household income by the square root of household size. Thus, for example, a person living alone and a four-person household with twice the total income of the single person would have the same adjusted income. Adjusted household income is used only to rank households in the income distribution. All dollar measures of income reported in the paper are total income, unadjusted for differences in household size. As a result, households of different sizes that are at the same point in the income distribution will have different incomes. Average income for a given segment of the distribution thus depends on the relative numbers of households of different sizes.

The choice of income measure and unit of observation affects how households are ranked within the income distribution.[6] Counting income from more sources moves households with income from those sources up the distribution relative to those not receiving such income. Using households rather than families as the unit of analysis lifts people in multifamily households up the distribution ahead of some people in single-family households. And adjusting income to account for the greater needs of larger households drops those larger households down the income distribution and consequently pushes smaller households up.

Quintiles contain equal numbers of people. Because households vary in size, quintiles generally contain unequal numbers of households. Income measures are broken down further by type of household: those with any members under age 18 (households with children), those headed by a person aged 65 or older and with no member under age 18 (elderly childless households), and all others (nonelderly childless households). The income and size of households vary more widely across those three groups than across all households; that means that the distributions of specific types of households among quintiles are more unequal than the distribution of all households.

▶ Pretax Income

Pretax household income increased by 42 percent in real terms between 1979 and 2000, with half of the gain coming in the last 5 years of the period (see Tables 1 and 2). Average income—measured in 2000 dollars—grew from about $52,300 in 1979 to $63,200 in 1995 and $74,200 in 2000.

Much of the gain, however, came from much more rapid increases in income of households with the highest incomes, and average incomes of all but the highest quintile climbed much more slowly than the overall average. Average real income of the lowest quintile increased 7 percent over the period, that of the middle quintile 13 percent, and that of the highest quintile 70 percent. The disparity in growth rates was just as pronounced within the top quintile: incomes of households in the 80th to 90th percentiles rose 32 percent over the two decades compared with 43-percent growth for those in the 90th to 95th percentiles, 60-percent growth for those in the 95th to 99th percentiles, and 185-percent growth for the top 1 percent of households.[7] Had average income of households in the top 1 percent grown at the same rate as that for all other households, overall growth would have been one-third less over the period: 28 percent.

Because of those differences in growth rates of incomes, differentials among incomes increased sharply over the period. In 1979, households in the top quintile had average income 2.6 times that of households in the middle quintile and 8.5 times that of households in the lowest quintile. By 1995, those ratios had increased by about one-quarter to 3.3 and 10.9, respectively, and by 2000, they had risen another fifth to 3.9 and 13.5.

▶ The Impact of Federal Taxes

Federal taxes reduce after-tax income below their pretax levels. Overall, Federal taxes claimed 23 percent of pretax income in 2000, slightly more than the 22 percent taken in taxes in 1979. Because that change in effective Federal tax rates was small, after-tax incomes grew at roughly the same rate as their pretax equivalents. The overall average after-tax income in 2000 dollars increased 40 percent from $40,700 in 1979 to $57,000 in 2000 (see Tables 3 and 4).

Federal taxes are highly progressive, claiming a larger share of high incomes than of low incomes (see Table 5). In 2000, for example, households in the lowest income quintile paid Federal taxes equal to about 6 percent of pretax income, while households in the middle quintile paid nearly 17 percent, and those in the highest quintile paid 28 percent. Effective tax rates for the bottom four quintiles fell between 1979 and 2000, thus yielding larger percentage gains in after-tax income than in pretax income. In contrast, tax rates were slightly higher in 2000 than 1979 for the highest quintile. At the very top of the income distribution, however, tax rates dropped sharply over the period; the top 1 percentile's tax rate fell from 37 percent to 33 percent, increasing the rate of growth of after-tax income 16 percentage points above that of pretax income.

The progressivity of the Federal tax system mitigates the differential of incomes across quintiles. Because tax rates changed little over the 1979-2000 period, however, taxes did little to offset the growing disparity of incomes across quintiles. Average after-tax income for the highest quintile was 3.4 times that for the middle quintile, substantially below the 3.9 level for pretax incomes but nearly 50 percent above the 1979 after-tax ratio of 2.3. Federal taxes in 2000 did little more to equalize incomes across the distribution than they did in 1979 and had very little effect on the rapidly widening income differences across quintiles.

▶ Changing Shares of Income

The different rates of income growth across quintiles resulted in a shift in shares of income from lower income categories to higher ones, both for pretax and after-tax incomes (see Table 6). The share of pretax income going to the lowest quintile declined from 5.8 percent in 1979 to 4.0 percent in 2000, and that for the middle quintile fell from 15.8 percent to 13.5 percent. In contrast, the share going to the highest quintile increased from 45.5 percent to 54.9 percent and that for the top 1 percent of households nearly doubled from 9.3 percent to 17.8 percent.

After-tax income shares showed similar patterns, starting from slightly greater equality but shifting relatively more. The share of after-tax income going to the lowest quintile fell from 6.8 percent in 1979 to 4.9 percent in 2000; for the middle quintile, the decline was from 16.5 percent to 14.7 percent. Gains at the upper end of the distribution were substantial: the share going to the highest quintile rose from 42.4 percent to 51.3 percent, and that going to the top 1 percent more than doubled from 7.5 percent to 15.5 percent, in part because of the 10-percent drop in their effective tax rate and in part because of the large income gains they experienced.

▶ Changes Across Types of Household

Average pretax income varies across types of household: households with children have incomes above those of nonelderly childless households, and both groups have average incomes above that of elderly childless households (see Table 7). In 2000, for example, households with children had an average pretax income of $85,300, about 19 percent greater than the $71,900 average for nonelderly childless households and 42 percent above the $60,100 average for elderly childless households.

Incomes of the three types of household grew at different rates over the past two decades. Average pretax income of the elderly climbed most rapidly, rising 57 percent between 1979 and 2000. In comparison, average income of households with children increased 43 percent, and that of other households rose 38 percent. Average income of households with children grew more slowly than those of the other two types of household during the 1980's but rose faster in the 1990's to make up some of the difference. The more rapid income growth for elderly households raised their average income from 73 percent of that for all households in 1979 to 81 percent in 2000.

The three types of household face quite different effective tax rates, largely because their incomes derive from sources that are taxed differentially but also because the tax code treats them differently. The elderly get more of their income in kind—principally from Medicare—and a large part of their Social Security benefits are not subject to income tax. Furthermore, because relatively little of their income comes from earnings, payroll taxes claim a smaller share of their income than is the case for households getting more income from work. At the same time, households with children are generally larger than other households and qualify for more dependent exemptions that serve to lower their income tax liabilities. As a result of those differences, elderly households face a lower than average effective Federal tax rate, and nonelderly childless households incur a higher than average rate (see Table 8). The effective tax rate for households with children was close to that for all households throughout the past two decades. Over that period, effective tax rates rose for both households with children and other nonelderly households but declined for the elderly.

Differential tax rates served to narrow the gap between incomes of the elderly and other households, and that effect grew over the past two decades (see Table 9). Average after-tax income of elderly households increased 60 percent from $30,500 in 1979 to $48,800 in 2000. Over the same period, average after-tax income of households with children rose 40 percent from $47,000 to $65,700, and that for other nonelderly households climbed 36 percent from $39,500 to $53,900. The more rapid income growth for the elderly raised their average after-tax income from 75 percent of that for all households in 1979 to 86 percent in 2000.

▶ Caveats

The preceding analysis should be viewed with caution for many reasons. First, the study compares income groups over time, showing how incomes have changed for each quintile. The composition of each quintile changes, however, from year to year. Over time, people join and leave households, enter and leave the labor force, and experience other changes that can alter their positions in the income distribution. Trends in tax rates and income that are discussed here reflect what has happened to people in the same parts of the distribution over time, not what has happened to the same people.

Second, expanding the income measure for calculating effective tax rates to include taxes paid by businesses, employee contributions to 401(k) plans, and in-kind benefits makes that measure larger than what many people think of when they consider their own incomes. As a result, it may be difficult for readers to determine their own placement within the reported distributions. Third, adjusting income for the size of households in order to rank them substantially reorders those units throughout the income distribution. Consequently, total household income can vary markedly among households of differing size, even though they are closely ranked in the distribution.[8]

Fourth, any choice of a period over which to assess changes in effective tax rates or incomes is arbitrary. Many of the comparisons made in this paper compare incomes in 1979 against those in 2000. Changes over other periods may show markedly different patterns. For example, between 1979 and 2000, average household income rose 42 percent. But it also rose 42 percent between 1983 and 2000, having lost and recovered 4 percent of its value between 1979 and 1983. What period is most appropriate depends on the question posed. The tables provide measures of income for every year of the 1979-2000 period and thus allow an evaluation of changes between any pair of years.

Finally, the study looks only at annual income and taxes. A better indication of the well-being of households at different points in the income distribution would cover a longer period—ideally, each person's lifetime. That kind of time frame would remove the effects of year-to-year variations and avoid the problem that information about a single year might differ markedly from average values for longer periods. For example, households realize capital gains irregularly over time, some years having large gains and other years having none. The annual income measures used in this paper count gains in the year they are realized, even though they may represent accumulations of wealth over many years. Using a lifetime measure of income would avoid that shortcoming and yield a flatter income distribution than the one shown here.

▶ Footnotes

[1] The views expressed in this paper are those of the author and should not be interpreted as those of the Congressional Budget Office.

[2] That breakdown reveals the effects of the disproportionately rapid growth of income that has occurred over the past two decades at the top end of the distribution. The analysis does not show a comparable subdivision of the lowest quintile because income moved in similar ways for households in different parts of that income group.

[3] Available data do not allow the counting of accrued gains, thus precluding that alternative measure of income.

[4] Other Federal taxes, representing about 5 percent of Federal revenue collections, are not counted. The omitted taxes include estate and gift taxes, customs duties, and other miscellaneous revenues.

[5] See Chapter 2 of Congressional Budget Office, *Effective Federal Tax Rates: 1979-1997* (October 2001) for further discussion of methodological questions concerning the appropriate unit of observation and measure of income.

[6] Previous work has compared the income measure used in this paper with alternatives to assess the effects of the choice of measure. See *ibid*.

[7] It is important to remember that households in a given income category in one year are not the same as those in that category in another year. Households move up and down the distribution over time, faring better or worse than average. At the same time, changes in income for quintiles or other parts of the income distribution do indicate how income gains are shared among households.

[8] Statistics based on household cash income that omits in-kind income and is unadjusted for household size may provide information that is more consistent with how most people think about their own tax and income situations. This paper does not examine such a measure. For information on unadjusted cash incomes, see Congressional Budget Office, *op. cit*.

Table 1. Total Pretax Household Income by Quintile, All Households, 1979-2000 (in 2000 dollars)

Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	All Quintiles	80th-90th Percentile	90th-95th Percentile	95th-99th Percentile	Top 1 Percent
1979	13,700	29,800	44,700	60,500	115,800	52,300	78,600	96,800	142,400	454,200
1980	13,200	28,600	43,300	58,700	112,100	50,700	76,500	95,300	136,900	428,400
1981	12,900	28,400	42,700	59,000	111,400	50,600	76,400	95,000	134,800	425,900
1982	12,500	27,900	42,200	58,700	112,400	50,700	76,400	95,200	135,300	447,800
1983	12,100	26,700	41,200	58,100	115,400	50,800	76,500	96,600	138,400	487,400
1984	12,500	28,300	43,000	60,600	123,600	53,100	80,200	102,400	148,400	537,100
1985	12,600	28,400	43,800	61,200	127,000	54,500	81,100	103,800	151,500	577,400
1986	12,600	29,100	44,800	63,500	143,300	58,500	85,700	109,500	166,300	751,500
1987	12,400	28,000	44,500	63,600	134,900	56,500	86,100	111,000	164,500	607,200
1988	12,700	28,600	45,200	64,300	145,200	59,000	88,100	112,800	170,500	765,000
1989	13,100	29,100	45,700	65,100	144,400	59,500	88,800	115,200	174,100	712,100
1990	13,500	29,900	45,500	64,300	140,300	58,800	87,400	112,400	168,400	683,400
1991	13,500	29,400	44,800	63,600	135,300	57,300	86,100	111,000	164,600	615,900
1992	13,200	29,200	45,100	64,200	141,300	58,900	87,100	112,800	171,600	698,600
1993	13,500	29,400	45,300	64,700	141,300	59,100	88,200	114,000	171,000	671,000
1994	13,400	29,600	45,700	65,900	144,200	59,800	89,400	116,000	175,500	692,100
1995	14,300	31,200	47,500	67,500	155,600	63,200	93,600	121,200	191,800	783,800
1996	13,900	30,900	47,500	68,000	158,500	64,100	93,400	122,800	190,700	841,000
1997	14,200	31,600	48,300	69,300	168,100	66,700	95,600	126,500	202,600	964,600
1998	14,900	33,000	49,500	72,000	178,300	69,900	99,000	131,400	213,400	1,083,300
1999	15,300	33,900	50,500	73,800	189,100	73,000	102,300	136,300	222,400	1,181,000
2000	14,600	33,300	50,300	74,500	196,900	74,200	103,700	139,000	227,800	1,295,300

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Table 2. Pretax Household Income Relative to 1979 by Quintile, All Households, 1979-2000 (in percent of 1979 income)

Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	All Quintiles	80th-90th Percentile	90th-95th Percentile	95th-99th Percentile	Top 1 Percent
1979	100	100	100	100	100	100	100	100	100	100
1980	96	96	97	97	97	97	97	98	96	94
1981	94	95	96	98	96	97	97	98	95	94
1982	91	94	94	97	97	97	97	98	95	99
1983	88	90	92	96	100	97	97	100	97	107
1984	91	95	96	100	107	102	102	106	104	118
1985	92	95	98	101	110	104	103	107	106	127
1986	92	98	100	105	124	112	109	113	117	165
1987	91	94	100	105	116	108	110	115	116	134
1988	93	96	101	106	125	113	112	117	120	168
1989	96	98	102	108	125	114	113	119	122	157
1990	99	100	102	106	121	112	111	116	118	150
1991	99	99	100	105	117	110	110	115	116	136
1992	96	98	101	106	122	113	111	117	121	154
1993	99	99	101	107	122	113	112	118	120	148
1994	98	99	102	109	125	114	114	120	123	152
1995	104	105	106	112	134	121	119	125	135	173
1996	101	104	106	112	137	123	119	127	134	185
1997	104	106	108	115	145	128	122	131	142	212
1998	109	111	111	119	154	134	126	136	150	239
1999	112	114	113	122	163	140	130	141	156	260
2000	107	112	113	123	170	142	132	144	160	285

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Table 3. Total After-Tax Household Income by Quintile, All Households, 1979-2000 (in 2000 dollars)

Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	All Quintiles	80th-90th Percentile	90th-95th Percentile	95th-99th Percentile	Top 1 Percent
1979	12,600	25,600	36,400	47,700	84,000	40,700	60,400	72,700	103,000	286,300
1980	12,200	24,600	35,200	46,100	81,500	39,400	58,300	71,100	98,900	280,300
1981	11,900	24,200	34,500	46,000	81,500	39,200	58,000	70,400	97,700	290,400
1982	11,500	24,100	34,700	46,700	85,200	40,200	59,200	72,800	102,200	326,200
1983	11,000	23,000	34,000	46,400	87,900	40,400	59,800	74,400	105,600	352,200
1984	11,200	24,100	35,300	48,200	93,600	42,000	62,400	78,600	112,600	385,500
1985	11,400	24,200	35,800	48,700	96,500	43,100	62,900	79,700	115,400	421,500
1986	11,400	24,800	36,800	50,500	109,200	46,300	66,300	83,900	127,000	559,900
1987	11,300	24,100	36,700	50,800	100,100	44,300	66,600	84,100	121,700	417,800
1988	11,600	24,500	37,100	51,000	108,000	46,100	67,700	85,800	126,900	537,900
1989	12,100	25,100	37,500	51,800	108,000	46,700	68,600	87,400	129,800	506,500
1990	12,300	25,500	37,400	51,000	105,000	46,200	67,400	85,400	125,800	486,800
1991	12,300	25,200	36,900	50,500	101,000	45,000	66,500	84,400	122,500	431,900
1992	12,100	25,200	37,200	51,200	105,100	46,200	67,400	85,700	127,600	484,900
1993	12,400	25,400	37,500	51,600	103,500	46,000	68,300	86,400	125,400	439,800
1994	12,600	25,700	37,800	52,400	104,700	46,500	69,000	87,400	128,100	444,500
1995	13,400	27,100	39,300	53,800	113,500	49,200	72,200	91,400	140,800	515,200
1996	13,100	26,800	39,300	54,200	114,100	49,500	71,900	92,100	137,800	538,200
1997	13,400	27,300	39,900	55,100	121,000	51,400	73,300	94,900	146,300	627,700
1998	14,000	28,700	41,200	57,300	129,100	54,100	76,200	98,600	154,500	721,100
1999	14,400	29,300	42,000	58,700	136,000	56,200	78,400	101,700	159,700	783,600
2000	13,700	29,000	41,900	59,200	141,400	57,000	79,500	103,600	163,200	862,400

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Table 4. After-Tax Household Income Relative to 1979 by Quintile, All Households, 1979-2000 (in percent of 1979 income)

Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	All Quintiles	80th-90th Percentile	90th-95th Percentile	95th-99th Percentile	Top 1 Percent
1979	100	100	100	100	100	100	100	100	100	100
1980	97	96	97	97	97	97	97	98	96	98
1981	94	95	95	96	97	96	96	97	95	101
1982	91	94	95	98	101	99	98	100	99	114
1983	87	90	93	97	105	99	99	102	103	123
1984	89	94	97	101	111	103	103	108	109	135
1985	90	95	98	102	115	106	104	110	112	147
1986	90	97	101	106	130	114	110	115	123	196
1987	90	94	101	106	119	109	110	116	118	146
1988	92	96	102	107	129	113	112	118	123	188
1989	96	98	103	109	129	115	114	120	126	177
1990	98	100	103	107	125	114	112	117	122	170
1991	98	98	101	106	120	111	110	116	119	151
1992	96	98	102	107	125	114	112	118	124	169
1993	98	99	103	108	123	113	113	119	122	154
1994	100	100	104	110	125	114	114	120	124	155
1995	106	106	108	113	135	121	120	126	137	180
1996	104	105	108	114	136	122	119	127	134	188
1997	106	107	110	116	144	126	121	131	142	219
1998	111	112	113	120	154	133	126	136	150	252
1999	114	114	115	123	162	138	130	140	155	274
2000	109	113	115	124	168	140	132	143	158	301

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Table 5. Effective Federal Tax Rates by Quintile, All Households, 1979-2000 (in percent)

Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	All Quintiles	80th-90th Percentile	90th-95th Percentile	95th-99th Percentile	Top 1 Percent
1979	8.0	14.1	18.6	21.2	27.5	22.2	23.1	24.9	27.7	37.0
1980	7.6	14.0	18.7	21.5	27.3	22.3	23.8	25.4	27.8	34.6
1981	7.8	14.8	19.2	22.0	26.8	22.5	24.0	25.9	27.5	31.8
1982	8.0	13.6	17.8	20.4	24.2	20.7	22.6	23.5	24.5	27.2
1983	9.1	13.9	17.5	20.1	23.8	20.5	21.8	23.0	23.7	27.7
1984	10.4	14.8	17.9	20.5	24.3	20.9	22.2	23.3	24.1	28.2
1985	9.5	14.8	18.3	20.4	24.0	20.9	22.4	23.2	23.8	27.0
1986	9.5	14.8	17.9	20.5	23.8	20.9	22.6	23.3	23.6	25.5
1987	8.9	13.9	17.5	20.1	25.8	21.6	22.7	24.3	26.0	31.2
1988	8.7	14.3	17.9	20.7	25.6	21.9	23.2	24.0	25.6	29.7
1989	7.6	13.7	17.9	20.4	25.2	21.5	22.7	24.1	25.4	28.9
1990	8.9	14.7	17.8	20.7	25.2	21.4	22.9	24.0	25.3	28.8
1991	8.9	14.3	17.6	20.6	25.4	21.5	22.8	24.0	25.6	29.9
1992	8.3	13.7	17.5	20.2	25.6	21.6	22.6	24.0	25.6	30.6
1993	8.1	13.6	17.2	20.2	26.8	22.2	22.5	24.2	26.6	34.5
1994	6.0	13.2	17.3	20.5	27.4	22.2	22.9	24.7	27.0	35.8
1995	6.3	13.1	17.3	20.3	27.1	22.2	22.9	24.6	26.6	34.3
1996	5.8	13.3	17.3	20.3	28.0	22.8	22.9	24.9	27.7	36.0
1997	5.6	13.6	17.4	20.5	28.0	22.9	23.3	25.0	27.8	34.9
1998	6.0	13.0	16.8	20.4	27.6	22.6	23.1	25.0	27.6	33.4
1999	5.9	13.6	16.8	20.5	28.1	23.0	23.3	25.4	28.2	33.6
2000	6.2	12.9	16.7	20.5	28.2	23.2	23.4	25.4	28.3	33.4

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: Effective tax rates include individual and corporate income taxes, payroll taxes, and excise taxes. They omit estate and gift taxes, customs duties, and other miscellaneous collections; revenues from those sources total about 5 percent of all Federal revenues.

Table 6. Shares of Pretax and After-Tax Household Income by Quintile, All Households, 1979-2000 (in percent)

	Shares of Pretax Income					Shares of After-Tax Income				
Year	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile	Lowest Quintile	Second Quintile	Middle Quintile	Fourth Quintile	Highest Quintile
1979	5.8	11.1	15.8	22.0	45.4	6.8	12.2	16.5	22.2	42.3
1980	5.7	10.9	15.7	22.0	45.7	6.7	12.1	16.4	22.2	42.6
1981	5.4	10.9	15.8	22.1	45.8	6.4	11.9	16.4	22.2	43.1
1982	5.1	10.6	15.7	22.1	46.5	5.9	11.5	16.2	22.1	44.3
1983	4.8	10.2	15.4	22.1	47.5	5.5	11.1	15.9	22.1	45.4
1984	5.0	10.2	15.3	21.9	47.7	5.6	11.0	15.8	22.0	45.6
1985	4.8	10.1	15.1	21.7	48.3	5.4	10.8	15.6	21.8	46.3
1986	4.5	9.5	14.6	21.1	50.3	5.1	10.2	15.2	21.1	48.3
1987	4.3	9.9	15.2	22.0	48.6	5.0	10.8	16.0	22.3	45.9
1988	4.2	9.7	14.7	21.4	49.9	4.9	10.6	15.4	21.7	47.4
1989	4.3	9.7	14.9	21.4	49.6	5.0	10.7	15.6	21.7	47.1
1990	4.5	9.9	15.0	21.5	49.2	5.2	10.7	15.6	21.6	46.8
1991	4.7	9.9	15.3	21.6	48.5	5.4	10.7	16.0	21.8	46.0
1992	4.4	9.7	14.9	21.4	49.6	5.1	10.6	15.7	21.7	46.9
1993	4.5	9.7	14.9	21.4	49.4	5.2	10.8	15.8	21.9	46.3
1994	4.4	9.7	15.1	21.5	49.4	5.3	10.8	16.0	21.9	46.0
1995	4.5	9.6	14.6	21.0	50.2	5.4	10.7	15.5	21.4	46.9
1996	4.3	9.4	14.4	20.8	51.1	5.2	10.5	15.4	21.4	47.5
1997	4.2	9.1	14.2	20.3	52.2	5.2	10.2	15.1	20.9	48.6
1998	4.3	8.9	14.0	20.1	52.7	5.2	10.0	15.0	20.6	49.2
1999	4.2	8.8	13.7	19.8	53.5	5.1	9.9	14.8	20.4	49.8
2000	4.0	8.6	13.4	19.5	54.6	4.9	9.7	14.5	20.1	50.9

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Table 7. Average Pretax Household Income by Quintile and Household Type, 1979-2000 (in 2000 dollars)

	Average Pretax Household Income				Pretax Income Relative to 1979			
Year	All	With Children	Elderly Childless	Nonelderly Childless	All	With Children	Elderly Childless	Nonelderly Childless
1979	52,300	59,500	38,200	52,200	100	100	100	100
1980	50,700	57,000	37,900	50,900	97	96	99	98
1981	50,600	56,500	39,100	50,600	97	95	102	97
1982	50,700	55,500	40,600	51,200	97	93	106	98
1983	50,800	55,600	41,300	51,100	97	93	108	98
1984	53,100	58,900	43,600	52,500	102	99	114	101
1985	54,500	59,200	45,100	54,900	104	99	118	105
1986	58,500	64,000	48,800	58,300	112	108	128	112
1987	56,500	61,700	44,700	57,800	108	104	117	111
1988	59,000	63,300	46,400	61,500	113	106	121	118
1989	59,500	65,000	47,100	61,000	114	109	123	117
1990	58,800	63,700	46,100	61,200	112	107	121	117
1991	57,300	62,800	44,100	59,300	110	106	115	114
1992	58,900	63,800	45,700	61,100	113	107	120	117
1993	59,100	65,400	45,700	60,100	113	110	120	115
1994	59,800	66,100	46,400	61,200	114	111	121	117
1995	63,200	70,900	50,400	62,900	121	119	132	120
1996	64,100	70,400	53,300	63,900	123	118	140	122
1997	66,700	73,400	56,100	66,100	128	123	147	127
1998	69,900	77,800	59,000	68,500	134	131	154	131
1999	73,000	82,100	60,900	71,200	140	138	159	136
2000	74,200	85,300	60,100	71,900	142	143	157	138

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Table 8. Effective Federal Tax Rates by Type of Household, 1979-2000 (in percent)

Year	All	With Children	Elderly Childless	Nonelderly Childless
1979	22.2	21.0	20.2	24.3
1980	22.3	21.4	19.0	24.4
1981	22.5	21.9	17.6	24.7
1982	20.7	20.5	15.0	22.9
1983	20.5	20.5	15.5	22.3
1984	20.9	20.9	16.5	22.7
1985	20.9	20.8	16.4	23.0
1986	20.9	21.1	16.0	22.8
1987	21.6	21.4	17.0	23.7
1988	21.9	21.3	17.2	23.9
1989	21.5	21.4	17.0	23.4
1990	21.4	21.4	16.5	23.7
1991	21.5	21.5	15.6	23.8
1992	21.6	21.2	16.2	23.9
1993	22.2	21.9	16.8	24.1
1994	22.2	21.9	17.5	24.7
1995	22.2	22.0	17.5	24.2
1996	22.8	22.2	19.1	24.7
1997	22.9	22.5	18.9	25.0
1998	22.6	22.0	18.5	24.7
1999	23.0	22.5	19.0	25.0
2000	23.2	23.0	18.8	25.0

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: Effective tax rates include individual and corporate income taxes, payroll taxes, and excise taxes. They omit estate and gift taxes, customs duties, and other miscellaneous collections; revenues from those sources total about 5 percent of all Federal revenues.

Table 9. Average After-Tax Household Income by Quintile and Household Type, 1979-2000 (in 2000 dollars)

	Average Pretax Household Income				Pretax Income Relative to 1979			
Year	All	With Children	Elderly Childless	Nonelderly Childless	All	With Children	Elderly Childless	Nonelderly Childless
1979	40,700	47,000	30,500	39,500	100	100	100	100
1980	39,400	44,800	30,700	38,500	97	95	101	97
1981	39,200	44,100	32,200	38,100	96	94	106	96
1982	40,200	44,100	34,500	39,500	99	94	113	100
1983	40,400	44,200	34,900	39,700	99	94	114	101
1984	42,000	46,600	36,400	40,600	103	99	119	103
1985	43,100	46,900	37,700	42,300	106	100	124	107
1986	46,300	50,500	41,000	45,000	114	107	134	114
1987	44,300	48,500	37,100	44,100	109	103	122	112
1988	46,100	49,800	38,400	46,800	113	106	126	118
1989	46,700	51,100	39,100	46,700	115	109	128	118
1990	46,200	50,100	38,500	46,700	114	107	126	118
1991	45,000	49,300	37,200	45,200	111	105	122	114
1992	46,200	50,300	38,300	46,500	114	107	126	118
1993	46,000	51,100	38,000	45,600	113	109	125	115
1994	46,500	51,600	38,300	46,100	114	110	126	117
1995	49,200	55,300	41,600	47,700	121	118	136	121
1996	49,500	54,800	43,100	48,100	122	117	141	122
1997	51,400	56,900	45,500	49,600	126	121	149	126
1998	54,100	60,700	48,100	51,600	133	129	158	131
1999	56,200	63,600	49,300	53,400	138	135	162	135
2000	57,000	65,700	48,800	53,900	140	140	160	136

Source: Congressional Budget Office tabulations of data from the 1979-2000 Statistics of Income (Internal Revenue Service) and from the March 1980-March 2001 Current Population Survey (Bureau of the Census).

NOTE: See text for discussion of methodology used to combine data bases.

Comments on Papers by Welniak; Strudler, Petska, and Petska; and Williams [1]

Eric. J. Toder, Internal Revenue Service

These three papers by analysts at the U.S. Bureau of the Census, the Statistics of Income (SOI) Division of the Internal Revenue Service, and the Congressional Budget Office (CBO) estimate changes in the distribution of income over the past two decades. In my remarks, I first address what the three papers have in common. Then, I discuss some of the main issues in measuring the distribution of income and compare how these three papers addressed these issues. I comment on the strengths and weaknesses of alternative approaches. Finally, I briefly discuss some implications of the authors' findings. For brevity, I reference the papers in the discussion by the institutional affiliations of the authors (Census, SOI, and CBO).

▶ Common Features of the Three Papers

All the papers measure changes in the distribution of income over the past two decades. Census estimates changes in the distribution of pretax income between 1979 and 2000 and also extends its analysis back to 1967. SOI estimates changes in the distribution of pretax and post-tax income between 1979 and 2001. CBO looks at the changes in a broader measure of pretax income that includes taxes paid by businesses between 1979 and 2000.

All the papers show that inequality has increased over the past two decades. Census shows that Gini coefficients and other commonly used inequality measures have increased. SOI shows that pretax and post-tax Gini coefficients have increased, that income cutoffs at the top percentiles of the income distribution have increased faster than income cutoffs for lower percentile groupings, and that the share of income going to the highest percentiles of the population has also increased. CBO shows that average pretax income has grown faster for the higher percentile groupings than for other population groups.

All three papers compare "snapshots" of the income distribution in different years. That is, they compare dispersions of income among samples of the population, but the individuals in the sample change over time. Thus, none of the studies is examining changes over time in the incomes of a fixed group of individuals, as would be done by a panel study. An alternative and more conceptually appealing way to look at income distribution is to measure the dispersion of lifetime incomes across a fixed population, but available data do not facilitate comparing how the dispersion of lifetime incomes changes over time. Compared with a distribution of lifetime incomes, the "snapshot" distributions in these papers overstate inequality for two reasons. First, they include some individuals whose incomes are temporarily high or low in a given year because of, for example, windfall gains or a spell of unemployment. Second, they include individuals at different ages; so, a portion of the inequality reflects the variation in incomes over a person's lifecycle and not lifetime difference in incomes among people. While the papers overstate the level of inequality, however, it does not follow that they overstate the increase in inequality.

▶ Methodological Issues in Measuring Income Distribution and How Papers Address Them

While the papers reach similar conclusions, they differ significantly in their approaches. This reflects the numerous methodological issues that researchers confront in measuring income distribution. The differences in part also reflect differences in the types of data produced by the agencies where the researchers work. In this section, I discuss the strengths and weaknesses of different approaches and compare the choices made in the three papers.

Choice of an Income Concept. The first question is what income concept to use, given the existence of taxes and Government transfer programs. The two conceptually pure alternatives are to look at income that people receive from market transactions—that is, income in the absence of Government taxes and transfers—or to look at income net of all Government taxes and transfers. The latter is the best measure of the well-being of

individuals, while the distribution of market income indicates how the income distribution might have changed, absent changes in the tax law.

None of the authors estimates the distribution of market income, while only CBO displays changes net of Federal Government taxes and transfers.[2] Instead, CBO measures income before taxes but including transfers, while Census and SOI measure income before individual income taxes but net of business taxes. SOI also measures income net of taxes, while CBO measures income net of both taxes and transfers.

Measuring pretax income is not straightforward. We observe reported income of individuals before taxes, but we do not really know whose incomes are reduced how much by taxes. In the case of the individual income tax, it is typical, though not strictly correct, for researchers to assume that the tax reduces the after-tax incomes of those who pay it, but does not affect anyone's observed pretax income.[3] But there are differing views on which individuals experience lower after-tax incomes as a result of taxes remitted by businesses. CBO allocates the employer portion of payroll taxes in proportion to wages received, and the corporate income tax in proportion to investment income (interest, dividends, capital gains) of individuals, and adds these taxes back to observed income to derive its measure of pretax income. These are reasonable assumptions, but not the only possible ones.

Inclusiveness of Income Measure. Economic income is defined as the sum of consumption plus changes in net worth. By this broad measure, income includes all sources of cash receipts, net of costs of earning income—wages, interest, dividends, rents, and business profits—plus changes in the value of assets (adjusted for inflation), income from noncash fringe benefits, and the net imputed value of consumption services from durable goods (principally houses). None of the authors uses this broad a measure of income, although the U.S. Treasury Department has used such a measure (called "family economic income") in analyses of the distributional effect of Federal taxes. See Cronin (1999).

All the authors include cash flow income (wages, interest, dividends, rent, profits) in their income measures. CBO, as noted above, also adds back business taxes to arrive at a broader measure of pretax income. No one counts accrued capital gains or other forms of accrued income (such as the inside buildup on pensions and life insurance reserves), but CBO and SOI include realized gains reported on tax returns. CBO and Census include cash transfer payments in their measures, but SOI does not. CBO also imputes some in-kind benefits received, such as the value of employer-provided health insurance.

Making the income measure broader improves it as a measure of economic well-being, but can come at a cost for items not reported in the primary data source (see below), but imputed from other data sets. Researchers confront a tradeoff between the quality of the income concept and the precision of the data.

Unit of Measurement. Another issue is how to define the unit of comparison. Because people who are related or live together typically pool their incomes, most researchers do not examine the distribution of income across individuals. Census and CBO use the household as their unit of analysis, while SOI uses tax filing units. In general, comparing incomes across households is preferable to comparison across tax units for two reasons. First, tax units exclude nonfilers, and therefore miss many households at the low end of the income distribution (although they do include low-income people without a filing requirement who file a return to get refunds of withheld taxes or to claim refundable credits.) Second, tax units include some individuals, such as many students, whose economic well-being is represented better by the incomes of their families than by their individual incomes.

A related issue, if the household or family is the unit of measure, is if and how to adjust for differences in family size. The same income supports differing standards of living for households of different compositions, and changes in the composition of households (by marital status and household size) over time can affect trends in measures of income distribution. Among the authors, only CBO includes an explicit adjustment for family size in the analysis. CBO also reports trends in income distribution within more homogenous subgroups—elderly childless households, nonelderly childless households, and households with children. In particular, they find that

incomes of elderly childless households have increased more over the past two decades than incomes of other household types.

Sources of Data. Not surprisingly, the authors use the sources of data their agencies produce—Census uses data from their Current Population Survey (CPS), while SOI uses administrative tax data from a sample of individual tax returns. CBO performs a statistical match between CPS and SOI data; the CPS sample is used as the basis for the CBO households, while SOI data are the basis for estimated incomes.

Each approach has strengths and weaknesses. Typically, administrative data are more accurate and complete than survey data; in particular, income from capital reported on tax returns is much larger than income from capital reported to CPS and much closer to totals in the National Income and Product Accounts. But SOI data are limited to what people are required to report on their tax returns, while the CPS collects a broader range of data and includes a representative national sample of households, not just tax filers. (SOI does include data on realized capital gains, which are not collected by Census.) CBO attempts to get the best of both worlds by merging tax return and CPS data, but the use of statistical matching procedures means that incomes are in part estimated rather than observed.

Measures of Inequality. Finally, the researchers use different indices to measure inequality. SOI and CBO (but not Census) examine changes in income shares among percentile groups. SOI also measures changes in the income levels at which percentile breaks begin. Both Census and SOI, but not CBO, estimate changes in the "Gini" coefficient, a commonly used overall index of inequality. Census also reports alternative summary measures that apply different weights to different parts of the income distribution. SOI and CBO, but not Census, compare changes in pre- and post-tax measures of inequality. In spite of this diversity, all the measures used show rising inequality over the past two decades.

▶ Concluding Comments

These are excellent papers and good examples of the careful and high-quality research performed within U.S. Government agencies. While the authors address difficult methodological issues in diverse ways, they reach broadly similar conclusions about trends in income distribution. Using measures of annual income, the dispersion of income has clearly increased. While this does not definitively establish that the distribution of lifetime income has become less equal, it certainly provides cause for concern about widening inequality in the United States.

How much this all has to do with Government fiscal policies, however, is not clear. Inequality widened in the 1980's, as tax rates, especially on high-income individuals, were falling. Inequality also widened in the 1990's, when tax rates on high-income individuals were increased. Since 2001, in the face of new tax cuts, measures of inequality may be narrowing as a result of the recent decline in stock prices, which disproportionately affects reported incomes (especially from capital gains) of high-income individuals. This suggests that tax policies, while modifying market outcomes, are probably not the major driver of the changes in income distribution.

▶ Notes and References

[1] See Welniak (2003); Strudler, Petska, and Petska (2003); and Williams (2003), this volume.

[2] The CBO measure does not include the effects of State and local taxes and transfers.

[3] This assumption is a good approximation, but does not hold in all cases. For example, tax-exempt municipal bonds pay lower interest rates than taxable securities of comparable risk. Recipients of income from tax-exempt bonds do not pay taxes to the Federal Government, but do receive lower incomes from those securities than they would have, absent a Federal income tax.

Cronin, Julie-Anne (1999), "U.S. Treasury Distributional Analysis Methodology," OTA Paper 85 (September).

Strudler, Michael; Petska, Tom; and Petska, Ryan (2003), "An Analysis of the Distribution of Individual Income and Taxes, 1979-2001, presented at the Joint Statistical Meetings, San Francisco, California (August).

Welniak, Edward J. (2003), "Measuring Household Inequality Using the CPS," presented at the Joint Statistical Meetings, San Francisco, California (August).

Williams, Roberton C., Jr. (2003), "The Distribution of Household Income: Two Decades of Change," presented at the Joint Statistical Meetings, San Francisco, California, (August).

2

Recent Development in Survey Methods

Wong ◆ Ho

Comparing Scoring Systems From Cluster Analysis and Discriminant Analysis Using Random Samples

William Wong and Chih-Chin Ho, Internal Revenue Service

Currently, the Internal Revenue Service (IRS) calculates a scoring formula for each tax return and uses it as one criterion to determine which returns to audit. The IRS periodically updates this formula from a stratified random audit sample. In 1988, such an audit sample was selected. The sample was used to derive a new scoring formula. This score is one of the criteria used to determine whom to audit. In Wong and Ho (2002), we examined the effect of changing sample size on the scoring formula from discriminant analysis. We now extend that work by examining a method of deriving scoring functions using cluster analysis with a variety of distance functions and other options. Those results are compared, and the best results are then compared against those from discriminant analysis. For the evaluation, random subsamples of edited returns are selected, scoring functions developed and applied, and average performances and variances calculated.

We discuss the design of our analysis, our data, and our goals. We then describe our cluster analysis and discriminant analysis approaches. The results of our analysis are presented, with the associated tables in the Appendix. Finally, we highlight our conclusions and future research.

▶ Basic Analysis Framework

We studied one examination class with a sample of 4,356 audited returns. For our study purposes, we selected a fixed set of 100 original variables. For the cluster analysis procedures, we primarily used a fixed subset of 15 of the "best" variables. We also compared using the 15 "best" variables with using the full set of 100 variables in the cluster procedure. In the discriminant analysis procedures, for each random subsample, we used SAS Proc Stepdisc to determine a subset of the 100 variables to use to create our discriminant function. We used a cross-validation approach to evaluate the performances of the scoring formulas.

We start by selecting stratified random subsamples of 2,500 from our 4,356 sample returns using three strata. These subsamples of 2,500 returns serve as the modeling data sets. Thus, for each of these subsamples, we create the cluster analysis and/or discriminant analysis models we wish to compare. Our modeling goal is to maximize the likelihood of identifying returns that exceed a minimum threshold discrepancy between the reported and audited tax amounts. (Due to disclosure sensitivity, the threshold dollar amount is withheld.) We now apply the resulting models on the test data sets of the remaining 1,856 (= 4,356 - 2,500) returns to score each return. Here, a higher score means the model is predicting a higher probability of the return achieving the threshold. The test data set returns are sorted by descending scores, and a cutoff percentage, c, of returns is selected for evaluation. The evaluation statistic, the "hit rate," is defined as the portion of the selected weighted returns achieving the threshold. Cutoff percentages of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, and 75 are analyzed. The cutoff percentage of 100 is also tabulated to provide the average hit rates over the entire test data sets. This procedure is repeated by reselecting 10 to 400 random subsamples, modeling, calculating hit rates for each cutoff percentage, averaging the hit rates over the subsamples, and calculating the variance of each average hit rate.

▶ Cluster Analysis Framework

Motivation: Our approach is to identify returns that exceed the discrepancy threshold, find where they cluster, and score the returns based on their shortest distance to the cluster centroids.

Our cluster analysis proceeds as follows:

- Obtain modeling data set: Select a stratified random subsample of 2,500 of the 4,356 returns.

- Identify those returns that exceeded the threshold tax discrepancy. Typically, this would be around 10 percent of the subsample.

- Create clusters of these "threshold exceeders": Using those returns that exceed the threshold tax discrepancy, run SAS Proc Cluster to create clusters. To create these clusters, we use most of the distance functions available in SAS Proc Cluster: average, centroid, complete, EML, flexible, McQuitty, median, single, and Ward. Distance functions average, centroid, median, and Ward also have "nosquare" options where the distances are not squared.

- Find the centroids of each cluster: For each cluster, obtain the means and standard deviations for each variable.

- Develop raw predicted score functions: For each return exceeding the threshold, calculate its standardized distance to each cluster centroid. Thus, for each variable, calculate the distance between the return value and the cluster mean and divide the result by the cluster standard deviation. Define the distance to each cluster centroid to be the square root of the sum of the squares of the distances across variables. The minimum of these distances across clusters is the raw predicted score. (When a cluster's average standard deviation is zero, the variable mean with a minimum of one is used.)

- Create cluster score adjustment factors: For each cluster, obtain both its average raw predicted score and its average real score, the tax discrepancy among its elements. The adjusted predicted score is then the raw score with a ratio adjustment to even out the cluster-to-cluster differences and prorate to the real score averages.

- Obtain the test data set: The test data set is the remaining 1,856 (= 4,356-2,500) returns.

- Score each test data set return: For each return, calculate raw scores using the same procedure as above and then apply the adjustment factors calculated above. Since a lower score currently means a higher likelihood of exceeding the threshold, the scores need to be inverted. Since the scores are used only in ranking returns, simply reverse the sort.

- Calculate hit rates for each cutoff percentage: After sorting the returns, apply the strata sampling weights to each return and calculate the weighted hit rates for each cutoff percentage.

- Select the next random subsample and repeat the procedure 10 or 400 times.

- Calculate average hit rates and standard deviations over the random subsamples.

▶ Discriminant Analysis Framework

For our study purposes, we selected 100 original variables and used SAS Proc Stepdisc to determine which variables to use for our discriminant function. Thus, the 100 variables are fixed, but the resulting subset of variables changes from sample to sample. The discrimination classification variable used is a zero-one indicator of whether a return exceeds the threshold tax discrepancy.

We start by selecting stratified subsamples of 2,500 from the 4,356 returns using three strata. The weighted samples are first processed through SAS Proc Stepdisc to determine which subset of variables will be used. This is done using two methods: stepwise with p=0.15 and forward discrimination with a maximum of 15 variables. The weighted subsamples are then processed through SAS Proc Discrim using only the variables identified by the Proc Stepdisc procedure. Only parametric discrimination is tested. These weighted subsamples serve as the discrimination modeling data set. The discrimination test data set is the remaining 1,856 (=4,356-2,500) returns. One output of Proc Discrim is the posterior probability of the test return exceeding the threshold. This posterior probability is used as the score. The test data set returns are sorted by descending scores and weighted, and hit rates are calculated for each cutoff percentage. This procedure is repeated over the 400 random subsamples, and average hit rates and their variances are calculated.

▶ Results

For each of the methods, the mean hit rates across the 10 or 400 subsamples were calculated for each percentage cutoff. Along with each mean hit rate, the standard deviation of the mean was also calculated. (The standard deviations calculated were to determine whether the differences between the means are significant and are not sampling error estimates. Those estimates would require correction factors for the large subsampling fractions.)

As indicated above, the basic scoring function for the cluster approach is an adjusted minimum distance between the return and the closest cluster centroid. Originally, the minimum cluster distances were not standardized. We found that standardized distances performed better. We tried various treatments of cluster variable means and variances when they were zero. We settled on replacing the standard deviation with the variable mean with a minimum of 1 when the standard deviation was zero. (This is needed to standardize the distance.)

We tested minimum cluster sizes of 1, 2, 3, 4, 5, 6, 8, 10, and 16. High minimum sizes performed poorly and often did not yield any clusters. The results for minimum cluster sizes of 2 and 4 are given in Appendix Table A. Since the main cutoffs of interest are 1 percent to 10 percent, we summarize the results by averaging the replicate Average Hit Rates (AHR) across these percentages and present them in Table 1. We see that a minimum cluster size of 2 performs better than 4. Furthermore, for distance functions: centroid nosquare, median nosquare, and singular, using a minimum cluster size of 4 did not yield clusters for every subsample.

Table 1. Average Hit Rate (AHR) Means Across Cutoff Percentages 1% to 10%, by Min Cluster Size, Using 10 Replicates of 10 Clusters with 15 Variables

	Min Cluster Size		Best
	4	2	Size
Average	12.96	15.51	2
Average Nosquare	13.20	14.13	2
Centroid	11.25	14.52	2
Centroid Nosquare		11.88	2
Complete	13.21	16.50	2
EML	15.17	18.71	2
Flexible	16.13	18.89	2
McQuitty	13.08	15.61	2
Median	12.04	14.94	2
Median Nosquare		11.41	2
Single		10.44	2
Ward	15.58	18.66	2
Ward Nosquare	17.28	17.60	2

In parallel with deciding minimum cluster size, we needed to determine how many clusters we should form. We tested different numbers of clusters up to 20, but the higher values did not consistently yield clusters. Table 2 compares the results for forming 10, 8, 6, and 4 clusters, using the thirteen distance measures. From the left-hand side of the table, we see that, if we average over the 1-percent to 10-percent cutoffs, the optimum number of clusters varies from 4 to 10. However, the 1-percent cutoff estimates are much larger than the rest. So, if the cutoffs of interest are likely to be in the 2-percent to 10-percent range, then the right-hand side of Table 2 shows that the optimum number of clusters is mainly 6 or 8. Most of the distance functions did reasonably well with 8 clusters; so, we pursued our analysis, using 8 clusters.

Table 2. Average Hit Rate (AHR) Means Across Cutoff Percentages to 10%, by Number of Clusters, Using 10 Replicates with Min Cluster Size of 2 and 15 Variables

	Mean of the AHR Over cutoffs 1% to 10%				Best	Mean of the AHR Over cutoffs 2% to 10%				Best
	Number of Clusters:					Number of Clusters:				
	10	8	6	4		10	8	6	4	
Aver	15.51	15.96	15.25	13.94	8	14.97	15.34	14.21	13.03	8
AvNs	14.13	16.56	15.01	13.94	8	13.72	16.08	14.05	13.04	8
Cent	14.52	14.59	16.08	13.79	6	14.07	14.24	14.87	12.90	6
CntNs	11.88	13.40	14.80	13.61	6	11.39	13.06	14.30	12.48	6
Comp	16.50	17.79	17.92	15.28	6	16.42	17.31	17.05	14.63	8
EML	18.71	18.71	16.14	14.98	10	17.83	17.79	15.84	14.56	10
Flex	18.89	18.55	19.25	18.21	6	18.51	18.24	19.01	17.69	6
McQ	15.61	17.56	17.43	13.54	8	15.37	16.75	16.39	12.89	8
Med	14.94	16.64	16.42	12.63	8	14.60	15.78	15.42	12.16	8
MdNs	11.41	13.71	14.78	13.71	6	11.05	13.12	14.02	12.91	6
Single	10.44	11.31	11.03	11.67	4	10.35	11.12	10.84	10.85	8
Ward	18.66	19.18	16.50	15.00	8	17.76	18.14	16.23	14.58	8
WdNs	17.60	17.94	18.05	18.63	4	17.36	17.74	17.85	17.58	6

Now, would using 100 variables instead of 15 yield better results? The results in Table 3 show that using 100 variables was sharply poorer than using 15. Perhaps the distance formula needs sharper differential weights by variable when there are so many.

Table 3. Average Hit Rate (AHR) Means Across Cutoffs Percentages of 1% to 10%, by Number of Variables, Using 10 Replicates of Forming 8 Clusters with Min Cluster Size of 2

	Using 15 vars	Using 100 vars	Best
Average	15.96	12.95	15
Average Nosquare	16.56	12.66	15
Centroid	14.59	11.92	15
Centroid Nosquare	13.40	11.85	15
Complete	17.79	12.12	15
EML	18.71	11.31	15
Flexible	18.55	10.71	15
McQuitty	17.56	12.91	15
Median	16.64	12.55	15
Median Nosquare	13.71	11.30	15
Single	11.31	8.10	15
Ward	19.18	12.53	15
Ward Nosquare	17.94	12.89	15

Just how stable are these average hits? Was using 10 replicates sufficient? Table 4 shows the mean Average Hit Rate and their ranks when using 10 replicates and 400 replicates. Although there is some difference in the means, their relative rankings changed only slightly. The top four distance functions: EML, flexible, Ward, and Ward nosquare, remained on top. The corresponding original tables and their standard deviations are given in Appendix Tables B and C.

Table 4. Average Hit Rate (AHR) Means Across Cutoffs of 1% to 10% and Their Ranks, by Number of Replicates, Using 8 Clusters with Min Cluster Size of 2 and 15 Variables

	Using 10 reps	Using 400 reps	Rank Using 10 reps	Rank Using 400 reps
Average	15.96	14.77	9	7
Average Nosquare	16.56	14.61	8	8
Centroid	14.59	14.29	10	10
Centroid Nosquare	13.40	13.30	12	11
Complete	17.79	15.99	5	5
EML	18.71	17.49	2	2
Flexible	18.55	17.46	3	4
McQuitty	17.56	15.25	6	6
Median	16.64	14.52	7	9
Median Nosquare	13.71	13.22	11	12
Single	11.31	10.71	13	13
Ward	19.18	17.47	1	3
Ward Nosquare	17.94	17.95	4	1

Finally, back to the original question of which is better, cluster analysis or discriminant analysis? Appendix Table D compares the best of the cluster analysis results with the discriminant analysis results. Discriminant analysis seems to do better, with forward discriminant doing the best. But, are we comparing the same things? Discriminant analysis used the package programs SAS Proc Stepdisc and Proc Discrim. Cluster analysis used the package program SAS Proc Cluster with a self-written scoring program. When writing the program, we noticed that the results were still rather sensitive to the parameters. These parameters need to be analyzed for improvement and robustness. Furthermore, we can interplay one method with the other and sharpen both results. We may also want to experiment with combining the methods with regression.

▶ Conclusions

- High minimum cluster sizes, high numbers of clusters, and high numbers of variables perform poorly. High sizes and numbers of clusters may be difficult to create. Using 8 clusters with a minimum cluster size of 2 and 15 variables appeared to perform best for our data set. Using 100 variables overwhelmed the scoring algorithm.

- Among the cluster methods, EML, flexible, Ward, and Ward nosquare performed the best.

- Using standard discriminant analysis currently performs better than our cluster scoring procedure.

▶ Future Research

In the future we would like to explore methods of enhancing our results, including:

- Combining the methods of cluster analysis, discriminant analysis, and regression for modeling.

- Studying alternative methods calculating and combining the distance functions between the test data set return and each cluster. One enhancement may be to tie the distance function to the function used in creating the clusters.

Finally, we need to test the different methods across years. Specifically, we wish to use one year's data to train the models and apply the results on a different year and then reverse roles. This will help determine the year-to-year deterioration of the models.

▶ Source

Wong, William and Ho, Chih-Chin (2002), "Evaluating the Effect of Sample Size Changes on Scoring System Performance" 2002 Proceedings of the American Statistical Association, Survey Research Methods Section.

▶ Appendix

Table A. Comparing Average % Hit Rates of 13 Clustering Methods by Minimum Cluster Sizes Using 10 Replicates of Forming 10 Clusters with 15 Variables

Cut-off Pct	Aver	Aver Nosq	Cent	Cent Nosq	Comp	EML	Flex	McQ	Med	Med Nosq	Sing	Ward	Ward Nosq
Using a minimum cluster size of 4:													
1	14.84	16.00	12.18	**	14.41	18.49	21.36	14.50	16.67	**	**	21.41	22.97
2	13.16	13.32	11.10	**	12.21	17.59	18.27	14.45	11.74	**	**	18.15	18.90
3	13.10	14.07	11.00	**	13.83	17.11	16.75	14.52	11.26	**	**	17.89	18.66
4	13.66	13.95	11.79	**	14.54	15.49	17.06	13.58	11.29	**	**	16.16	18.09
5	12.75	13.29	11.85	**	13.57	15.13	16.60	13.04	11.82	**	**	14.66	17.34
6	12.64	12.94	10.98	**	13.11	14.19	15.37	12.91	11.75	**	**	14.05	16.79
7	12.50	12.32	11.29	**	12.80	13.80	14.45	12.63	11.97	**	**	13.77	15.95
8	12.36	12.07	11.00	**	12.48	13.56	14.20	12.14	11.30	**	**	13.24	15.12
9	12.43	12.13	10.79	**	12.74	13.36	13.75	11.74	11.55	**	**	13.17	14.42
10	12.19	11.89	10.52	**	12.42	13.01	13.45	11.29	11.05	**	**	13.33	14.57
15	10.80	11.01	9.52	**	11.61	12.36	12.11	10.72	10.41	**	**	11.87	12.66
20	10.00	10.36	9.19	**	10.80	11.72	11.95	10.19	9.85	**	**	12.13	11.93
25	10.12	10.12	9.23	**	10.60	11.19	11.70	9.93	9.80	**	**	11.38	11.47
30	9.95	10.06	8.99	**	10.38	11.33	11.18	9.86	9.78	**	**	11.47	11.00
35	10.00	9.92	8.85	**	10.04	11.17	11.26	9.56	9.81	**	**	11.27	11.00
40	9.71	9.74	8.94	**	10.04	11.15	10.97	9.74	9.70	**	**	11.01	10.84
45	9.68	9.67	9.08	**	9.94	10.92	10.87	9.80	9.72	**	**	10.85	10.90
50	9.70	9.69	9.34	**	9.81	10.68	10.49	9.72	9.74	**	**	10.59	10.91
75	9.64	9.58	9.37	**	9.97	10.25	10.42	9.64	9.70	**	**	10.18	10.45
100	11.77	11.77	11.77	**	11.77	11.77	11.77	11.77	11.77	**	**	11.77	11.77
Using a minimum cluster size of 2:													
1	20.38	17.85	18.52	16.32	17.23	26.59	22.28	17.79	17.99	14.60	11.24	26.84	19.79
2	18.24	17.13	17.34	14.48	17.44	23.47	21.37	17.12	17.31	12.20	11.14	23.12	20.73
3	16.79	15.44	15.58	12.51	17.12	20.37	19.66	17.47	15.91	11.54	11.88	21.55	19.69
4	16.38	13.92	15.94	11.27	17.53	18.69	20.17	16.09	14.11	11.39	10.49	19.47	18.64
5	15.62	13.50	14.49	11.23	16.68	17.79	19.83	15.67	14.78	10.96	9.99	17.69	17.29
6	14.20	13.08	13.25	10.77	16.74	16.83	18.47	15.21	14.55	11.50	10.21	17.05	16.60
7	13.75	12.49	12.83	10.70	16.31	16.37	17.80	14.94	13.92	10.95	10.19	15.99	16.40
8	12.92	12.64	12.55	10.51	15.64	16.07	16.93	14.31	13.62	10.54	9.94	15.24	16.02
9	13.24	12.72	12.43	10.43	15.17	15.93	16.45	13.81	13.76	10.27	9.75	15.00	15.78
10	13.60	12.54	12.24	10.63	15.11	14.98	15.90	13.71	13.41	10.11	9.58	14.69	15.03
15	12.62	12.04	11.71	9.88	13.53	13.92	15.27	13.16	12.56	9.65	9.05	14.42	13.78
20	11.72	11.13	10.49	9.60	13.33	13.28	14.26	12.23	11.59	9.25	9.02	13.62	13.09
25	11.44	11.03	10.49	9.87	12.79	12.35	13.65	11.44	10.98	9.73	9.08	12.66	12.87
30	11.30	10.90	10.22	9.97	12.29	12.08	13.07	11.31	10.92	9.73	8.74	12.25	12.56
35	11.21	10.82	10.19	9.58	11.84	11.66	12.48	11.11	10.83	9.52	8.54	11.78	12.33
40	10.96	10.44	10.08	9.37	11.64	11.48	12.05	11.05	10.69	9.28	8.68	11.71	12.02
45	10.60	10.12	9.74	9.11	11.50	11.24	11.72	10.83	10.37	9.25	8.80	11.47	11.68
50	10.31	9.94	9.64	9.10	11.35	11.03	11.58	10.51	10.15	9.07	8.85	11.18	11.41
75	9.93	9.79	9.54	9.42	10.49	10.52	10.70	10.02	9.94	9.40	9.42	10.53	10.70
100	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77

Note: ** Ten clusters with cluster size >= 4 could not be formed for every replicate with this clustering method.

Table B. Comparing Average % Hit Rates of 13 Clustering Methods by Number of Replicates When Forming 8 Clusters with 15 Variables and a Minimum Cluster Size of 2

Cut-off Pct	Aver	Aver Nosq	Cent	Cent Nosq	Comp	EML	Flex	McQ	Med	Med Nosq	Sing	Ward	Ward Nosq
Using 10 Replicates:													
1	21.54	20.85	17.77	16.54	22.16	27.02	21.35	24.88	24.32	19.07	13.07	28.47	19.80
2	20.98	20.50	17.88	15.64	19.08	23.24	20.47	22.02	21.04	14.65	13.07	25.14	20.88
3	17.66	18.08	16.72	13.95	19.09	20.28	18.24	17.92	17.59	14.87	12.06	21.26	20.22
4	16.44	17.89	15.25	13.69	18.52	19.51	18.87	17.52	16.77	14.40	11.24	19.20	19.91
5	15.05	16.18	13.63	13.34	17.15	17.67	17.60	16.29	16.06	13.21	11.25	17.86	18.33
6	14.12	15.50	13.52	12.76	17.27	16.68	18.00	15.77	14.66	12.90	10.72	17.09	17.03
7	14.12	14.78	13.50	12.40	16.80	16.24	18.41	16.00	14.81	12.12	10.29	16.41	16.47
8	13.52	14.27	12.76	12.13	16.29	15.66	18.07	15.68	14.04	12.28	10.58	15.91	16.11
9	13.17	14.06	12.43	11.70	15.96	15.62	17.30	15.14	13.60	12.13	10.60	15.52	15.61
10	12.99	13.45	12.49	11.89	15.60	15.20	17.18	14.42	13.47	11.46	10.26	14.92	15.06
15	12.55	12.89	12.00	11.03	14.18	14.23	15.42	13.30	12.36	11.12	9.35	14.37	14.41
20	12.20	12.11	11.50	10.44	13.51	13.12	14.44	12.92	11.88	10.56	9.70	13.67	13.68
25	11.56	11.67	11.19	10.35	13.14	12.37	13.43	12.31	11.57	10.46	9.65	12.79	13.21
30	11.39	11.65	11.17	10.46	12.74	12.07	13.04	11.74	11.22	10.55	9.37	12.29	13.04
35	11.24	11.33	10.98	10.24	12.39	11.87	12.66	11.58	11.14	10.35	9.20	12.09	12.52
40	10.94	11.12	10.77	10.21	11.99	11.54	12.45	11.46	11.04	10.10	8.90	11.83	12.09
45	10.65	10.83	10.43	9.97	11.76	11.25	12.39	11.16	10.77	9.72	9.00	11.48	11.94
50	10.48	10.41	10.33	9.73	11.44	11.09	12.06	10.86	10.65	9.66	9.05	11.34	11.45
75	10.06	10.09	9.99	9.69	10.52	10.41	10.84	10.24	10.29	9.53	9.47	10.27	10.70
100	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77
Using 400 Replicates:													
1	18.68	18.79	17.70	15.64	20.43	23.78	21.96	19.53	18.39	16.21	13.25	23.64	23.36
2	17.53	17.18	16.72	15.30	18.47	20.95	19.92	18.06	17.18	14.97	12.38	20.95	20.67
3	16.24	15.89	15.61	14.49	17.13	18.92	18.89	16.39	15.79	14.27	11.55	19.06	19.30
4	15.17	14.99	14.63	13.72	16.38	17.75	17.88	15.54	14.79	13.48	10.79	17.65	18.23
5	14.42	14.23	13.91	13.13	15.64	16.78	17.03	14.80	14.10	13.03	10.31	16.77	17.52
6	13.84	13.70	13.46	12.68	15.19	16.19	16.48	14.24	13.65	12.54	9.97	16.13	16.87
7	13.41	13.30	13.12	12.31	14.71	15.72	16.08	13.94	13.20	12.27	9.76	15.71	16.47
8	13.11	12.95	12.86	12.07	14.30	15.32	15.74	13.62	12.98	11.96	9.74	15.27	15.98
9	12.78	12.64	12.51	11.87	13.95	14.95	15.44	13.32	12.71	11.80	9.65	14.89	15.74
10	12.51	12.41	12.34	11.74	13.68	14.58	15.16	13.06	12.45	11.67	9.65	14.67	15.40
15	11.86	11.78	11.78	11.14	12.74	13.55	14.15	12.30	11.87	11.15	9.13	13.65	14.37
20	11.49	11.43	11.35	10.64	12.14	12.86	13.45	11.78	11.44	10.68	9.02	12.98	13.63
25	11.14	11.08	11.02	10.36	11.71	12.38	12.89	11.48	11.10	10.46	9.06	12.44	13.08
30	10.86	10.80	10.77	10.16	11.40	11.97	12.46	11.21	10.83	10.26	8.90	12.02	12.56
35	10.68	10.61	10.58	10.01	11.10	11.63	12.09	10.96	10.68	10.16	8.76	11.68	12.18
40	10.53	10.51	10.44	9.89	10.97	11.39	11.80	10.79	10.55	10.03	8.71	11.45	11.87
45	10.36	10.31	10.26	9.76	10.79	11.22	11.59	10.65	10.39	9.85	8.76	11.27	11.61
50	10.14	10.09	10.06	9.62	10.56	11.04	11.38	10.43	10.14	9.71	8.78	11.08	11.41
75	9.84	9.84	9.83	9.63	10.01	10.28	10.59	9.99	9.88	9.67	9.26	10.32	10.62
100	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72

Table C. Comparing Std Dev (Average % Hit Rates) of 13 Clustering Methods by Number of Replicates When Forming 8 Clusters with 15 Variables and a Minimum Cluster Size of 2

Cut-off Pct	Aver	Aver Nosq	Cent	Cent Nosq	Comp	EML	Flex	McQ	Med	Med Nosq	Sing	Ward	Ward Nosq
Using 10 Replicates:													
1	2.12	3.28	2.35	1.44	3.70	2.51	2.18	4.23	3.94	2.28	2.41	2.64	2.23
2	2.16	2.14	2.09	1.35	1.18	2.50	1.20	2.30	2.53	1.17	1.11	2.34	1.64
3	2.15	1.53	1.31	1.54	1.29	1.39	1.00	1.50	1.83	1.31	0.94	1.86	1.32
4	1.68	0.97	0.98	1.26	1.12	1.18	1.01	1.54	1.72	0.87	0.84	1.29	1.41
5	1.36	1.05	0.87	1.15	0.95	1.17	0.96	1.39	1.37	0.94	0.71	1.07	1.01
6	1.25	0.76	0.74	0.91	0.87	1.00	0.85	1.51	1.12	0.95	0.74	0.81	0.86
7	1.10	0.78	0.77	0.79	0.79	0.79	0.82	1.30	1.02	0.69	0.60	0.99	0.78
8	0.96	0.67	0.54	0.74	0.90	0.78	0.62	1.10	0.97	0.77	0.41	0.85	0.63
9	0.93	0.48	0.47	0.71	0.89	0.84	0.69	1.07	0.98	0.67	0.44	0.67	0.53
10	0.89	0.51	0.46	0.68	0.87	0.71	0.57	1.05	0.83	0.55	0.34	0.71	0.57
15	0.87	0.39	0.71	0.59	0.67	0.50	0.51	0.85	0.64	0.46	0.38	0.58	0.49
20	0.61	0.31	0.52	0.54	0.45	0.55	0.37	0.69	0.66	0.48	0.40	0.70	0.45
25	0.47	0.30	0.42	0.39	0.55	0.51	0.21	0.56	0.59	0.37	0.33	0.57	0.36
30	0.56	0.29	0.44	0.48	0.52	0.48	0.22	0.55	0.54	0.30	0.36	0.46	0.33
35	0.48	0.26	0.38	0.40	0.44	0.43	0.30	0.52	0.49	0.24	0.31	0.34	0.25
40	0.33	0.25	0.36	0.28	0.34	0.38	0.29	0.37	0.37	0.26	0.25	0.36	0.25
45	0.29	0.19	0.30	0.29	0.34	0.35	0.30	0.36	0.36	0.21	0.22	0.36	0.23
50	0.27	0.17	0.25	0.29	0.36	0.30	0.35	0.29	0.30	0.19	0.24	0.30	0.23
75	0.20	0.17	0.22	0.23	0.27	0.24	0.23	0.22	0.24	0.17	0.20	0.25	0.16
100	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Using 400 Replicates:													
1	0.41	0.40	0.41	0.39	0.43	0.43	0.46	0.45	0.41	0.41	0.34	0.42	0.51
2	0.28	0.27	0.29	0.27	0.30	0.31	0.31	0.30	0.29	0.27	0.23	0.32	0.35
3	0.23	0.22	0.22	0.22	0.23	0.26	0.26	0.23	0.23	0.23	0.18	0.25	0.27
4	0.20	0.18	0.18	0.19	0.20	0.22	0.22	0.20	0.19	0.19	0.15	0.21	0.24
5	0.18	0.17	0.17	0.17	0.17	0.19	0.20	0.18	0.17	0.16	0.13	0.18	0.21
6	0.16	0.16	0.15	0.16	0.17	0.17	0.18	0.17	0.16	0.15	0.12	0.17	0.19
7	0.15	0.15	0.15	0.14	0.16	0.16	0.17	0.15	0.15	0.14	0.11	0.15	0.17
8	0.14	0.14	0.13	0.13	0.15	0.15	0.16	0.14	0.14	0.13	0.10	0.14	0.15
9	0.13	0.12	0.13	0.12	0.14	0.14	0.15	0.14	0.13	0.12	0.09	0.13	0.15
10	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.13	0.12	0.11	0.09	0.12	0.14
15	0.10	0.09	0.10	0.09	0.11	0.10	0.11	0.10	0.10	0.09	0.07	0.10	0.10
20	0.08	0.08	0.08	0.08	0.09	0.09	0.10	0.09	0.08	0.08	0.06	0.09	0.09
25	0.07	0.07	0.07	0.07	0.08	0.08	0.09	0.08	0.07	0.07	0.05	0.08	0.08
30	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.06	0.05	0.07	0.07
35	0.05	0.05	0.05	0.06	0.06	0.06	0.07	0.06	0.06	0.05	0.05	0.06	0.07
40	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05	0.04	0.05	0.06
45	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.04	0.05	0.05
50	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.05	0.05
75	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.04
100	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Table D. Comparing Average % Hit Rates(AHR) & SD(AHR) Among Select Cluster & Discriminant Methods Using 400 Replicates Where Clustering Is Done with 8 Clusters, 15 Variables, and a Minimum Cluster Size of 2

Cut-off Pct	Average % Hit Rate (AHR)						Standard Deviation (AHR)					
	Clustering				Discriminant		Clustering				Discriminant	
	EML	Flex	Ward	Ward Nosq	Step-wise	For-ward	EML	Flex	Ward	Ward Nosq	Step-wise	For-ward
1	23.78	21.96	23.64	23.36	27.03	27.65	0.43	0.46	0.42	0.51	0.49	0.49
2	20.95	19.92	20.95	20.67	27.47	28.85	0.31	0.31	0.32	0.35	0.33	0.34
3	18.92	18.89	19.06	19.30	27.29	28.42	0.26	0.26	0.25	0.27	0.28	0.28
4	17.75	17.88	17.65	18.23	26.70	27.44	0.22	0.22	0.21	0.24	0.24	0.23
5	16.78	17.03	16.77	17.52	26.06	26.56	0.19	0.20	0.18	0.21	0.21	0.21
6	16.19	16.48	16.13	16.87	25.38	25.79	0.17	0.18	0.17	0.19	0.19	0.18
7	15.72	16.08	15.71	16.47	24.85	25.17	0.16	0.17	0.15	0.17	0.17	0.17
8	15.32	15.74	15.27	15.98	24.23	24.63	0.15	0.16	0.14	0.15	0.16	0.16
9	14.95	15.44	14.89	15.74	23.76	24.02	0.14	0.15	0.13	0.15	0.15	0.14
10	14.58	15.16	14.67	15.40	23.29	23.49	0.13	0.14	0.12	0.14	0.14	0.13
15	13.55	14.15	13.65	14.37	21.29	21.38	0.10	0.11	0.10	0.10	0.11	0.11
20	12.86	13.45	12.98	13.63	19.68	19.86	0.09	0.10	0.09	0.09	0.09	0.09
25	12.38	12.89	12.44	13.08	18.69	18.71	0.08	0.09	0.08	0.08	0.08	0.08
30	11.97	12.46	12.02	12.56	17.80	17.79	0.07	0.07	0.07	0.07	0.07	0.07
35	11.63	12.09	11.68	12.18	17.09	17.05	0.06	0.07	0.06	0.07	0.06	0.06
40	11.39	11.80	11.45	11.87	16.45	16.42	0.06	0.06	0.05	0.06	0.06	0.05
45	11.22	11.59	11.27	11.61	15.89	15.90	0.05	0.06	0.05	0.05	0.05	0.05
50	11.04	11.38	11.08	11.41	15.40	15.42	0.05	0.05	0.05	0.05	0.05	0.05
75	10.28	10.59	10.32	10.62	13.34	13.41	0.04	0.04	0.03	0.04	0.04	0.04
100	11.72	11.72	11.72	11.72	11.72	11.72	0.03	0.03	0.03	0.03	0.03	0.03

3

New Developments in Tax Statistics and Administrative Records

Sailer ◆ Gurka ◆ Holden
Legel ◆ Bennett ◆ Parisi
Kilss ◆ Jordan
Dixon
Petska ◆ Kilss

Accumulation and Distributions of Retirement Assets, 1996-2000: Results From a Matched File of Tax Returns and Information Returns

Peter Sailer and Kurt S. Gurka, Internal Revenue Service,
and Sarah Holden, Investment Company Institute

Deductions for contributions to Individual Retirement Arrangements (IRA's) appeared on Form 1040 for the first time for Tax Year 1975, and the 1975 version of the annual report, *Statistics of Income—Individual Income Tax Returns,* duly recorded the number of returns with an entry on that line—1.2 million—and the amount deducted—about $1.4 billion (Figure 1). Twenty-five years and many tax law changes later, the 2000 Individual Income Tax Returns report still tabulated the entries on this line, amounting to 3.5 million returns and $7.5 billion in deductible traditional IRA contributions.

However important these statistics have been to the analysis of IRA's over the years, they have not told the full story. For example, during Tax Year 2000, in addition to the $7.5 billion in tax-deductible contributions to traditional IRA plans, $2.5 billion in nondeductible contributions were also made to such plans (Figure 2). Furthermore, other types of IRA received $26.3 billion in contributions. However, much more importantly, $225.6 billion of assets were rolled over into IRA's from other qualified pension plans and tax-sheltered annuities. These three statistics were taken from Forms 5498 filed with the Internal Revenue Service by IRA trustees.[1] Form 5498 also shows the total fair market value (FMV) of assets held in IRA's. At the end of Tax Year 2000, the total value of IRA assets stood at $2.6 trillion.

In the following paper, the authors use never-before-released IRS data1 from Form 5498, along with household survey and other information, to highlight key demographic and financial characteristics of traditional IRA owners and their traditional IRA assets. Historical trends will be noted. In addition, again using the matched file of tax returns and Forms W-2, some summary statistics on 401(k) plans are presented.

▶ Assets Held in IRA's

Types of IRAs. The predominant type of IRA is the traditional IRA, which was created with the Employee Retirement Income Security Act (ERISA) of 1974. Indeed, SOI estimates indicate that about 92 percent of all IRA assets were held in traditional IRA's at yearend 2000 (Figure 2). Roth IRA's (created in the Taxpayer Relief Act of 1997) represented about 3 percent of all IRA assets, while employer-sponsored IRA's (SEP IRA's—created in the Revenue Act of 1978, SAR-SEP IRA's—created in the Tax Reform Act of 1986, and SIMPLE IRA's—created in the Small Business Job Protection Act of 1996) held about 6 percent of the total. Education IRA's (created in the Taxpayer Relief Act of 1997), which are now known as Coverdell Education Savings Accounts (ESA's), accounted for a negligible share of the total.[2]

Types of Assets. Figure 3, based in part on IRS statistics and in part on surveys by the Investment Company Institute (ICI), shows that at yearend 2002, IRA assets amounted to an estimated $2.3 trillion, compared with $637 billion at yearend 1990. With more than half of IRA assets invested in equity securities, the effect of the stock market can also be seen in recent years. Indeed, IRA assets reached nearly $2.7 trillion at yearend 1999 during the bull market in equities, then began declining. IRA assets are invested in a variety of financial institutions. At yearend 2002, about 46 percent of all IRA assets were invested in mutual funds, another 34 percent were in non-mutual fund securities held through brokerage accounts, another 11 percent were held in bank and thrift deposits, and the remaining 9 percent were held in annuities at life insurance companies.[3]

▶ Demographic Composition and IRA Balances of Traditional IRA Owners

Data from Tax Returns and Information Returns

Age Distribution, by Taxpayer. The SOI data, which are based on a weighted sample of tax-return information,[4] allow analysis of IRA-owning taxpayers at yearend 1999 by age, marital status on the tax return,

gender, and income. Traditional IRA owners are predominantly middle-aged. Twenty percent of the 36.6 million taxpayers with traditional IRA's in 1999 were 35 to 44 years of age, and another 25 percent were age 45 to 54 (Figure 4). About 14 percent of taxpayers owning IRA's in 1999 were age 70 or older, which places them in the age group that must take required minimum distributions from their accounts.

Marital Status and Gender, by Taxpayer. The majority of traditional IRA-owning taxpayers are married filing joint returns, and half of traditional IRA- owning taxpayers are male. Among the 36.6 million traditional IRA-owning taxpayers at yearend 1999, nearly three-quarters were married (Figure 5).[5] Half of traditional IRA-owning taxpayers were married with both spouses holding traditional IRA's. Single women accounted for 16 percent of traditional IRA-owning taxpayers, and single men accounted for 10 percent.

Traditional IRA Balances, by Age of Taxpayer. There is a wide range of traditional IRA balances held by taxpayers around an average of $66,179 at yearend 1999.[6] Because older taxpayers have had more time to work and accumulate IRA assets, either from rollover at job change or from contributions over time, older taxpayers tend to have higher traditional IRA balances. The average traditional IRA balance held by taxpayers 25 to 34 years old was $12,435 at yearend 1999, while the average balance peaks among taxpayers aged 65 to 69 at $112,588, even though they may take distributions without penalty (Figure 4). Among taxpayers aged 70 and older, the average traditional IRA balance falls, perhaps because those individuals are taking withdrawals to fund retirement at least at the level of the required minimum distribution.

Traditional IRA Balances, by Marital Status and Gender of Taxpayer. There is a wide range of average traditional IRA balances by marital status and gender among taxpayers analyzed in the SOI data. Although it is difficult to interpret the significance of average account balances for a snapshot of one period in time, it is interesting to note that single taxpayers have similar average account balances regardless of gender (Figure 5). On the other hand, among married taxpayers, the husbands' average IRA assets were higher than the wives'. The traditional IRA assets at yearend 1999 represent an accumulation of activity starting possibly as far back as 1974 (when traditional IRAs were created). The lower average among wives may be driven by women's typically discontinuous work histories and therefore lower rollover amounts available to go into an IRA.[7] In addition, regulations restricting tax-deductible spousal contributions may also have damped wives' IRA assets. To gain a better understanding of the differential in average IRA assets between husbands and wives, data for the same individuals would need to be tracked for several years—monitoring contributions, rollovers, and workforce participation (and pension coverage therein). Unfortunately, such insight, while planned, will not be available for several years.

▶ Traditional IRA's: Comparison of SOI Data and Household Survey Information

Although it is difficult to match up household and taxpayer information,[8] this section makes some broad comparisons between the SOI data and household surveys conducted by ICI and the Federal Reserve Board. The IRS SOI tax return and information return data, which are based on a weighted sample of returns, find similar demographic characteristics for the typical (median) traditional IRA owner as these household surveys find. Furthermore, similar to household survey information, the median traditional IRA balance among tax returns with traditional IRA owners was $27,181 at yearend 1999 (Figure 6).

Age and Marital Status of Typical Traditional IRA Owner. The typical traditional IRA owner is about 50 years old. For example, ICI's June 1999 survey finds a median age among traditional IRA-owning households of 49 years, and ICI's June 2000 survey finds a median age of 53 years (Figure 6). Similarly, ICI tabulations[9] of the Federal Reserve Board's Survey of Consumer Finances (SCF) data indicate that the median age of IRA-owning households was 51 in both the 1998 and 2001 surveys. Likewise, the median taxpayer owning traditional IRA's in the SOI data was 53 years old in 1999. The typical traditional IRA owner is married. In the household surveys and the SOI tax return data, about two-thirds of households with traditional IRA's are married.

▶ Demographic Composition and Contribution Activity of 401(k) Participants

SOI Data Based on IRS Tax Returns and Information Returns

Although Section 401(k) of the Internal Revenue Code was created in the Revenue Act of 1978, clarification of the regulations did not occur until 1981. After that slow start, 401(k) plans have grown rapidly throughout the 1990's. At yearend 2002—401(k) plans held $1.54 trillion in assets (Figure 7). The key provision of 401(k) plans is the ability to defer salary by making before-tax contributions (deferrals) to an account maintained in the given participant's name. In most instances, the participant directs the investment of the account assets, which grow tax-free until they are withdrawn. In many cases, the plan sponsor may make a matching contribution (for example, contributing 50 cents for every dollar the participant contributes up to 6 percent of salary).[10] Contributions-by-plan participants depend on a variety of factors, including the regulatory framework under which 401(k) plans operate, personal participant characteristics, and plan design features.[11] Using IRS W-2 form information, a glimpse at elective deferrals by taxpayers into 401(k) plans is possible.

401(k) Elective Deferrals. W-2 Form information indicates that elective deferrals by 401(k) plan participants rose steadily from 1996 through 2002, from $61 billion to nearly $105 billion (Figure 7).[12] In addition, the number of taxpayers with deferrals and the average deferral amount also increased over the late 1990's. Indeed, average deferrals rose from $2,660 in 1996 to $3,408 in 2000.

Age of 401(k) Participants. The average and median age of taxpayer with 401(k) elective deferrals was 42 in 1999.[13] The bulk of deferrals in 1999 were made by taxpayers in their thirties, forties, or fifties. Average deferrals tend to rise with age through the age group in their fifties, and decline a bit among taxpayers in their sixties and older.

▶ Conclusions and Future Research

With nearly $4.0 trillion invested in IRA's and 401(k) plans at yearend 2003, these retirement savings vehicles represent a significant component of Americans' financial security. Taxpayers holding IRAs and 401(k) accounts cover a wide range of ages and incomes. This paper provided a glimpse at the rich detail available from the SOI sample of tax returns and information returns (focusing in detail on 1999). The results of this analysis of SOI data are encouraging. The typical IRA- owning taxpayer represented in the SOI data appears to be similar in basic demographic and financial characteristics to the typical IRA-owning household found in household surveys conducted by ICI and the Federal Reserve Board. The SOI 401(k) participant information in aggregate corresponds well to Form 5500 results, and the preliminary age analysis is similar to EBRI/ICI results.

Future research would extend both the IRA and 401(k) detailed analyses to more years. In addition, future data would analyze the taxpayers by type of IRA. Furthermore, longitudinal analysis tracking the behaviors of IRA owners and 401(k) participants over time should also be explored.

▶ Acknowledgment

The authors would like to extend their thanks to Marianne Cooley of the IRS Computing Center in Detroit, Michigan, for many years of stewardship of the SOI Information Returns data base.

▶ References

Federal Reserve Board. *Survey of Consumer Finances*. Available at: http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html

Holden, Sarah and VanDerhei, Jack, "Contribution Behavior of 401(k) Plan Participants." *ICI Perspective*, Vol. 7, No. 4, and *EBRI Issue Brief*, No. 238. Washington, DC, Investment Company Institute and Employee Benefit Research Institute, October 2001.

Holden, Sarah and VanDerhei, Jack, "401(k) Plan Asset Allocation, Account Balances, and Loan Activity in 1999," *ICI Perspective*, Vol. 7, No. 1, and *EBRI Issue Brief*, No. 230, Washington, DC, Investment Company Institute, January 2001, and Employee Benefit Research Institute, February 2001.

Investment Company Institute, "IRA Ownership in 2003," *Fundamentals*, Vol. 12, No. 3, Washington, DC, Investment Company Institute, September 2003.

———, "Mutual Funds and the U.S. Retirement Market in 2002," *ICI Fundamentals*, Vol. 12, No. 1, Washington, DC, Investment Company Institute, June 2003.

———, "IRA Ownership in 2002," *Fundamentals*, Vol. 11, No. 3, Washington, DC, Investment Company Institute, September 2002.

———, "IRA Ownership in 2001," *Fundamentals*, Vol. 10, No. 3, Washington, DC, Investment Company Institute, September 2001.

———, "IRA Ownership in 2000," *Fundamentals*, Vol. 9, No. 5, Washington, DC, Investment Company Institute, October 2000.

———, "IRA Ownership in 1999," *Fundamentals*, Vol. 8, No. 6, Washington, DC, Investment Company Institute, December 1999.

Sailer, Peter J.; Weber, Michael E.; Gurka, Kurt S., "Are Taxpayers Increasing the Buildup of Retirement Assets? Preliminary Results from a Matched File of Tax Year 1999 Tax Returns and Information Returns," p*aper presented at National Tax Association 95th Annual Conference* in Orlando, FL, November 15, 2002.

U.S. Department of Labor, Employee Benefits Security Administration, "Women and Retirement Savings." Available at: http://www.dol.gov/ebsa/publications/women.html

U.S. Internal Revenue Service, *Publication 590 Individual Retirement Arrangements (IRA's)*. Washington, DC, U.S. Department of the Treasury, Internal Revenue Service, 1999 and 2002. Available at: www.irs.gov/pub/irs-99/p590.pdf (1999) and www.irs.gov/pub/irs-pdf/p590.pdf

U.S. Internal Revenue Service, Statistics of Income, *2000 Individual Income Tax Returns,* Washington, DC, March 2003.

U.S. Internal Revenue Service, Statistics of Income, *1975 Individual Income Tax Returns,* Washington, DC, 1978.

U.S. Internal Revenue Service, Statistics of Income, *Individual Income Tax Returns, Publication 1304*, Washington, DC, various years.

▶ Footnotes

1 All SOI data are based on a stratified weighted sample of individual income tax returns with matching information returns. See Sailer, Weber, and Gurka (November 2002) for SOI data estimation methodology.

2 The estimate for education IRA's, or Coverdell ESA's, is underestimated in Figure 2, as nonfiling dependents are not included in the estimation.

3 Figure 3 reports aggregate IRA assets; it does not indicate what individual IRA owners may be holding. An ICI survey of IRA owning households in mid-2003 reports the incidence of the different types of financial assets held in IRA's (see Investment Company Institute (September 2003)).

4 See text footnote 1.

5 A small number of married-filing-separate returns are included as single. Some of these taxpayers are in fact in the middle of separation or divorce proceedings and are not, in fact, living together.

6 Among the 36.6 million taxpayers with traditional IRAs at yearend 1999, some 20.6 percent had traditional IRA balances of less than $5,000; 28.6 percent had traditional IRA balances between $5,000 and $20,000; 17.8 percent had balances between

$20,000 and $40,000; 18.7 percent had balances between $40,000 and $100,000; and 14.3 percent had traditional IRA balances of $100,000 or more.

[7] The U.S. Department of Labor's "Women and Retirement Savings" notes that women are more likely to work in part-time jobs that do not qualify for pension coverage or to have lapses in pension coverage because of interruptions in their careers to take care of family members.

[8] Exact comparison is not possible for several reasons including: (1) not all households file tax returns, (2) household units do not always correspond to tax return units, for example, a household with unmarried partners will appear as one household in a household survey, but as two single tax returns, (3) household surveys rely on self-reported information, which can suffer from participant recall, while the tax return information is unaudited tax return information, which may contain reporting errors, (4) the definition of income may vary across data sources, and (5) the timing of the surveys/returns varies.

[9] Special thanks to Michael Bogdan at ICI for tabulating the SCF data.

[10] See Investment Company Institute (June 2003).

[11] For a comprehensive study of 401(k) participant contribution activity, see Holden and VanDerhei (October 2001).

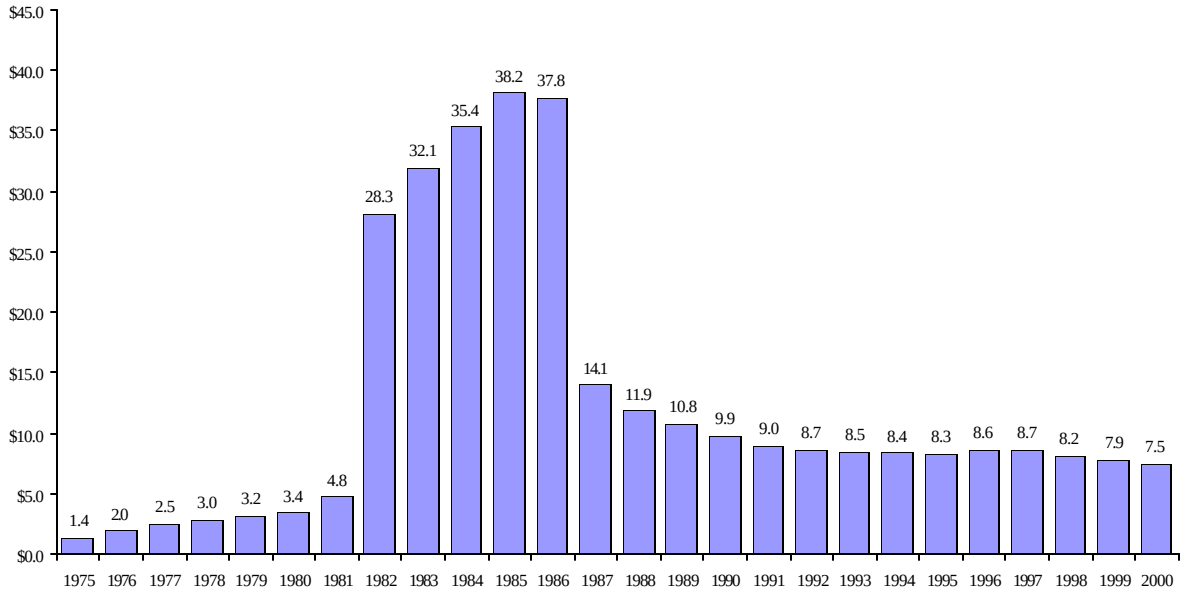
[12] The new tabulations from Forms W-2 produce similar estimates for aggregate deferrals as U.S. Department of Labor tabulations of employee contributions from Forms 5500 (Figure 7).

[13] About 13 percent of these taxpayers were in their twenties; 30 percent were in their thirties; 31 percent were in their forties; 20 percent were in their fifties; and 5 percent were in their sixties. This age distribution is essentially the same as the age distribution of active 401(k) plan participants in the collaborative data base maintained by the Employee Benefit Research Institute (EBRI) and ICI, known as the EBRI/ICI Participant-Directed Retirement Plan Data Collection Project (see Holden and VanDerhei (January/February 2001)).

▶ Note

The views in this paper are those of the authors and do not reflect those of the Investment Company Institute or its members, nor are they the official positions of the Internal Revenue Service. Any errors are solely the responsibility of the authors.

Figure 1
Deductible IRA Contributions to Traditional IRAs,* 1975-2000
(billions of dollars)



*Deductible IRA contributions reported on individual income tax returns (Form 1040).
Source: IRS, Statistics of Income Division, Individual Income Tax Returns, Publication 1304, various years

Figure 2. Individual Retirement Arrangement (IRA) Plans by type, Tax Year 2000

	Beginning of Year FMV[1]		Total Contributions[1]		Contributions Deducted[2]	
	Number of Taxpayers	Amount ($thousands)	Number of Taxpayers	Amount ($thousands)	Number of Taxpayers	Amount ($thousands)
	(1)	(2)	(3)	(4)	(5)	(6)
Total	43,063,085	2,651,203,109	15,124,569	36,331,114	5,397,588	12,207,520
Traditional IRA Plans	36,619,402	2,422,819,105	5,716,919	9,998,892	4,583,252	7,477,074
SEP Plans	3,146,153	142,873,671	1,735,666	10,068,405	683,861	4,198,700
SIMPLE Plans	1,177,084	9,126,960	1,489,333	4,675,650	130,475	531,746
Roth IRA Plans	7,031,194	76,242,001	6,812,129	11,509,407	n/a	n/a
Education IRA Plans[3]	182,000	141,372	155,253	78,761	n/a	n/a

Figure 2. IRA plans by type, Tax Year 2000 (continued)

	Rollovers[1]		Roth Conversions[1]		Withdrawals[4]	
	Number of Taxpayers	Amount ($thousands)	Number of Taxpayers	Amount ($thousands)	Number of Taxpayers	Amount ($thousands)
	(7)	(8)	(9)	(10)	(11)	(12)
Total	4,079,126	225,595,813	0	0	8,621,056	103,915,860
Traditional IRA Plans	4,079,126	225,595,813	282,387	-3,181,178	7,818,268	94,636,704
SEP Plans	n/a	n/a	n/a	n/a	276,861	3,822,337
SIMPLE Plans	n/a	n/a	n/a	n/a	158,713	822,171
Roth IRA Plans	n/a	n/a	282,387	3,181,178	365,186	4,632,735
Education IRA Plans[3]	n/a	n/a	n/a	n/a	2,028	1,913

Figure 2. IRA plans by type, Tax Year 2000 (concluded)

	Other Changes[5] Amount ($thousands) (13)	End of Year Fair Market Value (FMV)[1] Number of Taxpayers (14)	Amount ($thousands) (15)
Total	-180,341,354	46,269,312	2,628,872,822
Traditional IRA Plans	-153,972,936	38,076,500	2,406,622,990
SEP Plans	-15,094,078	3,313,204	134,025,661
SIMPLE Plans	-2,630,406	1,568,426	10,350,033
Roth IRA Plans	-8,733,303	9,485,189	77,566,548
Education IRA Plans[3]	89,369	241,238	307,589

Note: Except as noted, all data are from matched forms 1040 and 5498

[1] Tabulations of weighted sample of taxpayers represented on Form 5498.

[2] Amount of contribution shown on Form 5498, limited to amount deducted on Form 1040, either on line 23 (Traditional IRA) or line 29 (SEP or SIMPLE Plans)

[3] Education IRAs were renamed Coverdell Education Savings Accounts (ESAs) in July 2001; does not include Education IRAs owned by non-filing dependents

[4] Withdrawals are reported on Form 1099-R; does not include withdrawals for the purpose of rollovers to other IRA accounts, or Roth IRA conversions.

[5] Residual of change in fair market value minus all the enumerated changes.

Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2000
Statistics of Income Division, Internal Revenue Service.

Figure 3. Total IRA Assets by Institution, 1990-2002
(billions of dollars)

	Mutual Funds	Bank and Thrift Deposits[1]	Life Insurance Companies[2]	Securities Held in Brokerage Accounts[3]	Total
1990	140	266	40	190	637
1991	188	282	45	261	776
1992	237	275	50	311	873
1993	322	263	61	347	993
1994	349	255	69	383	1,056
1995	475	261	81	472	1,288
1996	596	258	92	520	1,467
1997	775	254	135	564	1,728
1998	976	249	156	769	2,150
1999	1,264	244	201	942	2,651
2000	1,247	252	202	929[e]	2,629[p]
2001	1,189	255	210	886[e]	2,540[e]
2002	1,068	263	208[e]	794[e]	2,333[e]

e=Investment Company Institute estimate

p=preliminary from SOI

[1] Bank and thrift deposits include Keogh deposits.

[2] Annuities held by IRA's, excluding variable annuity mutual fund IRA assets.

[3] Excludes mutual fund assets held through brokerage accounts (included in mutual funds).

Note: Components may not add to total because of rounding.

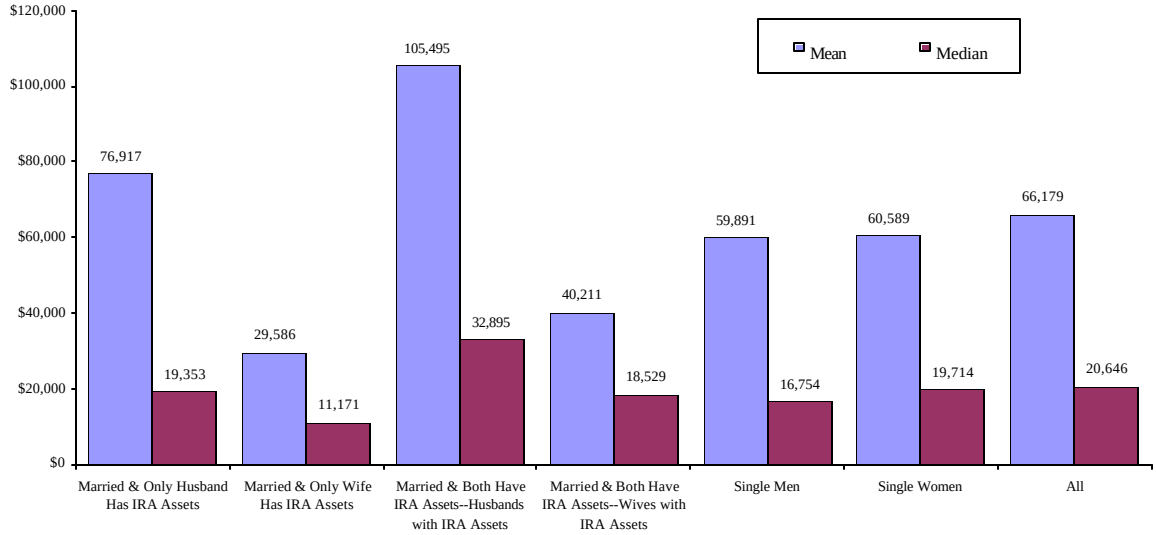
Sources: Investment Company Institute (ICI), Federal Reserve Board, American Council of Life Insurers, and Internal Revenue Service (see ICI, June 2003).

Figure 4. Age Distribution of Taxpayers with Traditional IRAs, 1999

	Taxpayers		Traditional IRA Assets	Traditional IRA Assets Per Taxpayer[1]	
Age Group	(millions)	(share)[2]	(millions of dollars)	Mean	Median
Younger than 18	0	0.1%	168.4	$7,735	$1,970
18 to 24	0	0.7%	936.6	$3,707	$2,191
25 to 34	2	7.7%	35,241.0	$12,435	$5,277
35 to 44	7.4	20.3%	232,433.4	$31,342	$12,103
45 to 54	9.0	24.6%	507,763.5	$56,377	$20,987
55 to 69	4.6	12.5%	371,395.1	$81,459	$27,012
60 to 64	3.9	10.7%	416,451.6	$106,771	$35,419
65 to 69	3.5	9.5%	392,328.4	$112,588	$39,310
70 to 74	2.8	7.8%	303,691.5	$106,902	$35,825
75 or Older	2.3	6.2%	161,856.9	$70,815	$21,245
All[3]	36.6	100.0%	2,422,266.5	$66,179	$20,646

[1] Among the 36.6 million taxpayers with Traditional IRA's.
[2] Percent of taxpayers with Traditional IRAs
[3] Average and median age of Traditional IRA-owning taxpayers was 53 years.
Source: IRS tabulations of weighted sample of taxpayers represented on IRS Form 5498.

Figure 5
Average (Mean) and Median Traditional IRA Balances by Marital Status on Tax Return and Gender, 1999



Note: Total of 36.6 million taxpayers with traditional IRAs in 1999. The small number of married individuals filing separately are shown as single.
Source: IRS, Statistics of Income Division tabulations of weighted sample of taxpayers represented on IRS Form 5498

Figure 6. Comparison of Tax Return and Household Survey Results on Traditional IRA Owners (Median)

	SOI Data[1]	ICI IRA Owners Survey[2]				Survey of Consumer Finances[3]	
Variable	**Dec. 1999**	**June 1999**	**June 2000**	**May 2001**	**May 2002**	**1998**	**2001**
Age of Owner[4]	53	49	53	50	50	51	51
Income[5]	$55,549	$60,000	$62,500	$62,500	$60,000	$55,000	$70,000
Percent of Owners Married	67%	70%	68%	65%	67%	66%	69%
Traditional IRA Assets[6]	$27,181	$20,000	$30,000	$30,000	$30,000	$20,000	$27,000

[1] SOI data tabulated on a tax return basis to approximate household units rather than individual taxpayers. SOI data based on a weighted sample of tax returns. In 1999, about 27.6 million tax returns had at least one IRA-owning individual.

[2] ICI conducts a household survey annually to track the demographic and financial characteristics of IRA owners. These and more recent survey results are available in ICI's *Fundamentals* newsletter.

[3] Tabulations of Survey of Consumer Finances data by ICI research staff.

[4] SOI data tabulated across ages of the primary taxpayer on returns with at least one traditional IRA owner. ICI survey data tabulated across ages of the primary or codecisionmaker. SCF data tabulated across ages of head of household.

[5] The SOI data column reports median adjusted gross income (AGI) of tax returns with traditional IRA owners. The household surveys are the self-reported household's previous year's income.

[6] The Survey of Consumer Finances tabulation includes household assets in all types of IRA's.

Sources: IRS, SOI Division, and ICI tabulations of Federal Reserve Board Survey of Consumer Finances data.

Figure 7. 401(k) Plan Participant Elective Deferrals[1] to 401(k) Plans,[2] 1996-2000

	1996	1997	1998	1999	2000
IRS SOI W-2 Tabulations:[1]					
Taxpayers (millions)	23.0	25.3	27.0	28.9	30.7
Amount Deferred ($billions)	61.2	71.2	82.6	93.1	104.5
Average Deferral	$2,660	$2,814	$3,053	$3,217	$3,408
DOL PWBA(EBSA) Form 5500 Tabulations:[3]					
Active Participants in 401(k) Plans (millions)	30.8	33.9	37.1	n/a	n/a
Participant Contributions ($billions)	64.5	72.5	84.9	n/a	n/a
401(k) Plan Assets ($billions)	1,061.5	1,264.2	1,541.0	1,798.0e	1,790.0e

e=ICI estimate

[1] Elective deferrals are before-tax contributions made by 401(k) participants reported on the taxpayer's W-2. They do not include employer contributions.

[2] Based on a weighted sample of IRS W-2 Forms for the tax years indicated.

[3] Based on the Pension and Welfare Benefits Administration (PWBA; renamed Employee Benefits Security Administration (EBSA) in 2003) annual tabulations of the IRS/DOL/PBGC Form 5500. Form 5500 information is filed by private pension plans on a plan-year basis, which may not coincide with the calendar tax-year basis reporting for the W-2 Form.

Sources: IRS, SOI Division tabulations of weighted sample of taxpayers represented on IRS W-2 Form, Department of Labor, EBSA Abstract of Form 5500 Annual Reports, and ICI.

The Effects of Tax Reform on the Structure of U.S. Business

Ellen Legel, Kelly Bennett, and Michael Parisi, Internal Revenue Service

The 1990's have been described as a period of immense and protracted profit-taking in the stock market. Mergers and acquisitions have impacted business demographics. Tax law changes have also had a marked effect by continually providing incentives and disincentives for certain business legal forms of ownership, such as those affecting the growth rates of companies moving from corporate to noncorporate status. Law changes, such as the landmark 1986 Tax Reform Act, the Small Business Job Protection Act of 1996, and the Omnibus Reconciliation Act of 1997, have had significant impacts on Subchapter C corporations, including small business (or Subchapter S) corporations; partnerships (general, limited, and limited liability companies LLC's); and sole proprietorships.

This paper is an examination of the changes in business demographics or "business organizational choice" of the various types of business during the 1990's and the changes in the historical trend from the 1980 period. Tax data will be used to focus on changes in the various business types, receipts, profitability, and tax rates over two recessions due to modifications in the tax code on administrative records sampled at Statistics of Income (SOI) Division of IRS.

The paper is divided into three sections. The first defines the various types of businesses. The second explains tax law changes during the 1990's. The third analyzes a time series dataset for the three distinct business types (and their subsets) based on tax filings with the IRS.

▶ Organizational Type

For this paper, corporations are divided into C corporations, those taxed at corporate rates, and S corporations, those taxed at individual income tax rates. Partnerships are divided into general partnerships, limited partnerships, and limited liability companies (LLC's). Since the tax treatment of the business organizational forms varies significantly, a brief synopsis follows.

Corporations. Corporation (or Subchapter C) income is generally taxed directly at the business level, and again at the shareholder level for receipt of dividend income. Income distributed to shareholders is only taxable on the after-tax profits earned by the corporation. However, after-tax corporate income is taxable at the shareholder level once it is distributed as dividends or the shareholder realizes capital gain.

Subchapter S. S corporations are incorporated entities that have many of the same attributes as the traditional C corporation, including limited liability, freely transferable ownership, and unlimited life span. Unlike the C corporations, income and losses are passed through to the shareholder and are subject to tax only at the owner level. S corporation shareholders report their shares of income or loss on their own tax returns. Therefore, any resulting tax liability is the responsibility of the shareholders. S corporations offer the benefits of partnership taxation without the liability. Subchapter S corporations must be compared with the limited liability company (LLC), which they resemble in operation and concept. Despite having several appealing characteristics, S corporations do face inherent limitations, including the number and type of shareholders, permitting only one class of stock, and exclusion of foreign, corporate, partnership, or LLC ownership.

Partnerships. Similar to the S corporation, a partnership does not pay tax on its income but passes through any income or losses to its partners. Partners include this passthrough income on their tax returns.

Partnerships may be general partnerships, limited partnerships, limited liability partnerships, and limited liability companies. Creditors of general partnerships, composed solely of general partners, may collect amounts owed to them from both the general partnership assets and the assets of the general partners. General partners are personally liable, limited to their personal resources, and actively participate in management of the business. Limited partnerships (LP's) have at least one general

partner. A limited partner is similar to a corporate shareholder, whose liability to third-party creditors is limited to the amount invested in the partnership. Limited liability partnerships (LLP's) are formed under State-limited liability partnership law. Limited liability partners, whose owners are general partners, are not personally liable for the debts of the LLP or any other partner, nor is the partner liable for the malpractice committed by other partners.

Limited liability companies (LLC's). The LLC is a State-formed entity with the limited liability of a corporation and the tax liability of a partnership. This hybrid entity has quickly become an alternative to the traditional partnership and corporate business structures. The members of the LLC are treated similarly to limited partners, in that income passes through an LLC to the members. The members include this passthrough income on their tax returns. Unlike general partners, the members of the LLC are not personally liable for the LLC's debts.

Data from LLC's have been collected since their first appearance on the partnership annual information return in 1993. LLC's are required to file on the partnership annual information return (Form 1065), although some file on the S corporation return. The LLC data displayed in this article are representative of the data gathered from the partnership annual information return only.

Sole Proprietorships. An owner of a non-farm sole-proprietorship summarizes the income and expenses of the business on Schedule C (or C-EZ) of the owner's individual income (Form 1040) tax return. The net income or loss from the business is added to the owner's personal income from all other sources and taxed at the applicable individual income tax rates.

▶ Tax Law Changes

The Tax Reform Act of 1986 (TRA86), the most comprehensive revision of the Internal Revenue Code since 1954, had a major impact on business decisions in the period after 1986 by broadening the tax base of both individuals and corporations by tightening the corporation "alternative minimum tax," limiting losses from passive activities, and repealing the long-term capital gain exclusion. The most marked effect has been on the changes in the individual and corporate marginal tax rates. In pre-TRA86, the highest individual rate (50 percent) exceeded the highest corporation rate (46 percent) by 4 percentage points.

TRA86 reversed this trend, starting in 1987 and continuing with the final lowered rates of 1988-1990 of 34 percent for corporations and 28 percent for individuals, a 6-percentage point reversal. For 1991 and 1992, this difference was cut in half when the individual rate was increased to 31 percent (Figure A).

In 1993 to the present, the top individual rate increased to 39.6 percent surpassing the highest corporation rate of 35 percent. Although both rates are lower than pre-TRA86, the difference of 4.6 percentage points between the individual rate and the corporation rate looks almost identical to the pre-TRA86 difference of 4 percentage points. The incentive to switch business types declined and reversed. With the reversal in incentives, was there renewed interest in the corporation type of business? We will investigate, using the SOI data for 1990-2000 for all three types of business entities.

Figure A. Top Marginal Tax Rates (Percentages) Corporations and Individuals, 1990-2000

Item	1990	1991-1992	1993-2000
Corporations	34.0	34.0	35.0
Individuals	28.0	31.0	39.6
Difference	6.0	3.0	-4.6

Note: These rates are for the highest levels of taxable income and do not reflect alternative minimum tax.

The Small Business Job Protection Act of 1996 (SBJPA) made several noteworthy changes that have significantly affected S corporation filings. First, the SBJPA increased the maximum number of shareholders from 35 to 75. Second, it enabled financial institutions, which did not use the reserve method of accounting for bad debts, to make an S election. Third, small business

trusts electing to be S corporations, were permitted to be shareholders in an S corporation. Finally, restrictions on the percentage of another corporation's stock that an S corporation might hold were eliminated. S corporations may now make an election to treat the assets, liabilities, income, deductions, and credits of wholly-owned subsidiaries as those of the parent S corporation.

Even though the SBJPA eased Federal tax restrictions on S corporations, the number of S corporation entities has not grown as fast as the partnership limited liability corporation. The IRS ruled in late 1988 (Revenue Ruling 88-76, 1988-2 C.B.360) that any Wyoming LLC would be treated as a partnership. Thus, the door was opened for other States to consider LLC legislation, and the growth of LLC's has not diminished since the IRS's 1988 ruling. By 1993, some 36 States had ruled to allow LLC's as a legal entity. In 1994, that number grew to 46 States, plus the District of Columbia. By 1997, all 50 States and the District of Columbia had enacted LLC legislation. The "check-the-box" regulations, implemented by the IRS in January 1997, relaxed the requirements for LLC's to obtain the favorable partnership tax classification, leading to a wider acceptance of LLC's.

▶ Analysis of Business Data

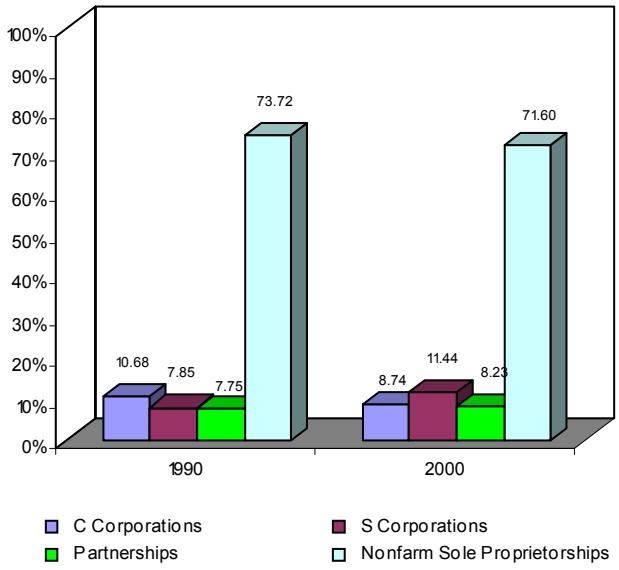
Data in this paper were collected in annual statistical studies by SOI and published in Table 1 (Number of Businesses, Business Receipts, Net Income, and Deficit by Form of Business, Tax Years 1990-2000).

Trends in the Number of Businesses

This segment of the paper places the spotlight on the number of entities and financial data for the 1990-2000 period.

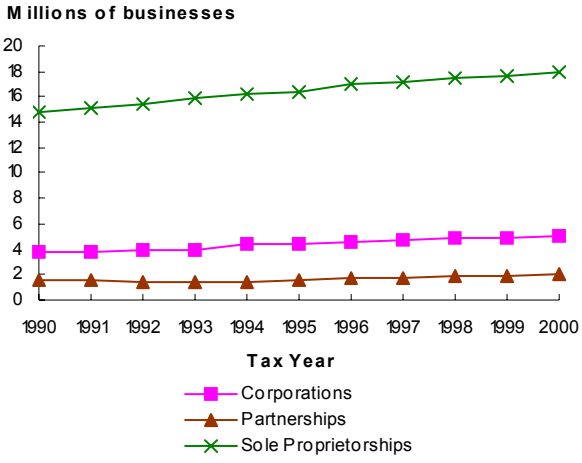
Over the decade of the 1990's S corporations displayed the largest percentage increase of all entities, representing 7.85 percent of all entities in 1990 and 11.44 percent in 2000 (Figure B). The increase in the S corporation percentage of all entities can be attributed to the large number of C corporations that elected to become S corporations after both TRA86 and the SBJPA of 1996. Over the same time period, the percentage of both C corporations and sole proprietorships, compared to all entities, declined. C corporations declined 1.94 percent during the same 10- year period, from 10.68 percent to 8.74 percent.

Figure B. Percent of Entities by Type, Tax Years 1990 and 2000



Number of Entities. Figures C-E present data on the number of entities. Figure C provides a picture of the number of entities by organizational type over time. Figure D displays the total number of entities for C corporations and S corporations. Figure E focuses on the number of entities by type of partnership.

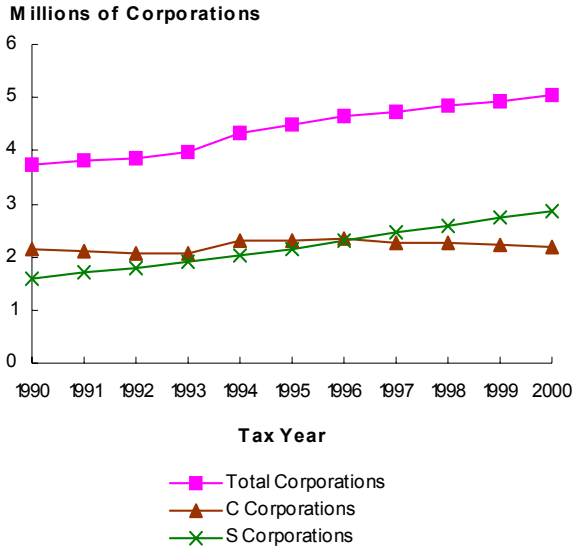
Figure C. Number of Corporations, Partnerships, and Nonfarm Sole Proprietorships, Tax Years 1990-2000



In the 1990's, sole proprietorships had the largest number of entities, (Figure C). Also, the overall growth of sole proprietorships in the 1990's was greater than the growth of corporations and partnerships, both of which grew 1.3 million and .5 million, respectively, compared to an increase of 3.1 million for sole proprietorships. Sole proprietorships grew on an average of 1.9 percent per year throughout the 1990's, with the largest increase of 3.2 percent taking place in 1996.

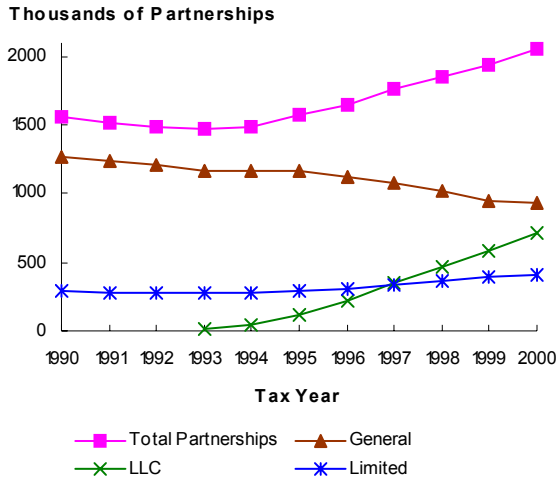
The number of S corporations is plotted in Figure D, which also shows C corporations and total corporations. S corporations are the single largest corporate entity type accounting for 56.7 percent of all corporations in 2000. The number of S corporations has steadily increased since TRA86 and surpassed C corporations in 1997 when the number of C corporations started to steadily decrease. The SBJPA of 1996 also played a role in the growth of S corporations over this time period.

Figure D. Number of Corporations by Type, Tax Years 1990-2000



General partnerships, the most prevalent of all partnership forms (45 percent of the total), have consistently declined since 1990, decreasing from 1.2 million to .9 million, or 26 percent. Limited liability corporations (35 percent of the total) have grown significantly since they first appeared on the partnership tax form in 1993, surpassing limited partnerships on 1997. Figure E shows an increase of nearly .7 million LLC's. Limited partnerships (20 percent of the total) have shown an overall gain of 46 percent since 1993. Prior to that time, limited partnerships displayed an annual decrease since the 1970's.

Figure E. Number of Partnerships by Type, Tax Years 1990-2000



Business Receipts. Figures F-H display data on business receipts by organizational type. Business receipts are plotted in Figure F for all organizational form types. Figure G focuses on corporations, while Figure H focuses on partnerships.

Business receipts for C corporations have always far outweighed receipts for partnerships ($2.1 trillion) and sole proprietorships ($1 trillion). Both show slight growth; and in 1996, partnerships passed sole proprietorships for the first time, as shown in Figure F.

Business receipts for C corporations have always surpassed S corporations, and, in the 90's, the gap has been growing progressively for C corporations from $6.7 trillion to $10.5 trillion, as shown in Figure G. Even though business receipts for C corporations have increased by 70.2 percent for the 90's, the number of C corporations has only increased 2.0 percent over the same period. S corporation business receipts have likewise increased by 123.8 percent while the number of S corporations has increased significantly, 81.9 percent. S corporations now comprise 20.1 percent of the total business receipts, compared to 14.3 percent for the beginning of the decade.

Figure F. Business Receipts of Corporations, Partnerships, and Nonfarm Sole Proprietorships, Tax Years 1990-2000

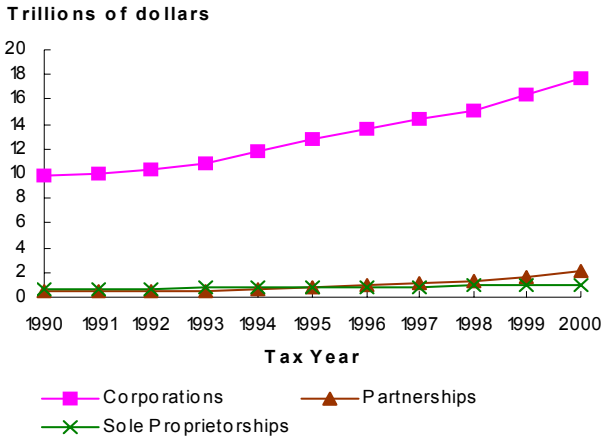
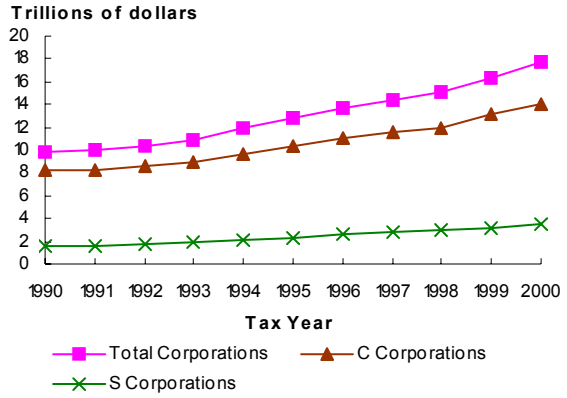
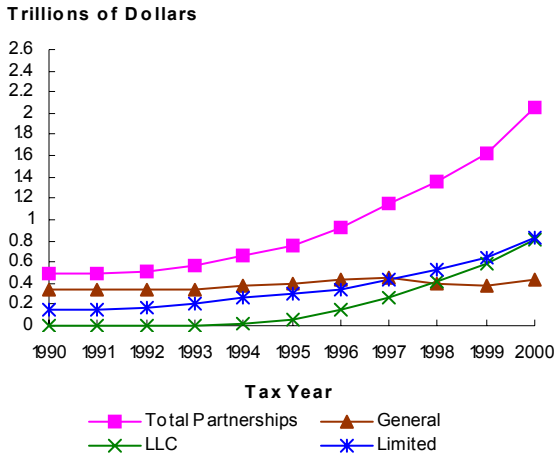


Figure G. Business Receipts by Type of Corporation, Tax Years 1990-2000



The growth of partnership business receipts is primarily due to the inception of LLC as a partnership entity choice, as shown in Figure H. In 2000, LLC's represented $805.5 billion (39.1 percent) of the $2,061.7 billion in partnership business receipts reported. Partnership business receipts have increased at an average annual rate of 33.4 percent since LLC's were first recognized on the tax form in 1993, or 267.5 percent over the 8-year period. Limited partnerships now account for more than $830.4 billion of partnership business receipts, or 40.2 percent, while barely representing one fifth of all partnerships, 402.2 thousand. General partnerships, the largest partnership entity, represent only $425.7 billion of total partnership business receipts, or

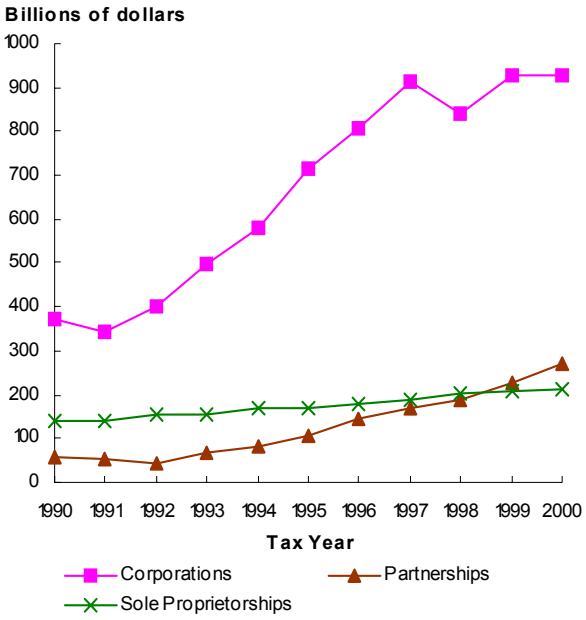
Figure H. Business Receipts by Type of Partnerships, Tax Years 1990-2000



20.6 percent. In 1998, Limited and LLC's surpassed General partnerships for the first time.

Net Income (less deficit). Figures I-K show overall trends in net income (less deficit) or profits by organizational type. Net income (less deficit) is displayed in Figure I for all business form types. Figures J and K focus on corporations and partnerships, respectively.

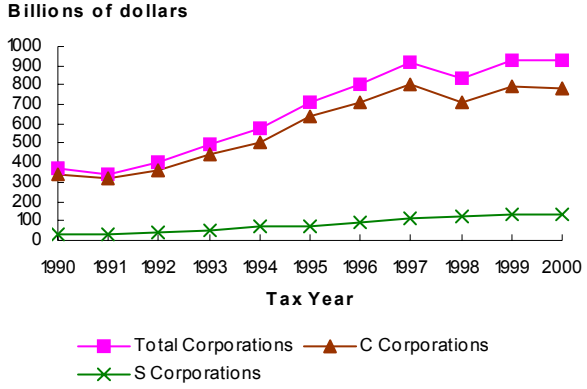
Figure I. Profits of Corporations, Partnerships, and Nonfarm Sole Proprietorships, Tax Years 1990-2000



Profits for corporations are now at $927.5 billion, compared to partnerships at $269 billion and sole proprietorships at $214.7 billion. Even though corporations dominate, the percentage decreased from 70 percent to 56 percent of the total for all business entities, due to an increase in deficit returns for C corporations. Over the decade, there has been a smaller increase for sole proprietorships, while partnerships have been progressively gaining in profits, and actually bypassing sole proprietorships in 1999 for the first time, as shown in Figure I. Partnerships, which started the decade at 3 percent of the total and grew to 19 percent, reached a new level at $269 billion in 2000. Sole proprietorships ended the decade at $214.7 billion, decreasing from 27 percent of the total in 1990 to 15 percent by 2000.

Corporate profits have grown steadily since 1991, peaking in 1997, with a slight downturn in 1998, but bounced back to near-1997 levels in 1999, and flat growth in 2000 (Figure J). Net income returns for both C and S corporations have grown steadily over the decade, peaking in 2000, with C corporations greater than S corporations in the entire decade. However, S corporations started the decade with 12.3 percent of the total and increased to 15 percent of the total by the end of the decade, due to the increase in the number of deficit returns for C corporations, from an average of $140 billion from 1990 through 1997, to $207 billion in 1998, $250 billion in 1999, and $348 billion by 2000.

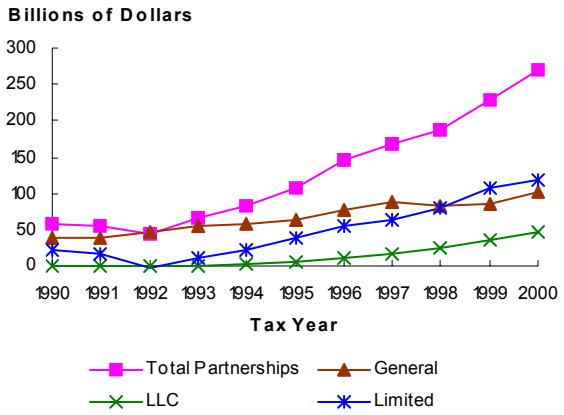
Figure J. Profits of Corporations by Type, Tax Years 1990-2000



Limited partnerships (LP's) accounted for the majority of growth seen in partnership profits (net income less deficit) during the 1990's, as shown in Figure K. Prior to 1993, LP's had been consistently decreasing.

Then, in 1993, LP's surpassed LLC's in profits and surpassed general partnerships by 1999. Net income returns for limited partnerships first surged ahead of general partnerships, with net income returns in 1998 accounting for the overall growth in all returns (net income and deficit return) by 1999. In 2000, profits continued to grow for LP's, accounting for $119.5 billion of the $268.9 billion in total profits reported by all partnerships, while still only representing one fifth of all partnerships. Over the decade, the profits of LP's grew at an average annual rate of 46.5 percent, displaying a steady stream of positive profits through the 1990's. While limited partnerships had the largest number of net income returns by the end of the decade, LLC's experienced an increase in the number of deficit returns, bypassing general partnerships in 1998 and limited partnerships by 2000.

Figure K. Profits by Type of Partnerships, Tax Years 1990-2000



▶ Conclusions

Sole proprietorships began and ended the decade with the largest percentage and number of business entities. However, the percent of the total dropped from 73.4 percent to 71.6 percent. S corporations had been steadily increasing since 1990, and by 1997 surpassed C corporations, which had been decreasing since 1994. Partnerships have gained in the decade, due to the growth of LLC's since their inception in 1993. During the same

time, general partnerships have been on a gradual downward turn.

The most obvious reason for the increase in LLC's is their advantages over other business entity types. All members of LLC's have the same limited liability protection; by comparison, a partnership must have at least one general partner that does not have limited liability. LLC's offer flexible management and flexible ownership, allowing members to participate in the LLC's management without exposing themselves to possible personal liability. LLC's offer several advantages over S corporations since there is no limitation on the number of owners, type of owners, or on the allowable types of interest. S corporations are limited to seventy-five shareholders, none of which can be nonresident aliens, and may only have one class of stock. The flexibility that an LLC offers helps attract a broader range of business investors than is possible with an S corporation.

The business receipts of corporations outdistanced partnerships and sole proprietorships due to the dominance of C corporations. However, C corporations have dropped from 86 percent of the total receipts of all corporations to 83 percent by the end of the decade. General partnerships have been decreasing since 1997, and by 1998 were surpassed by LLC's and limited partnerships.

For net income (less deficit) returns, C corporations are the largest of all entities, but dropped from 64 percent to 56 percent by 2000. Net income rose for all corporations due to C corporations, with the gap in the growth rate widening over S corporations. Partnership income also rose due to the steady increase in income for limited partnerships that surpassed general partnerships by 1999.

For C corporations, there has been a 68-percent increase in deficit returns since 1998. Likewise, for partnerships, losses have been increasing since 1994. General partnerships were surpassed by LLC's in 1998, and by limited partnerships by 2000.

Two factors accounted for the growth in S corporations and LLC's. In 1994, the number of States, 36, that permitted the formation of LLC's commenced a 7-year rise in the number of partnerships. The passage of the Small Business Job Protection Act (SBJPA96) increased the number of allowable shareholders from 35 to 75, and contributed to the growth of S corporations over C corporations, especially in the smaller asset categories. Both of these factors allowed income to be passed on to the individual and taxed at the lower individual rate rather than at the corporate rate.

▶ References

Bennett, Kelly, "S Corporation Returns, 2000," *Statistics of Income Bulletin,* Spring 2003, Volume 22, Number 4 (as well as *Bulletin* articles on these organizations for other years).

Gill, Amy and Wittman, Susan, "S Corporation Elections After the Tax Reform Act of 1986," *Statistics of Income Bulletin,* Spring 1998, Volume 17, Number 4.

Manolakas, Thomas G. (2000), *Partnerships and LLC's Tax Practice and Analysis*, CCH Incorporated, Chicago.

Petska, Tom, "Taxes and Business Organizational Choice: Déjà Vu All Over Again?"

Petska, Tom, "Taxes and Organizational Choice: An Analysis of Trends, 1985-1992," *Statistics of Income Bulletin*, Spring 1996, Volume 15, Number 4.

Pratt, William, "Partnership Returns, 2000," *Statistics of Income Bulletin*, Fall, 2002, Volume 22, Number 2 (as well as *Bulletin* articles on these organizations for other years).

Schlesinger, Michael (2000), *S Corporations Tax Practice and Analysis*, CCH Incorporated, Chicago.

Acknowledgment

The authors would like to extend special thanks to: Patrice Treubert, for developing the SAS data set; Bill Pratt and Tim Wheeler, for producing the partnership data. This article was a continuation of *The Effects of Tax Reform on the Structure of U.S. Business*, written by Tom Petska. Any errors are the responsibility of the authors.

Table 1. Number of Businesses, Business Receipts, Net Income, and Deficit by Form of Business, Tax Years 1990-2000

[All figures are estimates based on samples--returns are in thousands, money amounts are in millions of

Form of business	Tax Year										
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
All Industries											
Number of businesses	20,053	20,499	20,849	21,281	21,990	22,488	23,240	23,645	24,113	24,449	25,007
Business receipts	11,074,423	11,161,332	11,612,309	12,183,715	13,330,389	14,353,781	15,418,578	16,473,256	17,285,224	18,899,109	20,719,321
Net income (less deficit)	528,640	507,822	598,877	721,211	826,282	990,291	1,128,474	1,270,285	1,227,179	1,365,385	1,411,206
Net income	830,475	811,635	865,707	976,024	1,078,273	1,251,080	1,415,081	1,590,638	1,615,164	1,811,173	1,991,803
Deficit	301,835	303,813	266,830	254,813	251,991	260,788	286,608	320,353	387,985	445,788	580,597
Corporations											
Number of businesses	3,717	3,803	3,869	3,965	4,342	4,474	4,631	4,710	4,849	4,936	5,045
Business receipts	9,860,400	9,965,600	10,360,400	10,865,500	11,883,600	12,785,800	13,659,500	14,460,900	15,010,300	16,314,000	17,636,600
Net income (less deficit)	370,600	344,900	402,000	498,100	577,300	714,200	806,500	915,400	838,200	929,000	927,500
Net income	552,500	535,800	570,400	658,600	739,500	880,700	986,800	1,117,800	1,091,100	1,229,300	1,336,600
Deficit	181,900	190,900	168,400	160,500	162,200	166,500	180,300	202,400	252,900	300,300	409,100
C corporations											
Number of businesses	2,142	2,105	2,084	2,063	2,318	2,321	2,327	2,258	2,261	2,210	2,185
Business receipts	8,272,300	8,310,100	8,569,600	8,897,600	9,710,200	10,419,300	11,087,500	11,620,300	12,006,200	13,071,200	14,078,900
Net income (less deficit)	338,400	315,800	355,800	444,000	503,300	637,300	713,300	803,700	713,400	791,600	788,400
Net income	484,600	469,800	490,800	571,900	632,500	765,600	854,200	963,100	920,100	1,041,900	1,136,800
Deficit	146,200	154,000	134,900	127,900	129,200	128,300	140,900	159,400	206,700	250,300	348,400
S corporations											
Number of businesses	1,575	1,698	1,785	1,902	2,024	2,153	2,304	2,452	2,588	2,726	2,860
Business receipts	1,588,100	1,655,500	1,790,800	1,967,900	2,173,400	2,366,500	2,572,000	2,840,600	3,004,100	3,242,800	3,557,700
Net income (less deficit)	32,200	29,100	46,200	54,100	74,000	76,900	93,100	111,700	124,800	137,400	139,100
Net income	67,900	66,000	79,600	86,700	107,000	115,100	132,500	154,700	171,100	187,400	199,800
Deficit	35,700	36,900	33,500	32,600	33,000	38,200	39,400	43,000	46,300	50,000	60,700
Partnerships											
Number of businesses	1,553	1,515	1,485	1,468	1,494	1,581	1,654	1,759	1,855	1,937	2,058
Business receipts	483,417	483,164	514,827	560,999	656,159	760,618	915,844	1,141,963	1,356,656	1,615,762	2,061,764
Net income (less deficit)	16,610	21,407	42,917	66,652	82,183	106,829	145,218	168,241	186,705	228,438	268,991
Net income	116,318	113,408	121,834	137,441	150,928	178,651	228,158	262,373	297,874	348,468	409,973
Deficit	99,708	92,002	78,918	70,788	68,745	71,822	82,939	94,132	111,171	120,030	140,982
General											
Number of businesses	1,268	1,245	1,214	1,174	1,162	1,163	1,121	1,081	1,016	951	937
Business receipts	334,184	333,190	336,912	348,350	375,033	395,396	430,893	450,835	399,306	382,760	425,752
Net income (less deficit)	37,771	38,109	46,194	55,029	58,721	63,626	77,447	88,235	82,766	85,767	101,786
Net income	81,903	78,331	81,314	85,129	87,681	92,587	106,074	113,265	107,710	108,488	127,059
Deficit	44,132	40,222	35,119	30,100	28,959	28,961	28,627	25,030	24,943	22,720	25,272
Limited ([1])											
Number of businesses	286	271	271	276	284	299	312	328	369	397	402
Business receipts	149,233	149,975	177,914	205,554	257,887	302,337	338,916	423,969	534,249	644,247	830,430
Net income (less deficit)	-21,161	-16,702	-3,278	11,360	21,411	38,320	55,458	62,946	79,329	107,937	119,512
Net income	34,415	35,078	40,521	51,238	59,545	76,030	97,722	109,036	131,493	157,245	170,929
Deficit	55,576	51,780	43,798	39,878	38,134	37,710	42,263	46,090	52,165	49,308	51,417
LLC											
Number of businesses	N/A	N/A	N/A	17	48	119	221	349	471	589	719
Business receipts	N/A	N/A	N/A	7,095	23,239	62,885	146,036	266,990	423,101	588,755	805,582
Net income (less deficit)	N/A	N/A	N/A	263	2,051	4,884	12,313	17,060	24,609	34,734	47,692
Net income	N/A	N/A	N/A	1,073	3,702	10,035	24,362	40,072	58,672	82,736	111,984
Deficit	N/A	N/A	N/A	810	1,651	5,151	12,048	23,013	34,063	48,002	64,292
Nonfarm Sole Proprietorship											
Number of businesses	14,783	15,181	15,495	15,848	16,154	16,433	16,955	17,176	17,409	17,576	17,905
Business receipts	730,606	712,568	737,082	757,215	790,630	807,364	843,234	870,392	918,268	969,347	1,020,957
Net income (less deficit)	141,430	141,516	153,960	156,459	166,799	169,262	176,756	186,644	202,275	207,947	214,715
Net income	161,657	162,427	173,473	179,983	187,845	191,729	200,124	210,465	226,190	233,405	245,231
Deficit	20,227	20,911	19,512	23,524	21,046	22,467	23,368	23,821	23,915	25,458	30,515

N/ANot applicable

[1] For Tax Years 1993-1995, Limited Partnerships include Limited Partnerships, General Limited Liability Partnerships, and Limited Liability.

For Tax Years 1998-1999, Limited Partnerships include Limited Partnerships and Limited Liability.
For Tax Year 2000, Limited Partnerships include Domestic Limited Partnerships and Domestic Limited Liability.

Statistical Information Services at IRS: Improving Dissemination of Data and Satisfying the Customer

Beth Kilss and David Jordan, Internal Revenue Service

IRS's Statistics of Income (SOI) Division conducts statistical studies on the operations of the tax laws and publishes annual reports on corporations and individuals, the quarterly *SOI Bulletin*, an annual research report, special periodic reports and compendiums, and the annual IRS *Data Book*. The IRS World Wide Web site provides users an easy option for accessing these reports and other SOI data and also serves as a conduit for releasing other IRS information. Periodic news releases to the mainstream media announcing key products raise awareness of the data SOI makes available to the public. SOI's Statistical Information Services (SIS) office, comprised of statisticians and economists, has emphasized top-quality, customer-focused service throughout its 14-year history and strives to provide timely, accurate, and well-documented guidance on the availability of SOI data and other statistical services.

This paper will provide an overview of SOI efforts to improve and expand data dissemination. In the first section of the paper, some background information about the SOI Division and its Statistical Information Services office is highlighted and outlined. The second section discusses recent improvements to SOI's web site. The third section looks at innovations in data dissemination through the web site, the SIS office, and news releases. In the last section, some results from a recent survey of SIS customers are presented, along with how SOI is using these results to identify problem areas and improve customer service.

▶ Background Information

Congress created the Statistics of Income Division nearly 90 years ago in the Revenue Act of 1916, some 3 years after the enactment of the modern income tax in 1913. Since that time, the Internal Revenue Code has included virtually the same language mandating the preparation of statistics. Section 6108 of the Code currently states that "…the Secretary (of the Treasury) shall prepare and publish not less than annually statistics reasonably available with respect to the operations of the internal revenue laws, including classifications of taxpayers and of income, the amounts claimed or allowed as deductions, exemptions, and credits."

SOI's mission is to collect, analyze, and disseminate information on Federal taxation for the Office of Tax Analysis, Congressional Committees, the Internal Revenue Service in its administration of the tax laws, other organizations engaged in economic and financial analysis, and the general public. Its mission is similar to that of other Federal statistical agencies—that is, to collect and process data so that they become useful and meaningful information. However, SOI collects data from tax returns rather than through surveys, as do most other statistical agencies. These data are processed and provided to customers, in the form of tabulations or microdata files. Although the IRS uses SOI data, the primary uses for SOI data are outside of IRS, in policy analyses designed to study the effects of new or proposed tax laws and in evaluating the functioning of the U.S. economy.

Throughout its long history, SOI's main emphasis has been individual and corporation income tax information. However, growth has occurred over the years in the nature and number of studies undertaken. In addition to individuals and corporations, SOI Division also conducts statistical studies on partnerships, sole proprietorships, estates, nonprofit organizations, and trusts, as well as special projects or studies of international activities. In 1980, the SOI program consisted of 26 projects; now, in 2003, the SOI program consists of over 60 projects. While the number of projects has more than doubled over the past 20 years, this growth has been accompanied by even larger increases in the amount of data extracted from the various tax and information returns during that same timeframe.

SOI Products and Services. Statistics of Income information reaches thousands of outside tax practitioners and researchers, State and local governments, the media, the public at large, and staff within the IRS itself through SOI's published products and electronic media.

SOI user inquiries come from a wide array of interests. The detailed income and expenditure data provided on tax and information returns are highly regarded and more reliable than similar survey data because there are penalties for misreporting. SOI information is published in the quarterly *Statistics of Income Bulletin,* which contains four to eight articles and data releases of recently completed studies, as well as historical tables covering a variety of subject matter; separate annual "complete reports" on individual and corporation income tax returns, which contain more comprehensive data than those published in the *Bulletin;* and the annual *Corporation Source Book,* which presents detailed income statement, balance sheet, and tax data by industry and asset size.

Other SOI publications include special compendiums of research, which are published periodically on such topics as nonprofit organizations and estate taxation and personal wealth, and research articles published in a series of reports, usually annually, which document technological and methodological changes in SOI programs and other related statistical uses of administrative records. More recently, SOI Division has become the publisher of the Internal Revenue Service *Data Book* and the IRS Office of Research's annual research conference proceedings.

The IRS World Wide Web site offers easy access to these products and other services free of charge at www.irs.gov/taxstats. More specifically, at this site, users will find a combination of files presenting tax-related data on individuals, corporations, and other entities; articles about SOI data; information about SOI products and services; and non-SOI products, including the *Data Book,* Compliance Research projections, and nonprofit Master File microdata records. At present, over 1,734 files reside there.

Statistical Information Services Office. Over 14 years ago, SOI created its Statistical Information Services (SIS) Office to facilitate the dissemination of SOI data and reports and respond to all data and information requests. This office was established as a direct result of the management study mandated by the Office of Management and Budget Circular A-76 process in the late 1980's, which required SOI Division to determine its "most efficient organization." The establishment of the SIS office was one of a number of recommendations coming out of the A-76 process. Within 2 years of the decision to centralize responses to all data and information requests received in SOI, the SIS office opened for business in early 1989.

During the first 5 years, there was a steady increase each year in the number of telephone and written requests for SOI data and publications. Little by little, the SIS office began to establish a reputation for always providing an answer or at least a referral to someone who could provide an answer. In the midst of building up reference materials, setting up a library, and training new staff to help handle the growing workload, the technologies available were also changing. Word processors, typewriters, photocopy machines, and telephone were the main tools used to support this work at the beginning. Fortunately, within a few years, a computerized system for tracking and recording all customer requests was designed and implemented. The system is periodically updated and refined to keep pace with the changing functionalities of the SIS office, and it is used to permanently record all requests received, invoice customers for reimbursable products, and generate various reports about customers and their requests.

In more recent years, SOI has expanded the SIS function to better serve the public, first, by means of an electronic bulletin board (in 1992), and, more recently, by participating with the rest of IRS on the World Wide Web (in 1996). As electronic dissemination has grown dramatically, the number of written and telephone requests has declined, but questions that do not lend themselves to answers over the Internet have grown more complicated.

The current SIS staff handled nearly 2,800 information requests in Calendar Year 2002, and an equal number in Fiscal Year 2003. During FY 2003, about half of the requests were received from e-mails and faxes, while about 47 percent were over the telephone. The remaining requests were from letters or face-to-face meetings with "walk-in" visitors. The top three groups of requesters responded to directly were: consultants (22.5 percent), private citizens (17.2 percent), and other IRS offices (9.3 percent). After these categories, the next most frequent requesters were: universities, corporations, State

and local governments, Federal agencies, and the media, accounting for about 31.0 percent of all inquiries. The remaining requests (about 20 percent) came from students, nonprofit organizations, associations, law firms, accounting firms, Congress, banks, foreigners, and public libraries. While, in recent years, the overall level of direct requests has stayed fairly constant, Internet downloads continue to rise, which has enabled SIS staff to focus on more detailed research for customers seeking material not available from the web site. Many requests involve duplication of perhaps 25 or more years of historical material that is either not available all in one place elsewhere, or only available in hard copy. As always, any data provided, whether published or unpublished, are distributed free of charge, except for certain reimbursable products, which are sold to recoup dissemination costs.

▶ Improvements to SOI's Web Site

Not so long ago, delivering customer products and disseminating SOI data electronically meant providing data files on several magnetic tape reels or on diskettes to customers for use on their personal computers. In June 1992, the Division took a major step toward disseminating its data electronically when the SOI Electronic Bulletin Board was established. By dialing up the EBB, users had access to SOI files (primarily tabulations from *SOI Bulletin* articles, data releases, and the historical data section), files from IRS *Data Book* tabulations, IRS Master File microdata records of exempt organizations, and documents containing projection data produced by IRS's Office of Research.

Four years later, in the fall of 1996, a select group of SOI and other IRS products became available to the public in the "Tax Stats" area of the IRS home page. Initially, the site included over 700 files, which have more than doubled to 1,734 files currently. This year alone, 259 new files were added, including new unpublished files. SOI's Internet site offers a combination of files presenting SOI tables, articles about SOI data, and information about SOI products and services, as well as non-SOI products, including annual IRS *Data Book* tables, Compliance Research projections, and nonprofit Master File microdata records. Improvements to the web site have been slow in coming over the past 7 years, in large part due to the fact that SOI Division does not have direct control over the site, although, recently, this has begun to change. One major improvement is that SOI is able to upload files and make changes to the site within 30 minutes, whereas previously, the Division was forced to go through several channels to update pages, which could take 1 or more weeks.

More dramatic changes are on the horizon, although the extent of those changes to Tax Stats remains to be seen. However, the future looks promising because the SOI Director commissioned a Tax Stats Web Advisory Group—an inhouse team of Internet-savvy staff members working with several members of SOI's Consultants' Panel—to investigate various options for improving the site design.[1] The group is evaluating the current Tax Stats web site and recommending changes to improve accessibility, visibility, other important aspects of web design that enhance the site's capabilities, and overall effectiveness as a medium of data dissemination.

During FY 2003, the group evaluated the effectiveness of other U.S. Government statistical web sites and dozens of corporate and organizational sites and gathered the first ever data on customer satisfaction as part of the survey conducted by SOI's Statistical Information Services office, which is discussed later in this paper. In the near future, the group has plans to survey two specific user groups—the National Tax Association members and the Federation of Tax Administrators.[2] The Web Advisory group also helped develop prototype pages to experiment with content organization and layout, presented examples of prototype pages to the Consultants' Panel members of the group for their feedback, and began to work with the outside contractor who manages the web site to develop the taxonomy for organizing all irs.gov web content.

The following is a list of some of the specific enhancements that the advisory group is proposing:

- Develop a Tax Stats-specific search engine.
- Add data base and query capabilities so customers can create their own tables.
- Add scripting capabilities to support dropdown boxes, online surveys, and other functionality.

- o Identify the Tax Stats portion of irs.gov as SOI Tax Stats.

- o Allow SOI to use a greater variety of formats, font sizes, colors, typefaces, and graphics on all pages.

- o Allow the addition of a shopping cart so customers can select a number of different files before downloading.

Looking ahead, the goals and objectives of the group are to:

- o Continue development of prototype pages and eventually solicit feedback from other external users regarding effectiveness.

- o Continue evaluating whether capabilities within the current irs.gov environment are sufficient to satisfy distinct customers' needs.

- o Schedule writing classes to train SOI staff to "write to the Web."

- o Explore alternatives that would give SOI Division more control of site management.

The Advisory Group has an ambitious agenda, but progress is being made. This group is moving ahead with plans to conduct usability testing on proposed changes next March, which will allow them to develop guidelines for creating improved web pages by June, and begin programming new pages by next summer.

▶ Innovations in Data Dissemination

Data dissemination is an important part of SOI Division's mission. Webster's Dictionary defines disseminate as "to scatter widely" or "to spread out," which SOI has been able to do more successfully in recent years because of new technologies. Improved technologies have also allowed SOI to increase the amount of data produced over the years, as well as the speed with which they are produced, but these increases have also served to increase the expectations of users. Several innovations have been implemented in the past few years, and some quite recently to improve dissemination of SOI data. This section looks at innovations in data dissemination through the web site, news releases to the media, and the Statistical Information Services office.

- **IRS World Wide Web.** If all proposed improvements discussed above and others yet to be decided are implemented, the Tax Stats portion of the IRS World Wide Web site, www.irs.gov, will greatly improve SOI's ability to disseminate data online. Recently, intermediate steps have been taken to enhance data dissemination. For example, SOI's Webmasters used a different format to post *SOI Bulletin* material on Tax Stats. Instead of executable files, for each article or data release, there is now a PDF file for the entire article (including tables), plus separate links for each of the Excel tables.

  In addition, SOI's Webmasters have changed pages on Tax Stats that relate to the *SOI Bulletin*. They have added links and a separate page for the historical tables/appendix of the *Bulletin,* changed the "landing" page for the *Bulletin*, and added new pages for each issue of the *Bulletin*. These small steps will go a long way toward improving data dissemination—the historical data are now easier to find, and the table files on Tax Stats are much more user-friendly.

  Perhaps one of the more notable improvements during FY 2003 has been SOI's ability to make its published products available sooner because of a new printing contract. SOI staff now deal directly with their contract printer as opposed to many layers of other IRS and Government Printing Office staff. Furthermore, turnaround time on printing has improved to 2 weeks or less compared to 1 month or more. Hand in hand with this improvement is the more timely placement of *SOI Bulletin* articles and data releases on the Web because of improvements in, and more control over, placing files on the Web as noted earlier in this paper.

Another fairly recent improvement to the Web site is the addition of the Tax Stats Dispatch mailing list (to which users can subscribe to receive announcements about new products and services), which currently has around 3,000 subscribers.

- **News Releases and Other Marketing.** SOI Division has an abundance of tax-related data and information available for use by the general public. It is a unique data source that is well-known in the tax community and in the Federal statistical data arena, but is not commonly familiar to the public as, say, Census data are. In addition to increased awareness of SOI data, which has resulted from their availability via the Internet, SOI Division is taking further steps to promote the use of its data through other means. Within the past year, SOI staff began working with the IRS Media Relations office to improve news releases to the mainstream media, when a publication is about to be released to the public. News releases are now being written to focus on one or two things that are of interest in a particular publication. They are shorter and to the point and designed to attract the attention of a wider range of journalists. In particular, SOI is trying to get the attention of more than just the *The Wall Street Journal* and *The New York Times*, i.e., the Associated Press, Bloomberg, Dow Jones, Reuters, and *USA Today*, for example. SOI has also taken steps to expand news releases to cover other publications, products, and services beyond just the quarterly *Statistics of Income Bulletin,* which, for the most part, has been the only publication announced to the media. SOI Division staff are also asking in current and future publications for a specific citation when SOI data are identified with the hopes that repeated "branding" of our products and services will raise users' awareness and improve SOI's visibility as a producer of financial statistics from various tax and information returns. IRS Internal Communication Division is also helping SOI to expand the visibility of SOI data within the IRS itself by using multiple communications tools to make Service-wide IRS employees aware of SOI and what it has to offer.

- **Statistical Information Services Office.** The SIS staff is constantly working toward improving its ability to disseminate SOI products and services more quickly to more customers. The Web improvements already discussed have reduced the number of routine calls received by SIS staff, enabling them to improve response times and followups on more complex calls, which require research. SIS staff members can provide more data electronically on diskette or CD-ROM because of improved equipment to produce them. In addition, better mechanisms are now in place for responding to e-mails received via the Tax Stats Web site, which are forwarded to the SIS office, where SIS staff are able to respond more quickly. However, to ensure a better understanding of what SOI's customers need and want, and to enable those responding to customer inquiries to continually improve service to the customer, SIS staff conducted their first customer satisfaction survey in 2003.

Statistical Information Services Customer Satisfaction Survey

SOI Division has employed a variety of methods over the years to elicit customer feedback and expectations and to share that information with SOI staff so that they can improve service to the customer on many levels. One method to receive customer feedback about publications was through a user survey, which was included in certain publications. Another way to deal with concerns and expectations of the professional user community at large has been through the SOI Consultants' Panel (which SOI Director Tom Petska discussed in another paper in this session)—one of several forums that SOI uses to make long-term improvements in availability and accessibility of SOI information.[1]

Other customer feedback has been received through formal meetings with users, a notable example being the Public-Use File Users' Group (also mentioned in the

Petska-Kilss paper) and the Statistical Information Services (SIS) office through informal conversations with users. More recently, however, customer satisfaction has become a major part of the Internal Revenue Service Mission Statement:

> "Provide America's taxpayers top-quality service by helping them understand and meet their tax responsibilities and by applying the tax law with integrity and fairness to all."

To help achieve that mission and assess how it is perceived by those it serves, SOI Director Tom Petska has given his full support to the use of customer satisfaction surveys to evaluate SOI effectiveness as a data provider to its customers, including the Office of Tax Analysis and the Joint Committee on Taxation, which were first surveyed in 2000, and then the Bureau of Economic Analysis, which was first surveyed in early 2002. In late 2002, Tom requested that SOI further expand its survey efforts to include those SOI customers served by the Statistical Information Services (SIS) office. Thus was born the SIS Customer Satisfaction Survey, which was completed in late August 2003.

This survey, developed last fall by two of SOI's mathematical statisticians, Kevin Cecco and Diane Dixon, in close consultation with the authors of this paper, was approved by the Office of Management and Budget (OMB) in December 2002. A month later, SIS staff began implementing the survey, which they planned to give to a total of approximately 400, or 1 in 4, customers. These customers were being randomly sampled from among the daily roster of calls and e-mails, including requests from consultants, corporations, the media, academia, State and local governments, and other Federal agencies.

The survey period was originally set for January through July 2003. However, with six people sampling customers, there were difficulties keeping track of a sample rate of 1 in 4. Therefore, in order to increase the total number of customers sampled, SIS staff decided to extend the survey by 1 month.

Throughout SIS's 14-year history, staff has emphasized top-quality, customer-focused service and striven to provide timely, accurate, and well-documented multimedia products. They now hope to use the survey results to identify problems as well as successes and incorporate those results into plans for improving customer service. In particular, SIS staff hopes to evaluate its effectiveness as a data provider. The survey questions (17 of them) dealt with communication, characteristics of staff, opinions of products, and overall satisfaction, as well as timeliness, completeness of information provided, and usefulness of the Web site. Surveys were either faxed or e-mailed to sampled customers, and results were expected to help SIS:

- Determine if SOI products/data satisfied customer needs.
- Determine if SOI products/data were received timely.
- Determine if SOI's Web site is user-friendly and what would make it more so.
- Determine the type of media customers prefer for receiving SOI data.
- Determine the type of new products customers would be interested in receiving.

Results from the Statistical Information Services Office Survey. The following is a summary of results from the survey:

- Total Surveys Distributed…….288
- Surveys Completed…..……142
- Survey Response Rate…….49%
- Respondents Who Were First-Time Customers…….....……45%

Additional Results from the SIS Survey. Much has been learned from the survey. SOI mathematical statistician Diane Dixon analyzed the results extensively and met with the SIS staff to help interpret them. The question on overall satisfaction, for example, showed that customers are generally satisfied. About 87 percent rate

their overall satisfaction as "very high" or "high," while only 3 respondents, or 2.3 percent, rated their overall satisfaction as "low" or "very low." It also appears that 35 percent of respondents learned about the SIS office from the Tax Stats web site and that 45 percent of respondents were first-time users of SOI's Statistical Information Services office. Of those surveyed, the largest customer groups were Federal, State, and local government employees, consultants, and other researchers. Other more open-ended questions showed that customers want to receive notices of data releases, and an overwhelming percentage want to have access to downloadable files on the Web site.

There is more to learn from these results, and over the next few months, SIS staff will carefully sift through them to plan improvements to customer service. Planning is also under way for another customer satisfaction survey in 2004. It is expected that the survey will continue on a regular basis because of SOI's strong commitment to its customers.

▶ Summary and Conclusions

IRS's SOI Division is a world-class statistical organization with an abundance of tax-related data, which are available to the general public. Although these data are being disseminated widely, there is much more that can be done to broaden the distribution of available information. SOI is continuing its efforts to improve customer service, increasing its efforts to raise awareness about SOI data, working harder to make its data more accessible to users, and expanding efforts to disseminate its products and services more widely than ever before. This paper has been an overview of recent developments and provides a brief glimpse of activities to expand the customer base. It is hoped that by making this presentation at a professional conference such as this, SOI will be introducing even more analysts and researchers to the rich body of statistics so readily available from the Internal Revenue Service, and, with a bit of luck, it may even get suggestions for further improvements to give those products and services the audience they deserve.

▶ Notes and References

1 Petska, Tom and Kilss, Beth, "Recent Efforts To Maximize Benefits from the Statistics of Income Advisory Panel," paper presented at the Joint Statistical Meetings, San Francisco, California, August 2003.

2 The current irs.gov environment will not support online surveys, but that will be changing very soon. Once it does, the Web Advisory Group plans to survey Tax Stats users on a regular basis.

IRS Seeks To Develop New Web-Based Measurement Indicators for IRS.gov

Diane M. Dixon, Internal Revenue Service

In 1996, the IRS created and implemented its own website—irs.gov—to allow taxpayers easy access to IRS information and resources without having to call a customer service representative. Since the site's inception, the IRS has relied on web analytics to assess the site's usefulness and to make improvements to help enhance the customers' experiences and satisfaction. Serving customers and improving customer satisfaction within the diverse customer base of the IRS is a difficult task, but one to which the IRS is fully committed.

Although the goal is simply stated, there is no single approach to understanding the successfulness of a website or the level of satisfaction associated with it. With this in mind, the IRS has utilized several tools, including focus groups and customer surveys. However, in order to assess satisfaction on a large scale, the IRS has learned that understanding the underlying web activity is the key to designing a website that meets its customers' needs.

▶ Understanding Customers

As the customer demand for more functional websites increases, so does the need to understand how site usage affects an organization. There is a plethora of customer data that can be collected and used to interpret site usage. For some sites, a demographic customer profile is important because such information can help an organization define its market, which can aid in attracting new customers and generating revenue. However, simply collecting various demographics about customers will not result in a better understanding of customer needs. In order to understand customers, one must analyze customer behavior. Customer behavior data afford web administrators the ability to retain customers and predict future customer relationships.[1]

Making the decision to profile customer behavior is the first step; however, to make it work, an organization must first consider its specific needs in order to tailor results that will help with decision-making and planning. In order to choose measures that will be valuable, certain questions must be addressed, including, What is the website's purpose? How are website changes decided upon currently? What makes the website successful? Addressing these questions will help narrow down which measures will be most valuable to assess a site.[2]

▶ Introduction to Web Analytics

Following the evolution of technology, the way in which website traffic is analyzed has advanced greatly in the last decade. From primitive measures such as hit counts and files downloaded, web metrics have blossomed into a variety of different tools that are valuable both independently or combined into a suite of analysis tools. Depending on the data collection software, web administrators can collect the number of visits, unique visitors, and page views associated with a site, as well as various other web metrics, including path analysis, referral pages, and various customer demographics, while still collecting hits and downloads.

Since the launch of irs.gov, the IRS has recognized the importance of monitoring site activity. Using two of the most common web metrics at the time—hits and downloads—the IRS collected data to describe the web traffic on the site. In January 2002, the IRS launched WebTrends Reporting Center,® which gathers raw website data and transforms it into a collection of reports easily accessed via the Internet. WebTrends® has allowed the IRS to capture more site data, providing more insight to customer behavior. These additional metrics include visits and page views—displayed for a day, week, month, quarter, or year, depending on preference and need.

▶ Metric Analysis

Hits. A hit is a file that is requested by a visitor's computer. Each individual webpage consists of numerous hits—the HTML page itself counts as one hit, but each graphic or hyperlink is also interpreted as a single

hit. The amount of hits on each page is dependent on the page design.

The intended use of this metric is to measure website server workloads—how much stress is being placed on a server due to site usage. Depending on the size of the server, the amount of file requests could have a serious impact on the performance of the server, as well as the availability of the website. Therefore, knowing the volume of hits related to irs.gov is important to IRS information technology (IT) personnel. Using this data, they can assess server performance and make decisions concerning equipment needs.

As previously mentioned, the volume of hits is proportional to the design complexity of the website. Each individual page may consist of a varying number of hits, meaning graphic- or link-rich pages produce higher counts than simple pages which yield lower hit counts.

Due to limitations of the current version of WebTrends® running on the IRS system, hit counts for individual pages are not available. However, if these data were available, one would see the same number of visitors produce a higher amount of hits by visiting the "Where to File, By State" page of irs.gov than they would if they visited the "Retirement Plans—Educational Services Program" page.[3, 4] Both pages include all of the links contained on the top and left navigation bars, but the "Where to File" page has a graphic of the United States that contains 50 links, as well as a listing of each State, adding another 50 links. It also has a few other links and graphics. However, the page about the educational services program only consists of plain text and two other links. The significant difference in the number of links and graphics on the pages will notably alter the number of hits associated with each page, even if both pages are visited an equal number of times.

Downloads. A download occurs when a file is copied from the website server to the user's computer. Files are identified by their file extension (e.g., .xls is the file extension for a Microsoft Excel® file). Web analysts can program the software to count certain extension types so that they can filter out types of files that they do not want to include in the analysis.

For sites with numerous downloadable files, this metric can be extremely helpful in determining what is important to the majority of customers. This type of analysis can help site designers redesign navigation in order to guide customers to more popular files and products.

Downloads can also illustrate the effectiveness of recent marketing campaigns. By using historical data, analysts can calculate the increase in downloads for files promoted in campaigns and then determine the successfulness of the campaigns.

Using download counts, one can also determine which files are accessed often and which are not. This can help site designers analyze the setup of the current site. Files with the least number of downloads may be expected to be found at the bottom of the list, due to their age; however, if a designer expects more customers to access certain files that are currently not being accessed, the designer can alter the way in which these files appear on the site, to help improve accessibility. Then, using current and historical data, an analyst can determine if this change was helpful to customers.

This metric also allows analysts to see trends among the types of files downloaded during certain times of the year. Customers may want different information, depending on what month it is. This is certainly true for the IRS—the majority of IRS file downloads are predictable, following the filing seasons. However, there are portions of irs.gov, such as Tax Statistics, that are not as foreseeable. Files contained within Tax Statistics are produced by the Research, Analysis, and Statistics organization within the IRS. The way in which customers access these data files is not predictable. However, analyzing these data over time has allowed the web designer to better understand what customers want and when they want it. This knowledge has led to the discussion of designing navigation based on the time of year—using the landing page of Tax Statistics to spotlight certain data, making it easier for customers to locate desired information.

Visits. A series of actions that begin when a customer lands on his or her first page of the website and ends when s/he either leaves the site or remains idle for

more than 30 minutes is considered a visit. The "number of visits" may include multiple visits made by the same user.

In order for a visit to be tracked and counted, it is not necessary for a user to begin on the site's landing page. This is essential for many websites since many customers utilize the bookmark function for pages within sites that they visit often. Certain types of browsing behaviors, including jumping around a site, refreshing pages, and wrongly selecting pages, can greatly influence certain measures, leading to inaccuracies; however, these behaviors have no affect in the measurement of site visits. This ability makes the number of visits a valuable statistic to most website analysts.

Using this metric, a web analyst can determine how many visits are made to the site within a certain timeframe—an hour, a day, a month, a quarter, or a year. The number of visits can be analyzed historically to determine customer growth. Since it is possible to gather these data based on the time of day, this metric also allows IT personnel to determine the slowest periods of customer activity so that system upgrades and changes can be performed at a time that does not affect a large number of users.

Unique Visitors. Although visits are important in assessing a website, many businesses are interested in how many unique people are visiting their sites. Calculating this number allows a company to further determine the usefulness of its site.

These data are a further breakdown of the number of visits, allowing one to see how many individuals are behind those numbers. Tracked correctly, one could use this measure to determine the number of customers who visited a website within a certain timeframe. This differs from visits because, no matter how many times a customer visits within the timeframe, they are only counted once.

The ability to identify unique visitors also allows web analysts to assess repeat visitors, which further illustrates the usefulness of a site. The measure of repeat visitors may indicate satisfaction among customers, which may reduce or eliminate the need for an alternative site from which to obtain information.

Page Views. Each HTML page is tagged as a page. When a visitor accesses a page, it requests all of the hits on that page, including the page itself. In order to report the number of page views, the website analysis software separates the page hits from the other hits. These numbers make up the page view metric.

Much insight can be pulled from this statistic. One can assess which pages are accessed most, as well as those that are not. Although one cannot assume that the pages with the most views are the most useful to customers, these data can be useful during site redesign. If site owners have a general idea of what information is most appealing to customers, they will be able to determine if visitors are finding that information. Low page views for such pages could be an indicator of site navigational problems. This is similar to the information that downloads provide; yet customers need not download anything to obtain information concerning their interests. This measure equates sites with copious amounts of downloadable files to sites with few or none, thus allowing comparison between these two site types.

As with downloads, page views can also be helpful when determining if customers access types of information at certain times of the year, allowing for further navigational improvement.

Additional Metrics. Although the metrics described above do provide an immense amount of insight into customers' web-browsing behavior, there are other metrics that can further detail website usage, providing a more indepth understanding of one's customers.

Some software packages allow web analysts to track paths to certain information within a site. By monitoring these paths, analysts can determine if the site navigation is allowing customers to easily access information.

Another valuable tool is one that captures referring pages—the page the customer used to link to a site. Using this feature, one can determine which search engines are most popular among the majority of users. Web

managers can then contract with those engines to have their links appear closer to the top of certain searches. This metric also allows analysts to assess the success rate of certain partnerships with other sites, as well as whether or not that partnership should continue.

Although all previously mentioned metrics have focused on customer behavior, software can also collect demographic information about customers. This includes geographic regions, countries, cities, organizations, and domain names. Such information could be useful in various ways. Demographic information can help a site designer tailor a website to the audience. Understanding the audience and designing a site specifically for them will help attract customers and generate first-stage revenue.[5]

▶ Limitations with Web Analytics

Depending on website environments, policies, and restrictions, the usefulness of web analytics can be quite limited. Though the data might be insightful, analysts may not be able to fully appraise their sites using certain metrics, even in an unrestricted environment.

Interpreting Behavior. Complete interpretation of this data relies on some assumptions, which may not be reliable. For instance, the most downloaded file for a certain timeframe does not indicate that the file was useful to the customer, or even if it was what s/he was searching for. This concept also applies to other metrics, such as page views. Certain pages may be viewed frequently enough to appear in the listing of the top 50 pages viewed; however, this page may not be useful to most customers—it may even be an intermediate page that must be viewed before gaining access to any number of files. (For example, on the landing page of irs.gov, there is a link to the "Where's My Refund" feature. This link takes the visitor to an intermediate page that explains the information necessary to proceed. At the bottom of this page, there is another link that goes to the actual feature.) Using these assumptions, it is possible that a poorly-designed site could produce a significant amount of page views and downloads, which may lead some people to believe that the site is better than one that produces less because its navigation is better.

Another inherent problem is that web-browsing behavior can vary greatly among customers. Experienced Internet users may view fewer pages, download fewer files, and spend less time overall on a site. These users may also visit less frequently, as they may find everything they needed in one visit; whereas inexperienced customers may need to make several visits before finding everything. While web metrics may indicate otherwise, this behavior may not necessarily signify that their satisfaction with the site is lower.

Cookies. A cookie is a small text file placed on a customer's computer hard drive by a web server, usually unnoticed by the customer. This file allows the web server to identify individual computers—enabling a company to recognize returning users, track online purchases, or maintain and serve customized web pages. Cookies can also facilitate the collection of personal information, such as extensive lists of previously visited sites, e-mail addresses, or other information to distinguish individual customers.[6]

The Privacy Act of 1974 set regulations concerning the collection of personal information from a citizen.[7] Persistent Internet cookies are considered personally identifiable information and, thus, are covered by this Act. In 2002, the E-Government Act formally delegated responsibility to the Director of the Office of Management and Budget (OMB) to establish Government website policies.[8] However, even before the 2002 Act, OMB established a cookie-free policy, explained in Memorandum M-99-18.[9] In January 2002, the Department of Treasury clarified the policy, explaining that "persistent cookies shall only be granted when the bureau or office has presented documentation which details a compelling need to gather necessary data on the subject website." [10]

The inability to use permanent Internet cookies seriously restricts data interpretation. Without cookies, web analytic software must rely on Internet protocol (IP) addresses in order to collect data about customers. An IP address is a 32-bit numeric address written as four numbers separated by periods. This address is related to an Internet Service Provider's (ISP) server. Large ISP's, such as America Online (AOL) and the Microsoft

Network (MSN), have millions of customers sharing numerous servers, meaning that a single IP address may represent thousands of people. For example, if five AOL customers access irs.gov, they may be recognized as one, two, three, four, or five customers.

This notion has a serious effect on website data, especially since most IP addresses are dynamic (temporary) rather than static. This means that the majority of web users have a different IP address every time they visit a site. The problem is made worse by ISP's that allow a client's IP address to change with every new page, meaning that every page view will register as a new visit.

Caching. A cached file is one that has been previously stored on a system (e.g., a personal computer or an ISP server), making reuse of the page or object easier on the customer. When a visitor reaccesses a page or file that has been cached, their system accesses it from the cache location rather than the main web server that hosts the file. The objective of caching is to make efficient use of resources. Although this computer practice may positively affect a customer's experience when accessing a file (e.g., by significantly lessening the download time), it does negatively affect the site's web analytics, as hits, downloads, and page views of cached files will not be captured.

File Transfers. As mentioned previously, content- or file-heavy websites greatly rely on data concerning downloads. Such data can provide website owners with the best insight into understanding their customers. Sites with years of historical files, like irs.gov's Tax Statistics, are interested in understanding how downloads change over time, relying heavily on historical web analytics.

The problem with file transfers is that, depending on the software package, the way in which files are sent may differ. Some software packages allow all files to be sent as a single file—this is the ideal method of data transfer. With this method, 1,000 downloads correspond to 1,000 actual downloads. However, other software packages split a single file into multiple packets, each registering as an individual download, which greatly inflates the number of downloads reported. With this method, using the example above, 1,000 downloads reported represent the total number of packets sent, which corresponds to a much smaller number of actual files downloaded, depending on how many packets each individual file was split into. The latter method makes interpreting downloads more complex, leaving analysts to rely on other metrics to evaluate their sites.

▶ Educating Data Users

As explained above, the usefulness of web metrics can be severely restricted. Because of this notion, and a general confusion and lack of education surrounding web metrics, the IRS has begun an effort to educate website managers on definitions and usage of these metrics, as well as how certain limitations impact data interpretation. Only when there is understanding of data can it be utilized in such a way as to help improve the site and make more accurate interpretations of customer behavior.

To initiate this learning period, the IRS solicited information from members of the Web Facilitation Group (WFG)—a group of IRS employees responsible for setting IRS website policy—concerning how they use current data, the types of data wanted, how they plan to use additional data, what types of reports they generate using current data, and what types of decisions are made using web statistics. With this knowledge, the IRS will be able to determine the current level of knowledge among the WFG and decide where the education process should begin.

Future discussions with the WFG will focus on how irs.gov web statistics can and cannot be used to interpret customer behavior. Once members of the WFG have a better understanding of irs.gov data, they will be able to provide more accurate reports for their superiors and ensure that statements about the site are correct.

▶ Developing Detailed Reports

To aid the web statistics educational process, the IRS plans to develop new reports for irs.gov data. The new reports will contain a significant amount of annotation, allowing for easier and accurate interpretation of data. The IRS plans to develop individual reports for each

of the IRS business operating divisions (BOD's), as well as a report for all of irs.gov. By including definitions of certain measures, providing an initial data analysis, briefly explaining uses of each measure within these reports, and explaining the impact of limitations, the IRS hopes to help the BOD's make well-informed decisions concerning their respective sections of irs.gov, limit the amount of misinterpretation, and distribute the most accurate reflection of website usage.

▶ Upgrading Statistical Software To Improve Usability

In conjunction with the education effort, the IRS recently started researching new software options that offer additional functionality, as well as eliminate some of the limitations that currently hinder the interpretation of customer behavior. With the current version of WebTrends®, the IRS cannot generate metrics for individual BOD's. Instead, the software produces most data general to the whole site. As one would expect customer behavior to vary in each portion of the site, this makes customer behavior interpretation much more difficult.

Though most of the data generated by WebTrends® is whole-site-specific, the IRS can program the software to gather certain data specific to individual sections of irs.gov; however, this capability is still quite limited. With an upgrade, the IRS will be able to collect web statistics for various sections of irs.gov with ease. This new ability will aid the development of individual reports, as mentioned above.

▶ Conclusion

As one of the most powerful tools used to disseminate information, the Internet has created a world of faceless customers—people who seek information at their convenience. IRS.gov allows taxpayers 24-hour access to forms and filing information, which reduces the number of calls made to IRS call centers, changing the way in which taxpayers interact with the IRS. However, in order to sustain the success of this type of relationship, the IRS has to recognize the necessity of understanding web customer behavior.

By utilizing a software package to gather data on customer behavior, the IRS has been able to acquire, build, and sustain solid customer relationships without ever truly interacting with its customers. However, having these numbers alone is not the solution to interpreting customer behavior. IRS web analysts must understand the metrics, as well as the limitations associated with each. Education is a must when distributing reports about web analytics, as without such knowledge, misinterpretation of data is to be expected.

When utilized, analyzed, and interpreted correctly, web analytics can lead to a significant improvement in the usefulness and success of a website, allowing the IRS the potential to attract new customers, retain others, and maintain a high satisfaction rate among all.

▶ Notes and References

[1] Novo, J. (2002), *Drilling Down: Turning Customer Data into Profits with a Spreadsheet,* Bangor: Booklocker.com, Inc., Chapter 1.

[2] WebCriteria, "Executive Guide to Improving Website ROI: Questions Every Business Manager Should Be Asking," WebCriteria Online, available: http://www.webcriteria.com/company/white_papers/index.cfm [July 11, 2003].

[3] The IRS "Where to File, By State" page can be accessed at http://www.irs.gov/file/content/0,,id=105693,00.html.

[4] The IRS "Retirement Plans—Educational Services Program" page can be accessed at http://www.irs.gov/retirement/article/0,,id=96272,00.html.

[5] Novo, J. (2002), *Drilling Down: Turning Customer Data into Profits with a Spreadsheet*. Bangor: Booklocker.com, Inc., Chapter 1.

[6] An example of a well-known website that uses cookies is Amazon.com. After visiting Amazon for the first time and performing a simple search for one of its many products, Amazon will tailor its

main pages to better suit the customer's needs. This tactic is used with the hope of selling customers additional items that they may not have otherwise purchased.

[7] 5 U.S.C. § 552A (1996).

[8] H.R. 2458/S. 803.

[9] Lew, Jacob J. (1999) M-99-18: Privacy Policies on Federal Web Sites [Memorandum], Office of Management and Budget, available: http://www.whitehouse.gov/omb/memoranda/m99-18.html [July 11, 2003].

[10] The Department of Treasury (2002), 81-08: Certification Process for the Use of Persistent Cookies on Treasury Web Sites. [Treasury Directive], Available: http://www.ustreas.gov/regs/td81-08.htm [July 11, 2003].

Recent Efforts To Maximize Benefits From the Statistics of Income Advisory Panel

Tom Petska and Beth Kilss, Internal Revenue Service

The Internal Revenue Service's Statistics of Income (SOI) Division has had, for over 15 years, a Consultants' Panel whose membership consists of a distinguished group of individuals from academia, nonprofit organizations, State and local government, and the private sector. The Panel has always had a keen interest in helping SOI fulfill its mission by assisting SOI staff to improve its overall performance and in providing guidance and advice to make SOI ever more efficient, forward thinking, and responsive to its many customers in and outside of the public sector. In addition, the Panel has served as a management sounding board on issues, including strategic planning, data dissemination, and project prioritization.

For many years, this assistance was primarily as a result of periodic meetings in which SOI staff presented ongoing plans and operations to Panel members and invited guests from the public to solicit feedback, guidance, and direction. While these efforts were beneficial, both Panel members and SOI staff agreed that greater involvement in the core operations of SOI could be mutually beneficial. This paper is a progress report on how SOI has solicited greater involvement from its Panel members, what has been accomplished to date, and what approaches and initiatives are being planned for the future.

▶ Background Information

The SOI function goes back to the enactment of the modern income tax in 1913. In the 1916 Act, it was written that "the Secretary (of the Treasury) shall prepare and publish not less than annually statistics reasonably available with respect to the operations of the internal revenue laws." Despite many revisions to the tax law, the original requirement of that Act continues to this day.

The mission of the SOI program is to collect and process data so that they become meaningful information and to disseminate this information to customers and users. The SOI Division conducts for the Internal Revenue Service and Treasury Department studies on the operations of tax laws with respect to individuals, corporations, partnerships, sole proprietorships, estates, nonprofit organizations, and trusts, as well as specialized studies covering both inbound and outbound international activities.

The SOI Division has produced studies and published reports for over 85 years. However, around 1980, with the advent of modern computing and microsimulation modeling by policy and revenue estimation functions at Treasury and elsewhere, it became apparent that SOI had not kept up. In the first half of the 1980's, under the direction of then Director Fritz Scheuren, a major overhaul of SOI's methodologies and key business processes began. By mid-decade, several developments and accomplishments were of note:

- A renewed emphasis on quality that had come to the Internal Revenue Service was closely embraced within SOI.
- SOI began to attract many new technical staff who could help lead the retooling of projects.
- SOI began to develop its own "minicomputer network," replacing reliance on IRS mainframe technologies where statistical programs were a low priority.
- An SOI "Consultants Advisory Group" was formed to help guide and direct these efforts.

Minicomputers were placed in SOI's headquarters in Washington, as well as in two key field sites, Ogden, Utah and Cincinnati, Ohio. Pilot projects began to develop a one-pass approach to the complex data editing operations that were a substantial improvement over the multi-iterative approach used on the IRS mainframe computer systems. Each year, SOI staff made regular and continuous improvements to the systems, so that, over

time, SOI's data processing and analytical capabilities became quite sophisticated.

▶ Early Involvement by the Panel

The Consultants' Panel evolved from discussions and involvement with the late Joe Pechman, a scholar at The Brookings Institution, who saw the large public value in greater access to and dissemination of tax return data. In the early 1960's, SOI had developed its first public-use microdata file, a non-identifiable subset of the annual sample of individual tax returns, at Pechman's urging, as an invaluable tool for tax policy analysts outside of the Government who did not have access to SOI's rich data files. The success of the public-use file and the potential gain for the policy analyst community prompted him to host periodic meetings at Brookings and eventually suggest the formation of a formal advisory board in the mid-1980's.

The formation of an advisory group by IRS's Statistics of Income Division was not unique to the Federal statistical agency community. Most of the major Federal statistical agencies have a long history of using one or more advisory groups as a mechanism for inviting the participation of private citizens in their decision-making processes.[1] In 1986, the SOI Division was a relative latecomer to this arena and chose a less formal arrangement by forming an ad hoc group as opposed to an advisory committee operating under the provisions of the Federal Advisory Committee Act (FACA; Public Law 92-463, 92nd Congress, House of Representatives 4383, October 6, 1972). From the beginning, the SOI Division obtained advice from the individual members of its advisory group rather than from the group as a whole.

The SOI Consultants Panel formally began in the spring of 1986 with a general mission of use as a sounding board and a source of ideas and innovations. Planning sessions between Joe Pechman and then SOI Director Fritz Scheuren in preparation for the first meeting established the scope and character of the Panel. Initially, they determined that the main purpose of the Panel was to help shape the SOI program so its products would be given wider use in the research (academic/business/policymaking) community. The spring was chosen for the first meeting to coincide with the completion of SOI's multiyear planning process, and it focused on the current SOI program with special emphasis on future directions and changes. A follow-up meeting was suggested for the fall of that year because it would serve as a checkpoint on the degree to which SOI could incorporate changes into its plans in preparation for the next multiyear plan. It was also thought that a spring 1987 meeting would provide the forum for reviewing the new plans resulting from the above process and for deciding on the periodicity of meetings thereafter. Thus was born, 17 years ago, a framework for SOI to gain more systematic input about how well it was doing as an organization and how it might improve service to its customers that has become an integral part of the SOI culture.

Seventeen years ago, at the fall 1986 meeting, SOI concluded the session by giving all Panel members and participants an opportunity to identify those issues they felt the Division should be most concerned with. Many of the areas mentioned then were, and still are, a major focus of the SOI program, perhaps underscoring their significance, and yet also indicating SOI's need to continually work to improve what it does best. Some of the issues raised at that session that still resonate today include: risk of disclosure for microdata; public access and confidentiality issues; archiving and documenting historical files; the implications of electronic filing on the SOI program; and the development of more longitudinal data. It is also true that, while many topics have been repeated on the agendas over the years, there has also been much variety introduced into the sessions. Some noteworthy examples include conducting offsite meetings at two IRS service centers; inviting guest speakers; and organizing the session as a workshop. Many Panel meetings were organized around particular themes, some focused on technological innovations, while others included online demonstrations of SOI's computer systems.

The focus of the Panel meetings has included tax policy data needs, statistical disclosure research, computer modernization, microsimulation modeling, tax reform, and individual and corporation data. Early Panel meetings in the late 1980's and early 1990's regularly included updates on the SOI Division's individual, corporation, partnership, foreign, and special studies programs. These meetings frequently included presentations by Panel members on such topics as State tax sta-

tistics, household surveys with tax data, analyzing SOI panel data, individual tax model research, and economic statistics initiatives. Later, such topics as the earned income tax credit, tax gap, the Survey of Consumer Finances, use of SOI data in emerging tax issues, data sharing legislation, and data warehousing were covered.

▶ Panel Feedback and Input

While each Panel meeting agenda throughout the years comprised different, specific topics, which were interesting and educational, discussions of SOI programs or systems between SOI staff and Panel members were often the most revealing and beneficial. Though SOI periodically received much praise for the quality, usefulness, and importance of its data and the professionalism and caliber of its staff, it also recognized that it needed to do more, in some part because of the feedback and input from our Panel meeting discussions. As was hoped when the Consultants Panel was originally formed, ideas for improvements were presented, suggestions were made in a neutral environment, and general underlying themes kept recurring that eventually pinpointed a number of program shortcomings. One result was that SOI immediately undertook initiatives to address program limitations and deficiencies.[2] Former Director Fritz Scheuren and current Director Tom Petska, who was at the time Chief of the Division's Coordination and Publications staff, developed a list of items from these initiatives to present to Panel members for consideration and comment. These items were discussed as a "research and improvement agenda" at one of the Panel meetings in the early nineties. Discussions focused on tradeoffs among improvement priorities, and Panel members were polled for their own individual rankings. Five initiatives include the needs for greater program timeliness, improved data consistency, better tracking of demographic changes, preservation of historical information, and public access. These five are summarized below.

- **Timeliness.** The fact that users never have enough current information from tax returns is an inherent weakness of the SOI program. Timeliness of SOI studies has been a focus for improvement and one in which some success has been achieved. In all major SOI studies, there is an ongoing commitment to complete statistical processing within a minimum time after the close of the sampling period. Delivery dates have improved as a result. Preliminary data are also provided as early as possible.

- **Data Consistency.** Problems of data consistency are of two general types, statistical and conceptual. Despite extensive validity testing, inconsistent or erroneous data still escape undetected for a variety of reasons in some SOI data files. Efforts continue to rid these out of the system. Improving the conceptual clarity and year-to-year consistency of the content of tax and information returns is also a problem that has no easy solution. Where possible, efforts have been made to ensure consistency in time series data.

- **Tracking Demographic Changes.** The redesign of the individual program at the request of Treasury's Office of Tax Analysis (OTA) underscored the need to improve the longitudinality in SOI studies. Transactions such as capital asset realizations, that can have multiyear ramifications, can only be examined by means of a panel data base. A similar need for greater longitudinality also applied to business sector studies. Tax reforms, particularly those affecting individual and corporate tax rates, have increased the occurrence of changes of legal form, such as switching from a corporation to a limited partnership. Developing panel data in the individual and corporate areas has been a major focus of SOI work over the past 10 years.

- **Preservation of Historical Information.** Although current efforts are focused on better meeting current and future customer needs, SOI has become "keeper" of an abundance of tax information documents in a variety of media. Much of this information, though cumbersome to use, is irreplaceable. However, as new technologies become available, the cost of moving this information into more user-friendly formats will drop considerably. A difficult decision has been and continues to be how many current

resources should be diverted from present work to safeguard this historical information.

- **Public Access.** Tax returns are protected by law from public scrutiny, and strict procedures govern the handling of returns and computer tape files containing such information. Even after specific identifiers (e.g., name, address, and Social Security number) are removed, the remaining tax return data are usually still confidential. While SOI's primary customers are authorized to receive detailed tax return microdata files, other users may have only summary tabulations. Public-use microdata files of individual tax data have been produced regularly since 1960 and are the only source of certain information. An ongoing issue for SOI has been how to make more tax microdata publicly available to researchers outside of Government. This will continue to be studied in both the individual and corporate areas.

These items are all crucial to the growth, development, and success of SOI if it is to be considered a world-class statistical organization. Panel members' opinions on these topics during Panel meeting discussions over the years were certainly one of the factors that helped shape SOI's thinking and decision-making as the Division sought to make continual improvements to its programs.

▶ Panel Membership

With rotating membership, the Consultants' Panel has met virtually every year since 1986. The 10-15 members of the Panel represent academia, the corporate world, economic research centers, State governments, and nonprofit "think tanks." Attendees at Panel meetings include the members themselves, SOI staff, and invited guests. These include members of the Treasury's Office of Tax Analysis (OTA); the Congressional Joint Committee on Taxation (JCT); the Congressional Budget Office (CBO); the General Accounting Office (GAO); the Census Bureau; the Federal Reserve Board; and others from research organizations and academia. The daylong meetings are usually held at The Brookings Institution in the spring or fall. SOI reimburses Panel members only for travel and per diem; so, their advice and guidance are largely *pro bono*.

The Panel was originally chaired by the late Joe Pechman of Brookings and consisted of 12 additional members. Of those 12, the 4 who remain to this day are:

- Martin David, the University of Wisconsin and The Urban Institute
- Dan Feenberg, the National Bureau of Economic Research
- Gene Steuerle, The Urban Institute
- Bob Strauss, Carnegie Mellon University

In addition to the above, the current Panel membership has added the following members, all of whom have served at least 5-10 years:

- Bill Gale, The Brookings Institution and Panel Chair
- Steve Caldwell, Cornell University
- Virginia Hodgkinson, Georgetown University
- Tom Neubig, Ernst & Young
- George Plesko, MIT
- Joel Slemrod, University of Michigan
- Lin Smith, PricewaterhouseCoopers
- Phil Spilberg, California Franchise Tax Board
- Jenny Wahl, Carlton College
- Sally Wallace, Georgia State University

▶ Benefits from the SOI Consultants' Panel

Over the years, the SOI Consultants' Panel has become a critical part of the communication process between SOI and its customers. While other statistical

agencies, like the Census Bureau and the National Center for Health Statistics, hold major user conferences on a regular basis to receive input from their customers, SOI has chosen this small-scale and relatively inexpensive approach to keep its customers informed.

There have been many benefits to SOI from the Panel. These include:

- The Panel provides an opportunity to tap into an extensive knowledge base of tax experts, some of whom are regular SOI data users.

- SOI staff members have given presentations to the Panel on technological and methodological improvements in SOI programs or quality initiatives that affect SOI projects. These occasions have been valuable learning experiences for staff members and resulted in specific suggestions, which have led to further improvements.

- A continuous theme from the Consultants' Panel has been the need for more timely and electronically available data. In the early 1990's, this led to the development of the SOI Electronic Bulletin Board, the forerunner of the current Tax Stats on the IRS website.

- Demonstrations of online systems have led to improved understanding by users of how SOI data are processed.

- Discussions of statistical innovations by SOI staff have resulted in valuable comments that led to further improvements in SOI methodology.

- Panel members also strongly advocated the need for developing metadata systems, which more fully document a study's processes from start to finish.

- Input from microsimulation modeling experts has helped SOI to provide better data for its tax policy analysts at Treasury and the Joint Committee on Taxation.

▶ New Directions

All Panel members believe it is important to have public-use data on the functioning of the tax system and have given time and energy to ensure that SOI continually improves its capabilities to make available timely, high quality data from tax and information returns. Under the tutelage of new SOI Director Tom Petska, the Panel once again meets biannually, and members have been asked to get more involved in areas of SOI modernization.

As noted above, the SOI Division has produced annual Public-Use Files (PUF's) since the early 1960's, and, while there has been periodic and anecdotal feedback from PUF users on how SOI could best suppress the data to minimize analytical pursuits, the Division never had a formal PUF users' group. In the spring of 2001, a group was formed and, after 2 years, become an unqualified success. PUF data users have welcomed the opportunity to contribute to overall plans for disclosure suppressions.

The PUF users' group has six members from the user community, two of whom are Panel members. The success of the PUF Users' Group as a way to improve communications with users, to obtain users' advice, and to revise data files in a way most useful to data users is an excellent model for forming similar subgroups from the SOI Consultants' Panel membership. As a result, SOI Director Petska decided to seek additional Panel involvement and assistance to streamline SOI operations in four additional areas. His expectation was that every Panel member will become a member of one of these subgroups and help SOI explore possibilities for systematic improvements in its key operations. These areas are:

1. Modernizing SOI's website to efficiently disseminate data;

2. Guiding research in estates and gift taxation and personal wealth;

3. Improving SOI's publications and tables; and

4. Advising on how to improve training.

Each of these areas, and the roles for assistance from the Panel members, are described below.

1. **Web Modernization Team.** In the fall of 2002, an inhouse team of SOI's Internet "visionaries" was commissioned to scope out the best capabilities and Internet features and a new look for SOI's website, "Tax Stats." This team sought support from members of the Consultants' Panel as resources in making improvements to the website. One initial task was to visit the 60+ websites listed in FEDSTATS, the Federal agencies' primary source point for statistics, to scope out best practices and then broaden the search. The SOI goal is to implement the group's proposals by redesigning the SOI website. This team currently consists of nine SOI staff members and three additional Panel members who are familiar with tools, capabilities, and features of state-of-the-art websites to help this effort achieve SOI's goal of making Tax Stats the best website in its class.

2. **Evaluation of SOI Table Content and Publications.** SOI has a long history of publishing since its original mandate in 1916. Today, the Division publishes the quarterly *Statistics of Income Bulletin,* the annual Individual and Corporation complete reports, the annual *Corporation Source Book*, the annual report in the Methodology series, and the annual IRS *Data Book.* In addition, SOI publishes periodic compendiums and, most recently, the proceedings for the newly established annual IRS Research Conference. A tremendous amount of time and effort goes into publishing these reports, but considerably less time has been spent evaluating the content, frequency, and dissemination of the publications. Some of the tasks that a subgroup plans to undertake are: review content and frequency of all SOI publications, examine how to make them more useful, look at methods of advertising and disseminating, and look at what is not being published that perhaps should be, e.g., new types of *Bulletin* articles. A standing committee of senior SOI staff, working with three Panel members, has been formed to help anticipate these needs and make data more useful to a wider audience of researchers and practitioners. Expanding the regular statistical content of publicly available data in publications and/or the website would make SOI data more useful to a broader audience, and also eliminate needs for "ad hoc" data requests, which can be disruptive.

3. **Research in Estates, Gift, and Wealth.** The focus of research in the estate, gift, and wealth areas, including SOI support of the Federal Reserve Board's Survey of Consumer Finances (SCF), is closely tied to the needs of the Office of Tax Analysis and the Joint Committee on Taxation. For the SCF, a contractual agreement between the Federal Reserve Board and SOI regulates the use of administrative data and protects individuals from disclosure of their financial and tax data. However, it is beneficial to review the scope and direction, as well as the item content, of these areas of research. One Panel member already works with SOI staff members on the estate, gift, and wealth team. Three additional Panel members have recently joined this group to look at the recent body of work in these areas and help provide insights to SOI on the focus of this work. The group is also interested in exploring innovative ways to make these data more valuable and more available, not only to Treasury and the Joint Committee, but also to outside users.

4. **Teaching and Training SOI Staff.** Periodically, SOI has hosted invited speakers describing the importance of SOI data and how they use them in forecasting and economic or policy analysis. The two areas that have done this more systematically are SOI's new employee orientations and infrequent formal training classes. Concerning new employee organization, some years ago, a series of a dozen orientation briefings was developed for all new employees that concluded with presentations by SOI's principal external customers at Treasury, the Joint Committee on Taxation, and the Bureau of Eco-

nomic Analysis. Although new employees may not have been ready to grasp the complexities of the work of these agencies, other SOI employees asked to attend. Also in the early 1990's, a 10-part course in Public Finance was designed and taught by invited senior public policy analysts (including some Panel members) to lecture on their work in public finance practice and how it relates to theory. What this group seeks is to add Panel members to a new inhouse staff, initially to periodically brief SOI on policy analysis using tax and other microdata. Next, working with SOI staff, training needs in tax law, policy analysis, the Federal statistical system, and statistical project management will be assessed, and inhouse training modules and short courses will be developed.

With the four new subgroups, long-term plans are to encourage frequent, periodic meetings, as needed, of the subgroups but to host semiannual meetings for the entire Panel. As requested by some members, periodic reports from the subgroups will be distributed to all Panel members at least once between the semiannual meetings. The general Panel meetings have traditionally been open to the public and widely announced. However, as the workings of the small subgroups progress, some thought is being given to restricting the general meeting, which is open to the public, to once per year so that more attention can be given to improvements to SOI's internal operations and policies.

▶ Summary and Conclusions

The SOI Consultants' Panel is a capable and energetic group of distinguished tax scholars, policy analysts, academics, and researchers in the public and private sectors who have generously offered their assistance to improve SOI operations. Not to accept their offer of assistance would be a travesty. This paper is an interim report on how this work has been structured, with an expectation of tangible benefits in the not-too-distant future. As it learns of these initial attempts, SOI plans to refocus the talents of the Panel members to other aspects of SOI operations.

▶ Notes and References

[1] Eldridge, Marie D., "The Status of Advisory Committees to the Federal Statistical Agencies," *The American Statistician,* May 1990, Volume 44, Number 2, pp. 154-162.

[2] Petska, Tom (1995), "Statistics on Federal Taxation: The Statistics of Income Program of the IRS," *Turning Administrative Systems Into Information Systems: 1994,* Internal Revenue Service.

4

Survey Nonresponse and Imputation

McMahon

Regulatory Exemptions and Item Nonresponse

Paul B. McMahon, Internal Revenue Service

The regulations referred to in the title are those governing the filing of tax returns with the Internal Revenue Service. Some of the rules for filing the various forms permit item nonresponse if some set of conditions is met. For example, one need not report itemized deductions when claiming the Standard Deduction on the Individual Income Tax Return.

These regulations affect all of the electronic records derived from the tax filings; so, other Federal agencies that use extracts from the Service's Master Files to enhance, for example, their sampling frames are also affected. The impact of such regulations is more pronounced for the Statistics of Income programs, because they use these administrative records both for a sampling frame and as the source questionnaires for the studies. Thus, rules that permit nonreporting of various data may affect not only the sample design but the sample's estimates as well.

We will examine one such exemption that applies to partnerships, and, as with the itemized deductions, the exemption applies only to certain schedules, on asset holdings. This is an issue because a similar exemption has just been introduced for corporations.

▶ Background

The Statistics of Income Partnership study focuses on businesses that can have limited liability, like corporations, and be traded on the stock exchanges, like corporations, but are not corporations. One reason a firm might not incorporate is that, in its line of business, the State prohibits that form of organization. The States, after all, hold domain over the rules for incorporation, not the Federal Government. This leaves us with only a very general description of the population, beyond the requirement that they file a Form 1065, *Partnership Return on Income*, with the Internal Revenue Service.

That form is not a tax return, however, for partnerships are rarely taxed as an entity. Rather, the earnings, deductions, and tax credits flow through to the owners who are taxed. This might not be a direct linkage, though, for the owners can be other partnerships.

The chaining of groups of partnerships and corporations, trusts and individuals, and the allocation of the incomes, credits, and deductions raises interesting tax administration issues. The Department of the Treasury's Office of Tax Analysis and Congress's Joint Committee on Taxation use the microdata from the various Statistics of Income studies to evaluate the laws and revisions; so, these data from the tax forms are irreplaceable for their purposes. However, the Service does not provide, nor have these sponsors requested, imputed values for missing items on those microdata files.

The published tabulations[1] from this series of studies have two different audiences: advocates for various tax law modifications, and economic analysts. In the first case, there is a need to ensure that the advocates have the same benchmarks as our sponsors. This leads us to publish data that are uncorrected for missing data.

When the data are used in economic analysis, where only summary data are available, the pattern of missing information can be disruptive. When the magnitude of the unreported data, for example, varies over the years or is a large proportion of the "true" amount, estimates of rates of change or financial ratios can be mistaken. In this case, the filing rule allows companies that meet certain conditions to avoid reporting their assets on their balance sheets.

The original version of the balance sheet exemption, 20 years ago, had seven conditions to be met, including being in a selected industry, having 10 or fewer partners, and the relationships among the partners (both with respect to interest in the firm and its profits, and as family). This complicated and constrained balance sheet filing exemption led to only a relative handful of firms responding that they met all the various tests. Thus, the effect on the resultant statistics was too small to even

get a reliable measure of its size for Tax Years 1983 through 1990.

This exemption was relaxed and simplified for Tax Year 1991, requiring only that both receipts and assets were less than $250,000 (and that the Schedule K-1's were filed timely). Then, 2 years later, the current version, labeled Question 5 on Schedule B of the return, was introduced:

> ***"5.*** *Does this partnership meet* ***ALL THREE*** *of the following requirements?*
>
> ***a.*** *The partnership's total receipts for the tax year were less than $250,000;*
>
> ***b.*** *The partnership's total assets at the end of the tax year were less than $600,000;* ***AND***
>
> ***c.*** *Schedules K-1 are filed with the return and furnished to the partners on or before the due date (including extensions) for the partnership return."*

While "total assets" is well defined (at least five places on the form have a total assets value), there is no single reference to "total receipts." For Tax Years 1991 through 2001, no definition of this amount was provided, either on the form or in the instructions. The current edition of the instructions for Form 1065, though, provides a detailed computation[2] that requires 17 amounts from three schedules, which in turn reference still other forms and schedules. When this definition of total receipts is retroactively applied to the records in Tax Year 1998 through 2001 Studies, as shown in Figure 1 below, 65 percent to 70 percent of those who appear to meet the conditions for the exemption file a completed balance sheet anyway. Thus, there is sufficient response for us to estimate the difference between the published estimates and one adjusted for nonresponse.

Figure 1. Partnerships With Total Receipts Less Than $250,000 and Assets Less Than $600,000, Tax Years 1998-2001

	Tax Year			
	1998	1999	2000	2001
Exempt and Assets 0	356	342	359	348
Reported Assets	686	726	772	787
Assets 0, Nonexempt	39	34	34	34
Final Filings	150	157	152	155

(All estimates in thousands of returns filed.)

If one were to look only at the presence or absence of the balance sheet information among those records that meet the criteria for the exemption, then about half would be without those data. But about 12 percent are final reports (the companies ceasing business); so, their assets are zero by definition. Moreover, another 2.5 percent to 3 percent did not claim the exemption, yet reported no assets. We are inclined to believe that these reports are true, for there are cases where the partners bring their own tools to the job, and there are no jointly-owned properties in those companies.

In adjusting the estimates for the missing asset information, the final filings are considered to be outside the adjustment classes, the same as firms with large assets or receipts. Firms that did not claim the exemption yet had no assets were placed with those reporting balance sheet amounts.

There are a handful of records that do not meet the requirements for the balance sheet exemption, using the definition for Total Receipts found in the Tax Year 2002 instructions booklet. These cases are believed to be coding errors that occurred during data abstraction because, in all cases, the balance sheets were reported. This suggests that there are those in the adjustment classes who reported assets and answered Question 5, "yes." In these cases, we simply ignored that false "yes." (The verification procedures were modified, and this sort of error should now cease to appear.)

▶ Effect on Strata

The goal in creating strata is to form groups that are relatively homogeneous. This reporting regulation creates implicit boundaries within the population that, if ignored, could create heterogeneous strata with respect to a key set of data. Unfortunately, not all of the items needed to compute "total receipts" are available on the sampling frame, though all of the major components are present. To the extent possible, then, a proxy for that total receipts amount is computed, and the limits set by Question 5 are explicitly incorporated as strata boundaries.

The outline of the strata is shown in Figure 12 (after the footnotes). This design has strata below the boundaries of the area defined by the exemption. Those lower receipts categories are incorporated in the creation of the adjustment cells. Real Estate firms, more than a third of the population, are separately stratified, and, since there is a connection between industry and the allocation of assets among the balance sheet categories, this classification is also respected in choosing the cells.

This outline can only be followed so far, however, because the change to the North American Industry Classification System (NAICS) required a change in the industry groups used in the design,[3] starting with the Tax Year 2001 study. For non-real estate returns, NAICS industry divisions were used, even though they sometimes crossed the major stratification boundaries for the studies of Tax Years 1998 through 2000.

▶ Adjustment Procedure

The balance sheet exemption nears the border between item and unit nonresponse, in that while we are concerned with records that are mostly complete (with all the income and expense items reported), the items missing are contained on a schedule that is separable from the rest of the report. That is, few of the asset items are the results of computations reported on other parts of the return, and the calculations on the balance sheet affects no other schedule.

The goal is to assess the magnitude of the understatement caused by the reporting exemption in the published tables. Thus, viewing the balance sheets as a separate sample, the appropriate nonresponse correction policy is a weight adjustment strategy:

$$\hat{Y} = \sum w_i a_c x_{ijc}$$

where $w_i = N_i/n_i$, is the sampling weight, and a_c is the item nonresponse adjustment factor for class "c." This factor is:

$$a_c = \begin{cases} 0 & \text{if exempt and assets } 0 \\ 1 & \text{if not in an adjustment class} \\ \hat{N}_c / \hat{N}_{cr} & \text{otherwise} \end{cases}$$

An adjustment factor of 1 is assigned to final filings and those companies with total receipts or asset values that exceed the regulation's limits. The rest were divided into classes depending on the size of total receipts, using the strata boundaries to the extent possible, and the NAICS industry division, as noted above.

The operating assumption is that the exemption claimants have the same distribution as the respondents within the adjustment cells, with respect to their assets; so, we used the estimated populations ($\hat{N}_c$ and $\hat{N}_{cr}$ for the cell total and respondent populations, respectively) in computing the adjustment factors. Within the various adjustment cells, the sampling weights varied considerably, in one case from a low of near 5 to a maximum of over 250 (with the weights approximately equal to the inverse of the probability of selection).

Figure 2. Weight Adjustments for Balance Sheet Data

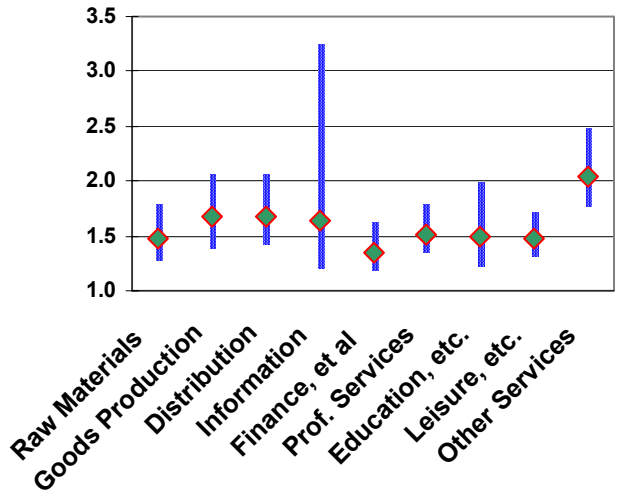
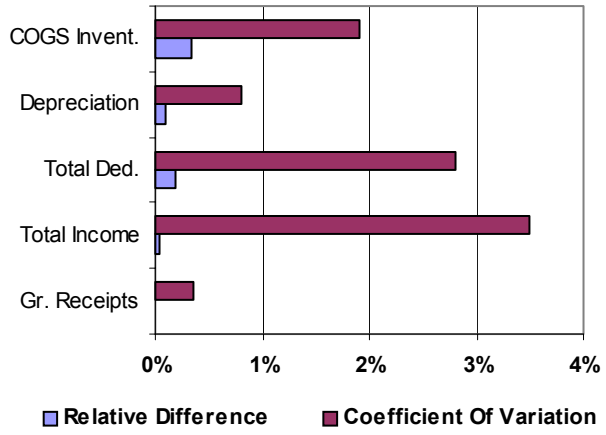


Figure 2 combines the adjustments for the 4 years to give a feel for the distribution of the factors. The factor for the Information Industry Division stands out, even though the average for that group (indicated by the lozenge) is quite reasonable because of the wide spread of the factors over the years. This is a small sample-size effect in the years after the conversion to NAICS, for, at the time the design was set, we had no usable data on the industry distributions.

▶ Validation of Adjustments

Do these adjustment factors provide reasonable estimates? The rule on not reporting selected data applies only to the Balance Sheet items; so, by computing alternate estimates for, say, income statement data, one can get a good measure on the reliability of this procedure, particularly if the items are somewhat related to balance sheet data.

Figure 3. Selected Estimates, Tax Year 2001

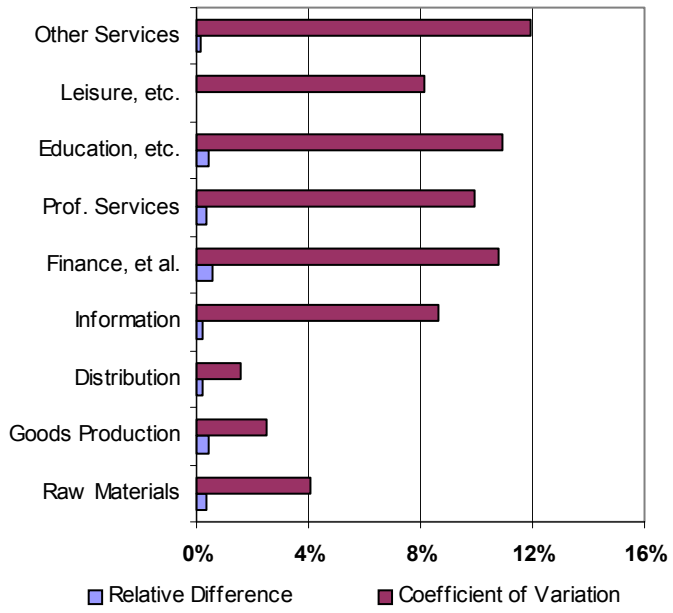


As seen in Figure 3, the absolute value of the ratio of the estimates under the adjustment procedure to the full sample estimates compares favorably to the relative errors at the national level. Cost of Goods Sold (COGS) Inventory and the Depreciation Expense are related to Inventory and Accumulated Depreciation on the balance sheet, respectively, but only comprise a part of those assets.

National comparisons can hide significant problems in critical subpopulations. Yet Figure 4 demonstrates, that, for COGS Inventory at least, the adjustments are very close to the full sample estimates for each of the industry divisions.

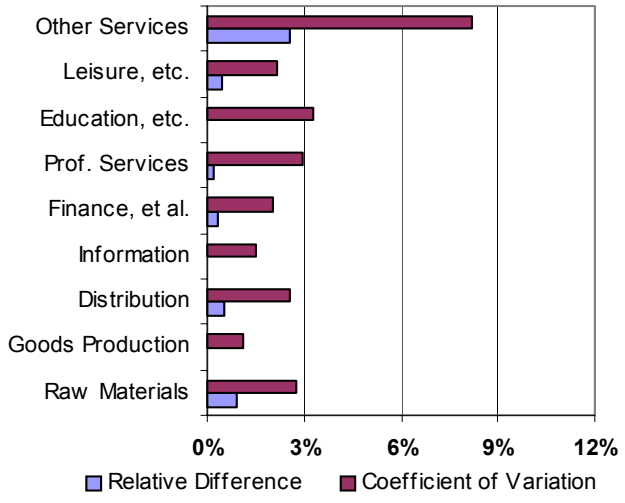
The scale for the Depreciation Expense, in Figure 5, is set to agree with that for Inventory, above. The Coefficients of Variation here are generally smaller because there is a greater dominance effect on the estimates by firms in the certainty strata. This effect is also apparent on the relative differences between the original figures

Figure 4. Cost of Goods Sold Inventory by Industry Division, Tax Year 2001



and the adjusted data. The exception is the division "Other Services," which has a small population and sample, as well as generally lesser amounts of total assets on average. These factors also affect the differences between the adjusted estimates from the respondents and the full sample estimates.

Figure 5. Depreciation by Industry Division, Tax Year 2001



Since the adjusted estimate for Other Services is still within 3 percent of the full sample estimate (and all the other data fall much closer to the mark), this method appears viable for the purpose of getting some measure of the size of the balance sheet estimates' understatement.

▶ Question 5's Impact

The Balance Sheet, shown in Figure 6, has two sections: the upper portion, which details the Asset holdings, and a smaller part on Liabilities and Equity. In the first part, there are four items that, though they are presented as positive values in the table, are subtractions from the total. These amounts, indicated by parenthesis, are: Bad Debts, Accumulated Depreciation, Accumulated Depletion, and Accumulated Amortization.

The two sections are, by accounting definition, equal, which is why we show the amount "Total Assets" in the break between them. The columns labeled "Relative Change" show the amount of the difference between the original and adjusted estimates as a percentage of the original estimate.

Although the size of the relative change is fairly small, particularly for Total Assets, there is little doubt that it is significant, as Figure 7 demonstrates. The increase in the coefficient of variation for Tax Year 2001 is the re-

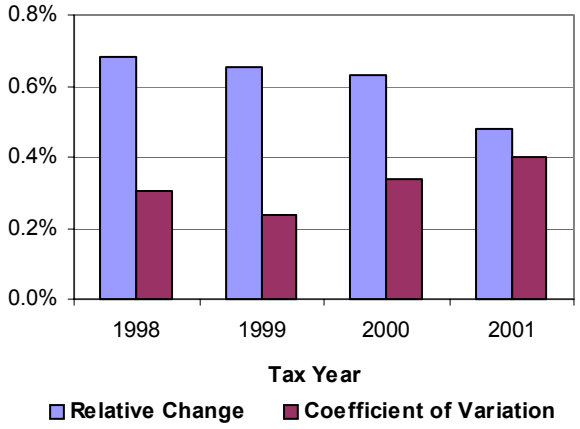
Figure 6. Adjusted Balance Sheet Estimates, Tax Years 1998 – 2001

	Tax Year 1998		Tax Year 1999		Tax Year 2000		Tax Year 2001	
	Adjusted Estimate	Relative Change	Adjusted Estimate	Relative Change	Adjusted Estimate	Relative Change	Adjusted Estimate	Relative Change
Assets								
Cash	185,162	1.82%	221,250	1.67%	267,031	1.64%	345,715	1.10%
Accounts Receivable	343,538	0.21	392,844	0.20	432,881	0.17	544,377	0.20
(Bad Debts)	6,194	0.75	7,478	0.01	9,494	0.06	12,027	0.39
Inventories	177,405	0.82	175,762	0.97	151,509	1.09	209,615	0.70
U.S. Obligations	95,784	0.03	79,280	0.05	72,952	0.14	156,399	0.04
Tax-Exempt Securities	28,132	0.03	23,158	0.04	26,304	0.08	33,500	0.01
Other Current Assets	700,299	0.30	828,183	0.27	837,555	0.26	1,261,821	0.18
Mortgages & Loans	52,239	1.86	48,798	1.82	61,052	1.11	71,778	0.84
Other Investments	1,586,214	0.26	1,980,991	0.26	2,281,339	0.26	2,890,034	0.20
Depreciable Assets	1,755,731	1.42	1,986,825	1.33	2,216,418	1.22	2,443,007	1.07
(Accum. Depreciation)	610,346	2.12	659,283	1.97	715,152	1.80	782,651	1.57
Depletable Assets	43,673	0.97	44,911	0.88	53,898	0.66	57,061	0.44
(Accum. Depletion)	18,308	0.92	14,790	1.51	16,146	0.97	17,182	0.76
Land	298,916	2.66	335,320	2.74	368,214	2.67	400,417	2.12
Intangible Assets	193,942	0.50	240,672	0.41	309,273	0.37	354,341	0.34
(Accum. Amortization)	52,522	0.66	55,676	0.66	66,971	0.45	81,126	0.52
Other Assets	367,838	0.42	417,278	0.42	465,767	0.41	593,507	0.35
Total Assets	5,161,503	0.68%	6,038,045	0.65%	6,736,429	0.63%	8,468,455	0.48%
Liabilities and Capital								
Accounts Payable	191,709	0.53%	245,213	0.59%	230,843	0.41%	362,413	0.18%
Short- Term Debt	233,044	1.36	235,057	1.40	255,593	1.33	292,238	1.03
Other Cur. Liabilities	935,377	0.46	966,930	0.46	927,837	0.43	1,578,613	0.20
Nonrecourse Loans	524,503	0.21	583,553	0.24	640,878	0.23	701,254	0.20
Long-Term Debt	896,685	1.38	1,000,853	1.23	1,144,654	1.10	1,298,752	0.96
Other Liabilities	399,503	2.09	449,410	1.15	522,613	0.91	630,073	1.22
Partners Cap. Accts.	1,980,682	0.25	2,557,030	0.44	3,014,010	0.51	3,605,113	0.33

(Amounts are in millions of dollars.)

sult of a smaller sample size arising from resource constraints. The change in the adjustment does not have an obvious source, on the other hand, though it seems connected to late filing firms of the sort that usually report losses.

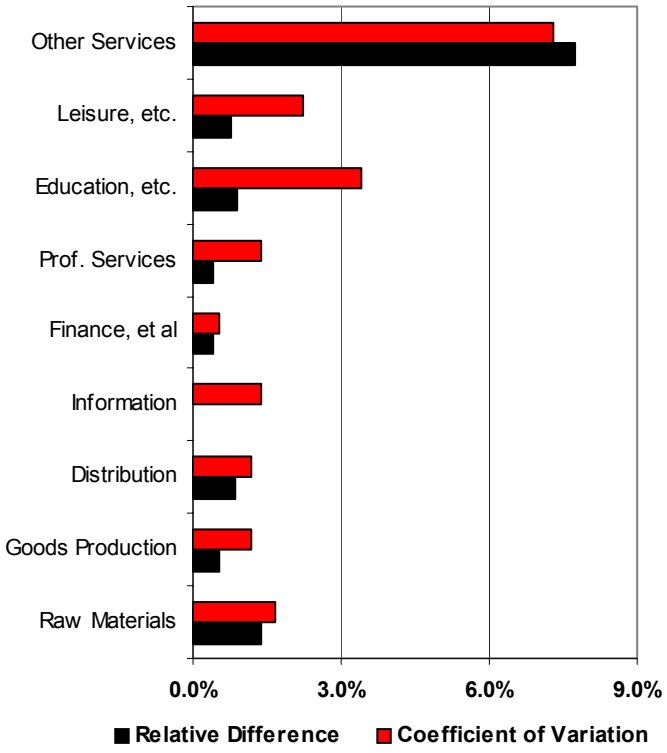
Figure 7. Relative Adjustment and Coefficients of Variation for Total Assets, Tax Years 1998-2001



At the same time, the general sizes of the relative adjustment and coefficient of variation are quite close, and small. This pattern of the close sizes appears to continue in the industry division estimates, as shown in Figure 8. The reason for this lies in the dominance of the largest firms. Such companies are selected with certainty for the sample and, hence, contribute nothing to the sampling error while reducing the coefficient of variation. Similarly, all of these firms have attributes that mean they do not meet the conditions set forth in Question 5; so again, the dominance reduces the effect.

The clearest example of this is in the Other Services and Finance Divisions. In the first case, Other Services, we have a small division without large firms. As a result, both the sampling error and adjustment are large compared to the estimate. The Finance Division, on the other hand, is dominated by firms with large amounts of assets and contains most of the partnership population. As a result of that dominance and size, the data for the Finance Division appear to have little significance in Figure 8. The values for both the adjustment and the coefficient, however, are very close to that for the all industries coefficient of variation and adjustment for Total Assets, demonstrating the inverse relationship in these data between the nominal size of the ratios presented and the importance of the underlying data.

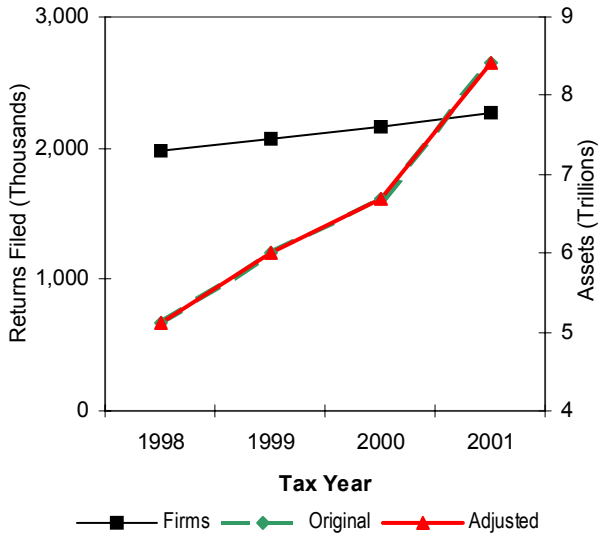
Figure 8. Adjusted Total Assets, by Industry Division, Tax Year 2001



Figures 6, 7, and 8, address the relative size of the adjustments. The size has an impact on ratios of estimates within a tax year, as is sometimes used in financial and accounting environments. The main purpose of the Statistics of Income data series, however, is to provide economic information, particularly on the effect of changes to the tax laws. In this situation, it is not the size of the adjustment itself that matters, but whether there is a large effect on the estimates of change.

When considering the estimates of change, one must bear in mind that the number of partnership returns filed, our population, has increased by a nearly constant 5 percent per year. The amount of total assets, on the other hand, has increased even faster, between 12 percent and 25 percent per year, as illustrated by Figure 9.

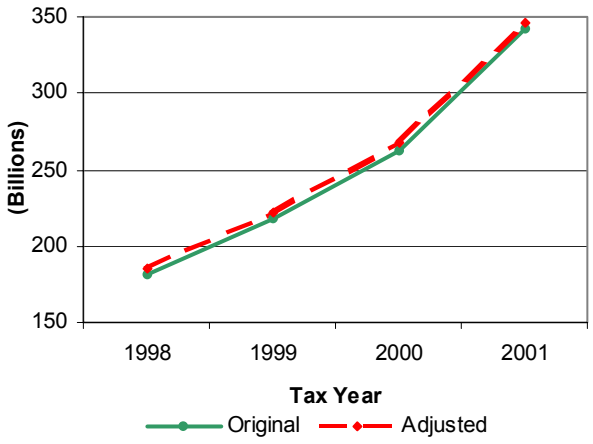
Figure 9. Change in Assets and Population, Tax Years 1998-2001



That figure, above, also shows the difference, or rather the lack thereof, between the original and adjusted estimates. On this scale, the difference between the two is barely discernible. This is not unexpected, for the relative differences are quite small and in the same direction (always greater).

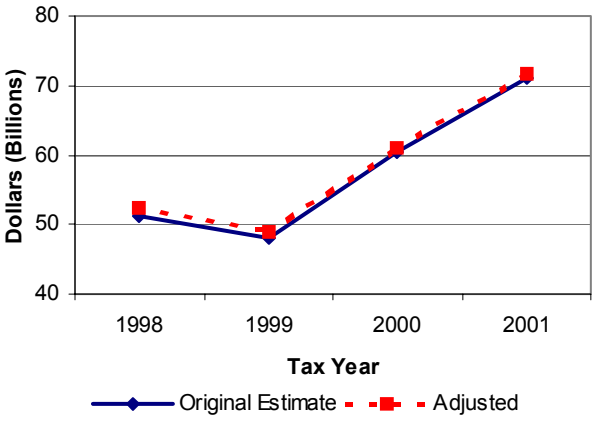
Both the scale required and the relative nearness of the two sets of estimates conspire to make the differences appear as they do. Perhaps better resolution could be obtained with smaller estimates where the departures are the greatest.

Figure 10. Cash, Original and Adjusted Estimates, Tax Years 1998-2001



Yet with the estimates for Cash, in Figure 10, we again see no real differentiation.

Figure 11. Estimated Mortgages and Loans, Original and Adjusted Estimates, Tax Years 1998-2001



This also holds true for the most extreme case, Mortgages and Loans, as seen in Figure 11.

▶ Conclusions

The method of weighting the balance sheet respondents is a reasonable procedure, given the response rate and the constrained circumstances of Question 5. The adjusted estimates of nonbalance sheet items from exempted firms, when compared to those from the full sample, lend credence to this adjustment strategy by the close agreement of those figures.

The adjusted balance sheet estimates are not greatly different from the original data, largely due to the dominance effect of the largest firms, but the differences do indicate a significant bias, as they are at least the size of the coefficients of variation. This bias is relatively constant; so, trends do not appear to be affected. However, the few years for which data are available suggest that this issue bears watching.

There are no plans to adjust the estimates the Service publishes to correct for these understatements, both

because the adjustment amounts for each item appear to be reasonably constant, and because the uncorrected totals provide a benchmark to external users of the data who review estimates from either the Office of Tax Analysis or the Joint Committee.

Nevertheless, we are considering adding a table to the annual publication comparing the full sample estimates to the adjusted results, mostly for the use of those researchers who focus on investment type ratios.

It is clear that, while the administrative systems do provide a very good source for population data, one has to be cautious about the existence of filing rules that can affect both sample designs and subsequent analysis.

▶ Footnotes

[1] Internal Revenue Service, *Statistics of Income Bulletin*, Fall 2002 (or other Fall editions), Washington, DC.

[2] Total receipts is the sum of:

Form 1065, pg .1: Gross Receipts, Ordinary Income From Other Partnerships, Net Farm Profit, Net Gain or Loss From the Sale of Business Property, and Other Income;

Schedule K: Non Real Estate Rents, Interest Income, Ordinary Dividends, Royalty Income, Short Term Capital Gains, Long Term Capital Gains (Taxed at the 28 Percent Rate), Other Portfolio Income, Income Under Section 1231, and Other Income;

Form 8825: Gross Real Estate Rents, Net Gain or Loss From the Sale of Business Property, and Income From Other Real Estate Partnerships.

[3] McMahon, Paul (2000), "Changing Industry Code Systems: The Impact on the Statistics of Income Partnership Studies," *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association.

Figure 12. Partnership Sample Design and Sampling Rates, Tax Year 2001

Extreme and Special Cases:

Total Assets $250,000,000 or more, or Receipts or Net Income $50,000,000 or more 100%

Publicly Traded Partnerships or Firms With 100 or more Partners 100%

Total Assets 100,000,000 Under 250,000,000 and Receipts or Net Income Under 50,000,000, or
Total Assets Under 100,000,000 and Receipts or Net Income 25,000,000 Under 50,000,000 . . . 35%

Real Estate

	Absolute Value of Receipts/Income ($)						
Assets ($)	Under 50,000	50,000 under 100,000	100,000 under 250,000	250,000 under 500,000	500,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 under 25,000,000
Under 250,000	0.12%	0.20%	0.30%	{	1.50%	}	
250,000 under 600,000	0.17	0.19	0.30	{	1.10	→	}
600,000 under 2,500,000	{ 0.27 }		0.35	0.50	{ 1.50 }		10%
2,500,000 under 5,000,000	{	0.50	}	0.80	0.90	1.90	
5,000,000 under 25,000,000	{	1.00	}	1.00	1.70	2.50	____
25,000,000 under 100,000,000	{			7.0%		}	15%

All Other Industries

	Under 40,000	40,000 under 100,000	100,000 under 250,000	250,000 under 1,000,000	1,000,000 under 2,500,000	2,500,000 under 5,000,000	5,000,00 under 25,000,000
Assets ($)							
Under 200,000	0.35%	0.50%	0.75%	0.12%	{ 3.8%	}	
200,000 under 600,000	0.40	0.80	0.95	1.40	{ 2.50	}	
600,000 under 2,000,000	{ 0.65 }		0.95	1.80	3.00	4.50	14.%
2,000,000 under 5,000,000	{ 1.50 }		2.50	3.00	{ 6.00	}	
5,000,000 under 10,000,000	{	2.50	}	3.00	5.00	6.50	
10,000,000 under 25,000,000	{	5.00	}	{ 6.00	}	10.00	____
25,000,000 under 100,000,000	{			14.%		}	30.%

Information, and Health, Education and Social Services

	Under 40,000	40,000 under 100,000	100,000 under 250,000	250,000 under 500,000	500,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 under 25,000,000
Assets ($)							
Under 150,000	0.35%	0.90%	1.50%	1.50%	{ 3.50%	}	
150,000 under 600,000	{ 3.00 }		20.0	{ 3.00 }		4.00	
600,000 under 5,000,000	{ 4.00 }		12.0	{ 3.00 }		7.00	13.%
5,000,000 under 25,000,000	{	25.0	}	{ 20.0 }		7.00	
25,000,000 under 100,000,000	{			40.%		}	30.%

Index of IRS Methodology Reports on Statistical Uses of Administrative Records

Special Studies in Federal Tax Statistics--2002
Selected papers given primarily at the 2002 Annual Meetings of the American Statistical Association in New York City and at the 2002 National Tax Association Conference in Orlando, FL. The volume is divided into seven major sections. It begins with two papers on recent IRS research. Section 2 includes a group of four papers on methodological and analytical advances in tax statistics. Section 3 presents two papers on statistical uses of administrative records. Section 4 contains a paper on disseminating IRS locality data. Section 5 includes a paper on confidentiality and data access issues. Section 6 presents a paper on measuring the quality of IRS responses to taxpayer inquiries. Finally, Section 7 includes two papers on distributional theory and computation.

Special Studies in Federal Tax Statistics--2000-2001
Selected papers given primarily at the 2000 and 2001 Annual Meetings of the American Statistical Association in Indianapolis, Indiana and Atlanta, Georgia, plus one other paper presented at the International Conference on Establishment Surveys II in Buffalo, New York in 2000. The volume is divided into four major sections. The book begins with five papers on statistical applications. Section 2 presents two papers on confidentiality and data access issues. Section 3 presents two papers on changing industry codes. Finally, Section 4 includes five papers on analyses of Federal tax and information returns.

Turning Administrative Systems Into Information Systems--1999
Selected papers given at the 1999 Annual Meetings of the American Statistical Association (ASA) in Baltimore, MD. In addition, the report includes one paper presented at the 1998 ASA conference in Dallas, TX. The volume is divided into six major sections. The book begins with a complete ASA session analyzing administrative records from the U.S. tax system. It contains four papers, as well as a set of comments on the presentations. Section 2 presents four papers on the statistical uses of administrative records. Section 3 includes two papers, which focus on employee satisfaction and customer satisfaction surveys at the IRS. Section 4 contains two papers, one of which was presented at the 1998 ASA conference, that provide an update on the Survey of Consumer Finances. Section 5 presents one paper that looks at the feasibility of preparing State corporate data by matching receipts and employment data by State and industry. Finally, the volume concludes with a paper on distributional theory and computation.

Turning Administrative Systems Into Information Systems--1998-1999
Selected papers given at the 1998 Annual Meetings of the American Statistical Association in Dallas, Texas. In addition, the report includes a session of papers presented in 1999 at the Annual Meetings of the American Economic Association (AEA) plus one other paper. The volume is divided into five major sections. The book begins with the AEA session in memory of the late Dr. Daniel B. Radner, Social Security Administration economist. It contains four papers on new empirical findings in the distributions of personal income and wealth, as well as two sets of introductory remarks and two sets of comments on the presentations. Section 2 presents two papers on data measurement and data bases for economic research. Section 3 includes two papers, which focus on sample design, estimation, and imputation research. Section 4 explores issues dealing with public-use files, including the potential for disclosure. Finally, Section 5 concludes the volume with a paper verifying the classification of public charities in the 1994 Statistics of Income Study Sample. (It is the only paper not presented at the ASA or AEA meetings.)

Turning Administrative Systems Into Information Systems--1996-1997
Selected papers given primarily at the 1996 and 1997 Annual Meetings of the American Statistical Association in Chicago, Illinois and Anaheim, California, plus one non-ASA article. The volume is divided into nine major sections. The book begins with a paper originally printed as a textbook article on inheritance and wealth in America. Section 2 presents papers on using administrative records for generating national statistics. Section 3 contains two sets of panel reports on the statistical uses of administrative records. Section 4 focuses on methodological research. Section 5 explores issues dealing with quality improvement in government. Section 6 presents a panel discussion on Customer Satisfaction Surveys. Section 7 focuses on the effect of downsizing on Federal statistics. Section 8 explores the privacy area. Finally, Section 9 concludes with seven papers on statistical disclosure limitation.

Turning Administrative Systems Into Information Systems--1995
Selected papers given primarily at the 1995 Annual Meetings of the American Statistical Association in Orlando, Florida and another conference. The volume is divided into five major sections. The book begins with a paper on SOI migration data, giving an example of how this unique dataset can be used by demographers and policy researchers. Section 2 presents papers on sample designs and redesigns, as well as on SOI efforts in the corporation and partnership areas. Section 3 contains papers on weighting and estimation research. Section 4 focuses on analytical approaches to quality improvement, from graphical techniques to cognitive research. Finally, Section 5 concludes with papers from an invited session on record linkage applications for health care policy, a session organized by SOI in view of its long-term interest in improving matching techniques for administrative and survey data.

Turning Administrative Systems Into Information Systems--1994
Selected papers given primarily at the 1994 Annual Meetings of the American Statistical Association in Toronto, Ontario, Canada. The volume is divided into nine major sections. The book begins with an overview of the Statistics of Income Programs, describing the origins and customers of various SOI data and highlighting our products and services. Section 2 presents the descriptive results from two recent studies--one on sales of capital assets and one on self-employed nonfilers. Section 3 contains papers and discussion from a session on privacy issues involved in using administrative record data. The next two sections are much more methodical in nature: Section 4 focuses on sample design and estimation work in SOI, beginning with a reprint of a 1963 paper by W. Edwards Deming, which presents an evaluation of the SOI sample. Section 5 presents data on record linkage. Section 6 draws together the papers from a session on nonresponse in Federal surveys. Section 7 is a more statistical section, which contains a collection of papers on imputation methodology in a number of different arenas. Section 8 focuses on another long-time theme of these volumes--quality improvement efforts. Finally, Section 9 presents two unrelated papers on data preparation techniques.

Turning Administrative Systems Into Information Systems--1993
Selected papers given at the 1993 Annual Meetings of the American Statistical Association in San Francisco, California and other related conferences. The volume contains seven major sections, each focusing on a somewhat different area of research. The first section begins with a paper that presents a view for the future of the Federal statistical system. This effort is part of a dialogue with other agency leaders to redefine a cohesive plan for Federal data producers and users. Section 2 contains several descriptive papers based on tax data about individuals, and Section 3 looks at similar uses of tax data for businesses. Section 4 focuses on sample design issues for several SOI projects, while Section 5 presents information on improvements to analytical techniques. Finally, Sections 6 and 7 describe a number of different studies SOI is involved in to improve the quality and productivity of other areas of IRS.

Turning Administrative Systems Into Information Systems--1991-1992
Selected papers given mostly at the 1991 and 1992 Annual meetings of the American Statistical Association, held, respectively, in Atlanta, Georgia and Boston, Massachusetts. Papers chosen for this volume exemplify some of the

basic changes that are occurring in the Statistics of Income program during the 1990's, including discussions of methodological improvements and applications currently under way in the U.S. Federal statistical community. The volume contains seven general areas of interest: information from tax return data; the 1989 Survey of Consumer Finances; estimation and methodological research in the SOI business program; sample design and weighting issues in the SOI individual program; some quality improvement applications; some technological innovations for SOI research; and a look to the future data needs for the Federal sector. Previous volumes in the series were called Statistics of Income and Related Administrative Record Research (see below). The title was changed to more clearly reflect how the Internal Revenue Service's Statistics of Income function is adapting to better meet the informational needs of its many customers.

Statistics of Income and Related Administrative Record Research--1990
Selected papers given primarily at the 1990 Annual meeting of the American Statistical Association in Anaheim, California. Papers selected for this volume contain discussions of methodological improvements and applications currently under way in the U.S. Federal statistical community. In particular, the focus is on work being done by the Statistics of Income Division of the Internal Revenue Service (IRS). The volume covers five general areas: longitudinal panel data and estimation issues; analytical research using survey and administrative data; design issues for Federal surveys; information on the conclusions of the Establishment Reporting Unit Match Study; and a look at future data needs for the Federal sector.

Statistics of Income and Related Administrative Record Research--1988-1989
Selected papers given mostly at the 1988 and 1989 Annual Meetings of the American Statistical Association in New Orleans, Louisiana and Washington, D.C., respectively. Papers for the volume focus on perspectives on statistics in government--in celebration of ASA's 150th anniversary; improvements in income and wealth estimation; methodological enhancements to administrative record data; some looks at the effects of tax reform; and technological innovations for statistical use.

Statistics of Income and Related Administrative Record Research--1986-1987
Selected papers given, for the most part, at the 1986 and 1987 Annual Meetings of American Statistical Association in Chicago and San Francisco, respectively. Papers focus on ongoing wealth estimation research and U.S. and Canadian efforts regarding methodological enhancements to corporate and individual tax data and recent refinements to disclosure avoidance techniques.

Record Linkage Techniques--1985*
The Proceedings of the Workshop on Exact Matching Methodologies held in Arlington, Virginia, May 9-10, 1985. Includes landmark background papers on record linkage use and papers describing methodological enhancements, applications, and technological developments, as well as extensive bibliographic material on exact matching.

Statistical Uses of Administrative Records: Recent Research and Present Prospects*
A two-volume reference handbook on research results involving the use of administrative records for statistical purposes from 1979 through 1982:

- Volume I (March 1984) focuses on general considerations in administrative record research, applications of income tax data, uses based on data from other major administrative record systems, and enhancements to statistical systems using administrative data.

- Volume II (July 1984) focuses on comparability and quality issues, access to administrative records for statistical purposes, selected examples of end uses of linked administrative statistical systems, and a status report that sets goals for the future.

Statistics of Income and Related Administrative Record Research--1984*
Selected papers given at the 1984 Annual Meeting of American Statistical Association in Philadelphia. Papers focus on future policy issues, applications, exact matching techniques, quality control, missing data, and sample design issues.

Statistics of Income and Related Administrative Record Research--1983*
Selected papers given at the 1983 Annual Meeting of American Statistical Association in Toronto. Papers focus on use of administrative records in censuses and surveys, applications for epidemiologic research and other statistical purposes, and statistical techniques involving imputation and disclosure and confidentiality

Statistics of Income and Related Administrative Record Research--1982*
Selected papers given at the 1982 Annual Meeting of American Statistical Association in Cincinnati. Papers focus on statistical uses of administrative records, resulting methodologic advances, and estimates and projections for intercensal updates.

Statistics of Income and Related Administrative Record Research*
Selected papers given at the 1981 Annual Meeting of American Statistical Association in Detroit. Papers focus on applications and methodologies with an emphasis on IRS's Statistics of Income Program, the Small Business Data Base, nonprofit and pension data, and on Canada's Generalized Iterative Record Linkage System.

Economic and Demographic Statistics*
Selected papers given at the 1980 Annual Meeting of American Statistical Association in Houston. Papers focus on evaluation of the 1977 Economic Census, CPS hot deck techniques, and efforts to upgrade Social Security's Continuous Work History Sample.

---

*Out of print--Copies of selected papers can be obtained upon request.

NOTE: The IRS Methodology Reports on statistical uses of administrative records are now being offered free of charge. To obtain copies, write to:

Statistical Information Services (SIS)
Statistics of Income Division (RAS:S:SS:SD)
Internal Revenue Service
P.O. Box 2608
Washington, DC 20013-2608
Phone: (202) 874-0410
FAX: (202) 874-0964
E-mail: sis@irs.gov