Special Studies
in
Federal Tax Statistics

2004

▽

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

Selected Papers Given in 2004
at the Annual Meetings of the
American Statistical Association
and Two Other Professional Conferences

PREFACE

This is the fourth edition of a new set in the IRS Methodology Report series, now entitled *Special Studies in Federal Tax Statistics: 2004* (and formerly called *Turning Administrative Systems Into Information Systems*).

The papers included in this volume were presented at the 2004 Annual Meetings of the American Statistical Association (ASA) held in Toronto, Ontario, Canada, and two other professional conferences--the Luxembourg Wealth Study Workshop in Perugia, Italy, and the Conference on Privacy in Statistical Databases in Barcelona, Spain.

◆ Content

This year's compilation has been divided into five areas of interest:

❑ The volume begins with a group of four papers on recent developments in Statistics of Income research;

❑ The second section includes five papers on quality assessment of administrative records data;

❑ The third section presents a paper on estimates of income and wealth from survey and tax data;

❑ The fourth section presents a paper on disclosure protection techniques;

❑ The final section contains a paper on some current theoretical research on multivariate analysis presented in a poster session at ASA.

◆ Basic Format

Ten of the articles in this volume were prepared by each author for publication in the 2004 Proceedings of the American Statistical Association. Therefore, the format conforms basically to that required by the ASA, with the exception that we have not imposed a strict page limitation. Hence, in some cases, additional explanatory material may be included that is not available in the Proceedings.

◆ Copy Preparation

The contents of the papers included here are the responsibility of the authors. They followed ASA's peer review guidelines for Proceedings papers and then sought additional comments from colleagues either within SOI Division or elsewhere within IRS. Also, during the compilation of this material, any major problems observed were brought to the attention of the authors. Minor editorial changes were considered the prerogative of the editors. In no sense, therefore, have the articles in this volume been "refereed."

◆ Acknowledgments

The editors of this collection, James Dalton and Beth Kilss, would like to thank Lisa Smith for her invaluable contribution in laying out all the papers in this volume.

Thomas B. Petska
Director
Statistics of Income Division
Internal Revenue Service

July 2005

Special Studies in Federal Tax Statistics
2004

CONTENTS

Special Studies in Federal Tax Statistics: 2004 Online

Special Studies in Federal Tax Statistics: 2004 is available online on the IRS Internet site at:

http://www.irs.gov/taxstats/productsandpubs/article/0,,id=125133,00.html. From this page, click
on "Special Studies in Federal Tax Statistics" and select "2004."

1

Recent Developments in Statistics of Income Research

Weber

Strudler ◆ Petska ◆ Petska

Sailer ◆ Holden

Cecco

The Statistics of Income 1979-2002 Continuous Work History Sample Individual Income Tax Return Panel

Michael Weber, Internal Revenue Service

Since 1979, the annual SOI Individual Income Tax Return Cross-Sectional Sample has had at least one Continuous Work History Sample (CWHS) Social Security Number (SSN) ending embedded in the sampling framework. The CWHS utilizes a feature of the SSN numbering system where the last four digits of the number have the properties of a random number. Thus, by sampling on the last four digits, a random sample can be obtained.[1] The CWHS sample was embedded in the sample in order to create an occupational match study with the Social Security Administration. It was envisioned that the study would be repeated, and, eventually, longitudinal occupation data could be produced.[2] The match study never came to fruition, but the CWHS sample remained embedded in the annual SOI cross-sectional sample. Limited use was made of the longitudinal aspects of the CWHS portion of the SOI sample until the mid-1980's when a tax return panel was created. This panel, which began with 1979 data, was then made into a public-use file. Data for the public-use file was released through Tax Year 1990. SOI stopped the public release of data beyond 1990 due to disclosure concerns. Subsequently, SOI turned its attention to the creation of stratified panels: a 1985-2001 Sales of Capital Assets Panel, a 1987-1996-based Family Panel, and two ongoing 1999 based panels. The 1979-1990 CWHS panel was never updated until now.

▶ The Cross-Sectional Problem

Before turning to the 1979-2002 CWHS Panel, one cross-sectional issue needs to be addressed. Table 1 shows the weighted total return counts for the SOI Individual Income Tax Return Cross-Sectional Sample and the CWHS subsample for the SOI years 1979-2002.[3] An interesting feature of this table is that the CWHS cross-sectional totals are always less than the full SOI cross-sectional totals. One would expect some random differences between two samples with the CWHS cross-section sometimes estimating more and sometimes fewer tax returns than the full SOI cross-section. But one would not expect the CWHS cross-section total to always be less than the full SOI cross-section. However, since the CWHS sample involves the same SSN's each year, and since the SOI sample is based on a transformed SSN, both of these samples in a given year have a high degree of overlap with the samples in all other years. Thus, if there were some systemic error involved with the CWHS sample as compared to the full SOI sample, one would expect that same error and the sign of the error to occur from year to year. The question then becomes what is the source of this consistent shortfall in the CWHS. One source of this shortfall may reside in the IRS issuance of Individual Taxpayer Identification Numbers or ITIN's to individuals who do not have SSN's. The IRS system of issuance for ITIN's may not produce numbers where the last four digits are random numbers. While this is a possible explanation, the issue merits further research. Interestingly, the difference between the CWHS and the full SOI cross-section significantly diminished as the CWHS sample was expanded to five endings for SOI Year 1998.

▶ The Use of the Primary SSN as the Unique Panel Identifier

Panel files require the use of a unique control number to identify the temporal observations associated with the unit that is being followed. Since taxpayers are required to provide their SSN's on their tax returns, the SSN is a good candidate for this unique person control number. In IRS computer systems, the use of the primary taxpayer SSN as a unique individual identifier is generally very accurate. This is due to fact that IRS returns processing rules do not permit duplicate primary SSN's to be "posted" or moved onto the IRS Individual Master File after the initial tax return transcription process has been completed. Thus, for a given tax year, there is one tax return for each primary SSN and one primary SSN for each tax return. No system, however, is error-free, and duplicate primary SSN's do slip in.[4] But for the entire 24 years of the panel, there were, approximately, only

700 cases or less than 0.1 percent of the sample where a deletion was required due to multiple returns using the same primary SSN for the same tax year.

▶ Eliminating Tax Returns Incorrectly Linked to an SSN

If one assumes that any taxpayer or IRS transcription errors found with primary SSN's are random, then each tax return found in the SOI CWHS sample is a valid sample record for cross-sectional purposes. Some returns in the sample should not be there, but a like number of returns that should be there are not. Longitudinally, however, sampled returns bearing the same primary SSN are useful only if they actually represent the same individual. Mistakes, intentional and unintentional, do occur in the use of SSN's as unique personal identifiers on tax returns. In a longitudinal, sample, this situation must be corrected. The question then becomes how to identify these situations. The most easily identifiable situation is where multiple returns show the same primary SSN for the same tax year. Fortunately, as mentioned earlier, this problem accounts for only a very small portion of the sample.

The next step is to identify and separate the true owner of the SSN from the incorrect user(s) of that SSN. Fortunately, SOI has a few tools at its disposal for this purpose. First, taxpayers are required to list their full names on the tax return; thus, a simple comparison of taxpayer names solves many problems. Unfortunately, SOI did not retain the full name listed on the tax return until 1988, and then only for special studies. For the CWHS panel, the full names for all members of the panel exist only for returns filed for SOI Year 1999 and later. What has been retained for all years is the IRS-generated name control, which is derived from the full name listed on the return. A name control is the first four digits of an individual's last name.

Second, IRS has access to an extract of the SSA's Numident file, which contains information on all of the name controls legally used with a given SSN. This file also contains a date of birth, gender, and, if applicable, a date of death.[5] As a general rule, for this paper, a taxpayer incorrectly uses an SSN when the name control listed for that SSN by SSA does not correspond to the name shown on the tax return, while a taxpayer correctly uses an SSN when the name control listed for that SSN by SSA does correspond to that shown on the return. In most SSN multiple-use cases, the taxpayer who incorrectly uses the SSN is readily identifiable. For example, for a year where two returns were filed using the same primary SSN, one return will have a name control that does not correspond to any of the valid SSA name controls, while the name control listed on the other return does correspond with a valid SSA name control.

Once the duplicate return situation has been resolved for the particular tax year in question, the rest of the returns for the remaining tax years in that SSN sequence need to be checked because a taxpayer may have filed using an incorrect SSN for years without causing a multiple return problem. This is most likely to occur due to one of three situations. The first situation occurs when the age associated with the SSN is under 21. If a taxpayer incorrectly uses a given SSN for a number of years, and then the true owner of the SSN enters the workforce after high school or college and begins to file returns as a primary taxpayer, multiple returns appear. The returns in the sequence filed prior to the first filing by the true owner must be removed. This can also happen in reverse when a taxpayer retires and perhaps is no longer required to file a tax return. A third situation occurs when a single woman files as an unmarried person and thus reports her SSN in the primary position and then marries and files returns as the secondary taxpayer. While she is married, another taxpayer incorrectly uses her SSN. If the woman subsequently divorces and again files as unmarried, a multiple return situation occurs. Approximately 75 returns were removed from the sample because, after finding at least one duplicate situation in a particular year, other returns in other years were found to have been filed by the same "incorrect" taxpayer but without causing a duplicate SSN problem.

In all of these cases, multiple returns using the same SSN within a tax year trigger the review process. A more difficult problem arises when multiples are not present but two different taxpayers are represented within the same longitudinal sequence of tax returns. This situation can be found by examining a sequence of returns using the same primary SSN but where the IRS name controls differ between years. First, let us examine the

case of males. Generally, males have only one SSA name control since men seldom change their last names. Consequently, all CWHS SSN's listed as Males were checked if the IRS name controls changed between any combinations of years. Once again, using the SSA name controls and the full name found on the tax return, this problem can be readily resolved. Approximately 225 returns were removed as a result of this test.

Returns where a woman is the owner of the primary SSN are more complicated because additional name controls are added to the SSA name control list when a woman changes her name due to marriage. Therefore, these returns were reviewed for name control problems only when an IRS name control did not match any of the valid SSA name controls. Approximately 500 returns were removed due to this check.

In the end, as shown in table 2, 1,517 records were removed from the sample, or 0.23 percent of all sampled returns.

▶ An Implication of Removing "Bad" Returns

As noted above, some returns selected for the SOI CWHS sample were selected because the SSN's listed on the returns were incorrect. In other words, the SSN's were SOI CWHS SSN's but they did not belong to the taxpayers who used them on the tax returns. Over time, as taxpayers resolve these SSN problems and begin to use their correct SSN's, they disappear from the CWHS sample. If SOI was able to perform real-time SSN resolution, SOI could continue sampling those taxpayers using their correct SSN's. Since this is not currently possible, these individuals were removed from the sample since, at a minimum, their longitudinal observations are incomplete. Conversely, individuals whose true SSN's are SOI CWHS SSN's but who filed returns using incorrect SSN's are not included in the SOI CWHS sample, and no realistic attempt could have been made to find them and follow them as they continued to use incorrect SSN's. The net result of these two situations is that the weighted totals generated by the CWHS panel sample, when weighted using the inverse of the sampling rate, are shy of the true population totals. It is possible that a post sampling weighting adjustment could be made for

each possible base year of the panel, but such an attempt will require more research.

▶ The Gender Bias Problem

A very unfortunate implication of a panel based on sampling primary SSN's is that it produces a profound gender bias. Table 3 shows the gender of the primary taxpayers in the SOI CWHS and of the spouses listed as secondary taxpayers on those CWHS returns that show a joint filing status. Table 4 shows the gender of just the primary taxpayers. The source of the difference between table 3 and 4 is shown in table 5; Over 95 percent of joint returns are filed with the male listed as the primary taxpayer. This does not create a cross-sectional problem, as the total number of taxpayers (primary and secondary) by gender will still be correctly represented as shown in Table 3.

Longitudinally, however, this is an enormous problem because taxpayers are followed solely on the basis of the primary SSN. If taxpayers never change their marital status from an initial base year state the gender bias problem would not exist. However, people do get married and divorced. Thus, from a panel perspective, if one wishes to study individuals who never get married or who are married to the same person for the period under study, the gender bias created by sampling on primary SSN's is not a problem. For all other situations, the problem is inescapable.

▶ From Filer to Nonfiler to Death

When analyzing a longitudinal sample, a user must always be aware of, and have a strategy for, dealing with missing observations and panel attrition. The most important piece of information a user needs in order to develop such a strategy is an explanation of what happened to the missing observations.[6] Suppose a taxpayer files returns for 3 years then vanishes never to file again; what happened to this individual? Did the individual die, retire, or marry? The answer to these questions affects the meaning of any analysis developed using a panel.

One possible explanation is that the taxpayer was a woman who married and subsequently filed as the secondary taxpayer on a joint return. As a result, she

disappears from a panel of primary taxpayers. This is the gender bias problem discussed above. Fortunately, for 2 base years, we are able to solve this problem. In 1987 and in 1999, SOI began panels where the base year primary SSN's were followed in future years whenever they appeared in either the primary or secondary positions. But a limitation of these two panels is that, unlike the primary SSN-based CWHS panel, in which any year from 1979 to 2002 can be used as a base year, the beginning, or base year, is limited to 1987 and 1999. In addition, the 1987 panel ended in 1996.

There are other legitimate reasons why a taxpayer may disappear from the CWHS primary SSN panel, or any other tax return panel for that matter. Two primary reasons are: an income insufficient to require the filing of a tax return; and, death. Fortunately, we have some tools to help with these situations. Someone once said there were only two things certain in life--death and taxes--but our income tax system provides a third possibility. It is possible to be alive and be the recipient of income and not be required to file a tax return or pay income tax. This situation occurs most often with individuals living on Social Security whose incomes are below the filing thresholds for the income tax system. But for purposes of tax return panels, these individuals disappear. Fortunately, IRS creates something called the Information Returns Master File, which contains information documents (Form W-2, Form 1099, Form 1098, etc.) that show whether an individual received any income from a variety of sources during a given year. So, for individuals whose only source of income is Social Security Benefits, and who thus do not file tax returns, SOI has evidence that they are alive and receiving income. Unfortunately, such data are only available for the years 1989, 1993, and 1996 to the present. The use of the IRMF has been the subject of previous ASA papers.[7] Finally, the same SSA files that provide information on name control and gender also provide us with dates of death.

▶ The 1979-2002 SOI CWHS Primary SSN Panel -- The Conclusion

To summarize:

- SOI has created a panel of primary taxpayers that begins in 1979 and continues to the present.

- Duplicate returns and erroneous returns have been removed to the extent possible.

- Age, gender, and date of death information are available for these individuals.

- Base year 1987 primary taxpayers are followed even if they file as secondary taxpayers through 1996.

- Base year 1999 primary taxpayers are followed in future years even if they file as secondary taxpayers.

- Information Returns data are available for all individuals in this panel for the years 1989, 1993, and 1996 through the current year.

▶ Footnotes

[1] Smith, Creston M., "The Social Security Administration's Continuous Work History Sample," *Social Security Bulletin*, Social Security Administration, Office of Research and Statistics, October 1989, Volume 52, Number 10.

[2] Sailer, Peter; Orcutt, Harriet; and Clark, Phil (1980), "Coming Soon: Taxpayer Data Classified by Occupation," *1980 Proceedings of the American Statistical Association, Government Statistics Section*, 1981.

[3] The SOI year is one less than the calendar year or processing year. For example, taxpayers generally filed their Tax Year 2003 returns during Calendar Year 2004. Thus, the returns filed in Calendar Year 2004 would be included in the 2003 SOI file. Over 97 percent of the returns sampled for the 2003 SOI file will be for Tax Year 2003.

[4] It is possible that the source of many of these primary SSN duplicates is the SOI sampling process itself. SOI samples tax returns on a weekly basis throughout a given processing year. It does not receive later IRS corrections to those weekly sample extracts. Thus, if, in January, a taxpayer uses a specific primary SSN, and, at a later date, another taxpayer lists the same primary SSN, IRS

will resolve this situation. For example, if the second occurrence of the SSN was determined to be incorrect, the return would not be posted to the IRS master file, and that return would never be subject to SOI sampling. But if the first occurrence of the SSN was determined to be wrong, SOI would still have the tax return listing the first occurrence in its sample, as well as the second tax return. This would produce a duplicate use of a primary SSN in SOI files.

[5]   IRS does not receive all of the death information contained on the NUMIDENT file. The death information SSA obtains from approximately half the States, and for which SSA cannot independently verify the date of death, cannot be shared with IRS due to restrictions placed on that information by these States. Fortunately, SSA is able to independently verify a significant number of the deaths in these States due to the administrative process of stopping Social Security Benefit payments for the deceased individuals. At this time, SSA is not able to provide an estimate of the number of missing entries for date of death, but a reasonable guess would place it below 5 percent.

[6]   For some data on CWHS panel attrition and ideas on how to use a panel of tax returns, see Sailer, Peter; Weber, Michael; and Wong, William, "Attrition in a Panel of Individual Income Tax Returns, 1992-1997," 2000 *Proceedings of the American Statistical Association, Government Statistics Section*, 2001.

[7]   Sailer, Peter; Weber, Michael; and Yau, Ellen, "How Well Can IRS Count the Population," 1993 *Proceedings of the American Statistical Association, Government Statistics Section*, 1994.

Sailer, Peter and Weber, Michael, "The IRS Population Count: An Update," 1998 *Proceedings of the American Statistical Association, Government Statistics Section*, 1999.

Table 1

			All Records			
SOIYR *	CWHS Endings in SOI	Unweighted Count	Weighted Total	SOI Complete Report (CR)	SOI CR less CWHS Total	SOI CR less CWHS Total %
1979	3	27,248	90,826,576	92,694,302	1,867,726	2.01%
1980	3	27,684	92,279,908	93,902,469	1,622,561	1.73%
1981	3	27,799	92,663,241	95,396,123	2,732,882	2.86%
1982	1	9,353	93,530,000	95,337,432	1,807,432	1.90%
1983	2	19,155	95,775,000	96,321,310	546,310	0.57%
1984	1	9,752	97,520,000	99,438,708	1,918,708	1.93%
1985	2	20,207	101,035,000	101,660,287	625,287	0.62%
1986	1	10,138	101,380,000	103,045,170	1,665,170	1.62%
1987	2	21,238	106,190,000	106,996,270	806,270	0.75%
1988	2	21,718	108,590,000	109,708,280	1,118,280	1.02%
1989	2	22,379	111,895,000	112,136,673	241,673	0.22%
1990	2	22,694	113,470,000	113,717,138	247,138	0.22%
1991	2	22,759	113,795,000	114,730,123	935,123	0.82%
1992	2	22,609	113,045,000	113,604,503	559,503	0.49%
1993	2	22,730	113,650,000	114,601,819	951,819	0.83%
1994	2	22,965	114,825,000	115,943,131	1,118,131	0.96%
1995	2	23,469	117,345,000	118,218,327	873,327	0.74%
1996	2	23,878	119,390,000	120,351,208	961,208	0.80%
1997	2	24,172	120,860,000	122,421,991	1,561,991	1.28%
1998	5	62,318	124,636,000	124,770,662	134,662	0.11%
1999	5	63,435	126,870,000	127,075,145	205,145	0.16%
2000	5	64,677	129,354,000	129,373,500	19,500	0.02%
2001	5	64,910	129,820,000	130,255,237	435,237	0.33%
2002	5	64,858	129,716,000	130,076,443	360,443	0.28%

* SOIYR is defined as the Calendar Year of IRS Processing minus one. Thus, the returns filed and sampled in 1980, of which most are for Tax Year 1979, are found in the SOIYR 1979 Individual Income Tax Return File.

Table 2

	All Records less Deleted Records					Deleted Records	
SOIYR *	Unweighted Count	Weighted Total	SOI Complete Report (CR)	SOI CR less CWHS Total	SOI CR less CWHS Total %	Records Deleted	Weighted
1979	27,162	90,539,909	92,694,302	2,154,393	2.32%	86	430,000
1980	27,566	91,886,575	93,902,469	2,015,894	2.15%	118	590,000
1981	27,720	92,399,908	95,396,123	2,996,215	3.14%	79	395,000
1982	9,303	93,030,000	95,337,432	2,307,432	2.42%	50	250,000
1983	19,078	95,390,000	96,321,310	931,310	0.97%	77	385,000
1984	9,694	96,940,000	99,438,708	2,498,708	2.51%	58	580,000
1985	20,118	100,590,000	101,660,287	1,070,287	1.05%	89	445,000
1986	10,084	100,840,000	103,045,170	2,205,170	2.14%	54	540,000
1987	21,119	105,595,000	106,996,270	1,401,270	1.31%	119	595,000
1988	21,634	108,170,000	109,708,280	1,538,280	1.40%	84	420,000
1989	22,314	111,570,000	112,136,673	566,673	0.51%	65	325,000
1990	22,641	113,205,000	113,717,138	512,138	0.45%	53	265,000
1991	22,688	113,440,000	114,730,123	1,290,123	1.12%	71	355,000
1992	22,537	112,685,000	113,604,503	919,503	0.81%	72	360,000
1993	22,658	113,290,000	114,601,819	1,311,819	1.14%	72	360,000
1994	22,906	114,530,000	115,943,131	1,413,131	1.22%	59	295,000
1995	23,411	117,055,000	118,218,327	1,163,327	0.98%	58	290,000
1996	23,835	119,175,000	120,351,208	1,176,208	0.98%	43	215,000
1997	24,146	120,730,000	122,421,991	1,691,991	1.38%	26	130,000
1998	62,269	124,538,000	124,770,662	232,662	0.19%	49	98,000
1999	63,389	126,778,000	127,075,145	297,145	0.23%	46	92,000
2000	64,645	129,290,000	129,373,500	83,500	0.06%	32	64,000
2001	64,879	129,758,000	130,255,237	497,237	0.38%	31	62,000
2002	64,835	129,670,000	130,076,443	406,443	0.31%	23	46,000

* SOIYR is defined as the Calendar Year of IRS Processing minus one. Thus, the returns filed and sampled in 1980, of which most are for Tax Year 1979, are found in the SOIYR 1979 Individual Income Tax Return File.

Table 3
SOI CWHS - Unweighted Taxpayer Counts by Gender

SOI Year	All Taxpayers	Male	Female	Percent Male
1979	40,434	20,137	20,131	49.8%
1980	40,852	20,276	20,427	49.6%
1981	41,071	20,316	20,602	49.5%
1982	13,839	6,773	7,023	48.9%
1983	28,259	13,842	14,316	49.0%
1984	14,385	7,046	7,305	49.0%
1985	29,591	14,516	14,992	49.1%
1986	14,800	7,235	7,530	48.9%
1987	30,592	15,042	15,496	49.2%
1988	31,184	15,336	15,792	49.2%
1989	31,944	15,766	16,138	49.4%
1990	32,284	15,916	16,304	49.3%
1991	32,342	15,939	16,340	49.3%
1992	32,092	15,786	16,238	49.2%
1993	32,187	15,797	16,305	49.1%
1994	32,474	15,980	16,424	49.2%
1995	33,108	16,205	16,826	48.9%
1996	33,490	16,448	16,997	49.1%
1997	33,840	16,596	17,220	49.0%
1998	87,035	42,509	44,485	48.8%
1999	88,233	42,998	45,208	48.7%
2000	89,707	43,777	45,902	48.8%
2001	90,216	44,034	46,158	48.8%
2002	90,399	43,917	46,461	48.6%

Table 4
SOI CWHS - Primary Taxpayer Unweighted Counts by Gender

SOI Year	All Returns	Male	Female	Unclassified	Percent Male
1979	27,162	19,899	7,097	166	73.3%
1980	27,566	20,058	7,359	149	72.8%
1981	27,720	20,080	7,487	153	72.4%
1982	9,303	6,686	2,574	43	71.9%
1983	19,078	13,660	5,317	101	71.6%
1984	9,694	6,957	2,703	34	71.8%
1985	20,118	14,331	5,704	83	71.2%
1986	10,084	7,149	2,900	35	70.9%
1987	21,119	14,852	6,213	54	70.3%
1988	21,634	15,154	6,424	56	70.0%
1989	22,314	15,567	6,707	40	69.8%
1990	22,641	15,700	6,877	64	69.3%
1991	22,688	15,723	6,902	63	69.3%
1992	22,537	15,561	6,908	68	69.0%
1993	22,658	15,541	7,032	85	68.6%
1994	22,906	15,722	7,114	70	68.6%
1995	23,411	15,898	7,436	77	67.9%
1996	23,835	16,145	7,645	45	67.7%
1997	24,146	16,298	7,824	24	67.5%
1998	62,269	41,719	20,509	41	67.0%
1999	63,389	42,190	21,172	27	66.6%
2000	64,645	42,900	21,717	28	66.4%
2001	64,879	43,076	21,779	24	66.4%
2002	64,835	42,860	21,954	21	66.1%

Table 5
SOI CWHS Joint Returns - Unweighted Counts by Gender

SOI Year	All Returns	Male	Female	Unclassified	Percent Male
1979	13,272	13,034	188	50	98.2%
1980	13,286	13,068	170	48	98.4%
1981	13,351	13,115	190	46	98.2%
1982	4,536	4,449	77	10	98.1%
1983	9,181	8,999	156	26	98.0%
1984	4,691	4,602	82	7	98.1%
1985	9,473	9,288	164	21	98.0%
1986	4,716	4,630	77	9	98.2%
1987	9,473	9,283	177	13	98.0%
1988	9,550	9,368	173	9	98.1%
1989	9,630	9,431	193	6	97.9%
1990	9,643	9,427	202	14	97.8%
1991	9,654	9,438	204	12	97.8%
1992	9,555	9,330	211	14	97.6%
1993	9,529	9,273	235	21	97.3%
1994	9,568	9,310	248	10	97.3%
1995	9,697	9,390	290	17	96.8%
1996	9,655	9,352	295	8	96.9%
1997	9,694	9,396	295	3	96.9%
1998	24,766	23,976	783	7	96.8%
1999	24,844	24,036	807	1	96.7%
2000	25,062	24,185	875	2	96.5%
2001	25,337	24,379	954	4	96.2%
2002	25,564	24,507	1,054	3	95.9%

Further Analysis of the Distribution of Income and Taxes, 1979-2002

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 sixth 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.[1,2,3,4,5] 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 and using panel data. 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.[1,2,3,4,5] Law changes effective for 1997 substantially lowered the maximum tax rate on capital gains. The newest law changes, beginning for 2001, lowered marginal rates and the maximum tax rate on long-term capital gains, as well as decreased 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 24 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.[6,7,8,9] 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. [1,2,3,4,5] 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,

including nontaxable portions that were excluded from AGI. Social Security benefits (SSB) 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 2002. 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.[10,11,12] 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 database, 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, $1,278,479 were needed for entry into this class for 2002. This represents more than a 400-percent increase. Also, while $79,679 of retrospective income were needed to enter the top 1-percent size class for 1979, $315,937 were needed for entry into this size class for 2002, an increase of 297 percent. For the top 20 percent, the threshold increased by 162 percent, and, for the bottom 20 percent, the increase was only 130 percent. Since much of these increases is attributable to inflation, we computed constant dollar thresholds, using the Consumer Price Index.[13]

What is most striking about these data are the changes between 1979 and 2002 for the various income-size percentile thresholds (see Figure A). For example, the threshold for the top 0.1 percent grew (using a 1982-1984

Figure A-Constant Dollar Income Thresholds, 1979-2002 (1982-84=100)
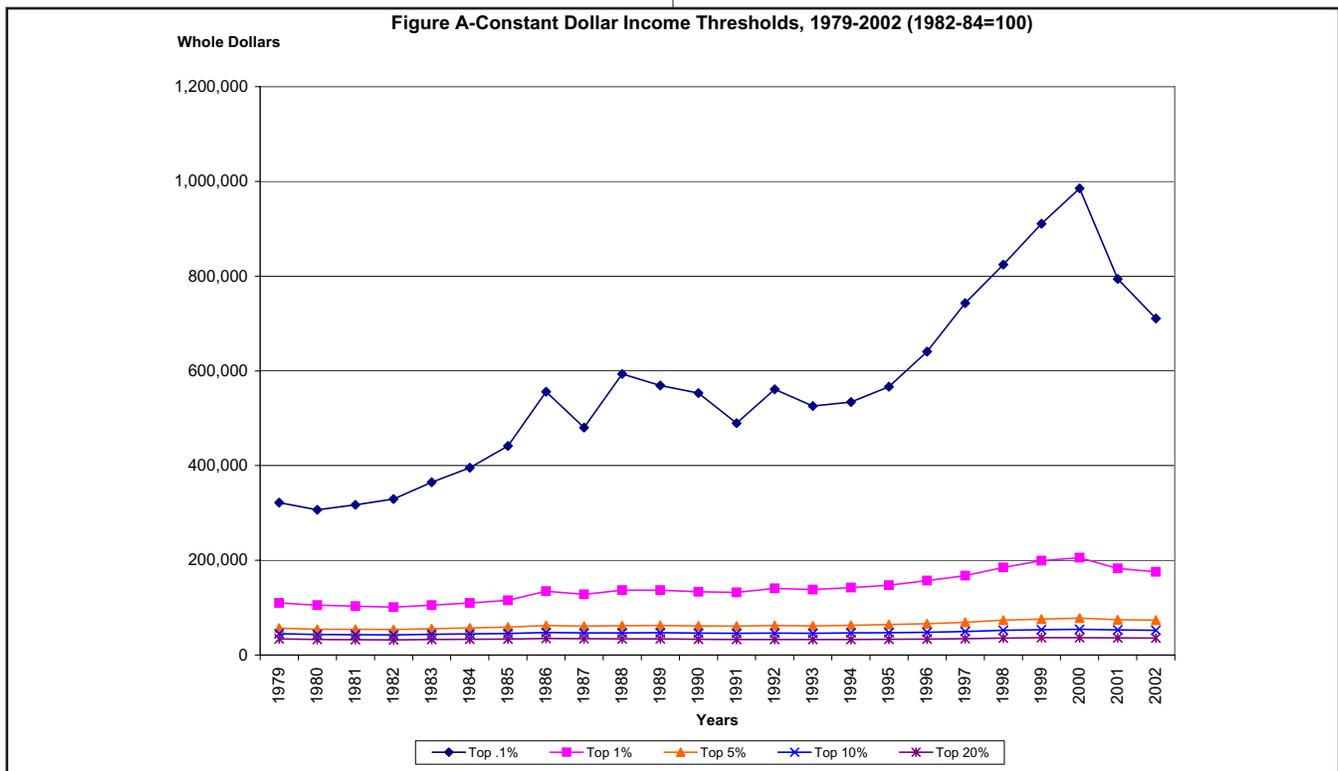
base) from $321,679 for 1979 to $710,661 for 2002, an increase of 121 percent. Similarly, the threshold for taxpayers in the 1-percent group rose from $109,751 for 1979 to $175,618 for 2002, an increase of just over 60 percent. However, the thresholds for each lower percentile class show smaller increases in the period; the top 20-percentile threshold increased only 5.6 percent, and the 40-percent and all lower thresholds 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 a high of 21.55 (10.49 for the top 0.1 percent) for 2000. With the recession and, then, the stagnating economy of 2001 and 2002, this share had declined to 16.89 percent (7.10 for the top 0.1 percent) for 2002. While this increase has been mostly steady, there were some significantly large jumps, particularly for 1986, due to a surge in capital gain realizations after the passage, but prior to implementation, of the Tax Reform Act of 1986 (TRA). The top 1-percent share also increased rap-

idly 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 15.14 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 11.28 percent for 2002. 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 24-year period (see Figure B).

Tax Shares--Income Tax

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

Figure B-Income Shares by Income Percentile Size-Classes, 1979-2002

Figure C-Income Tax Shares by Income Percentile Size-Classes, 1979-2002
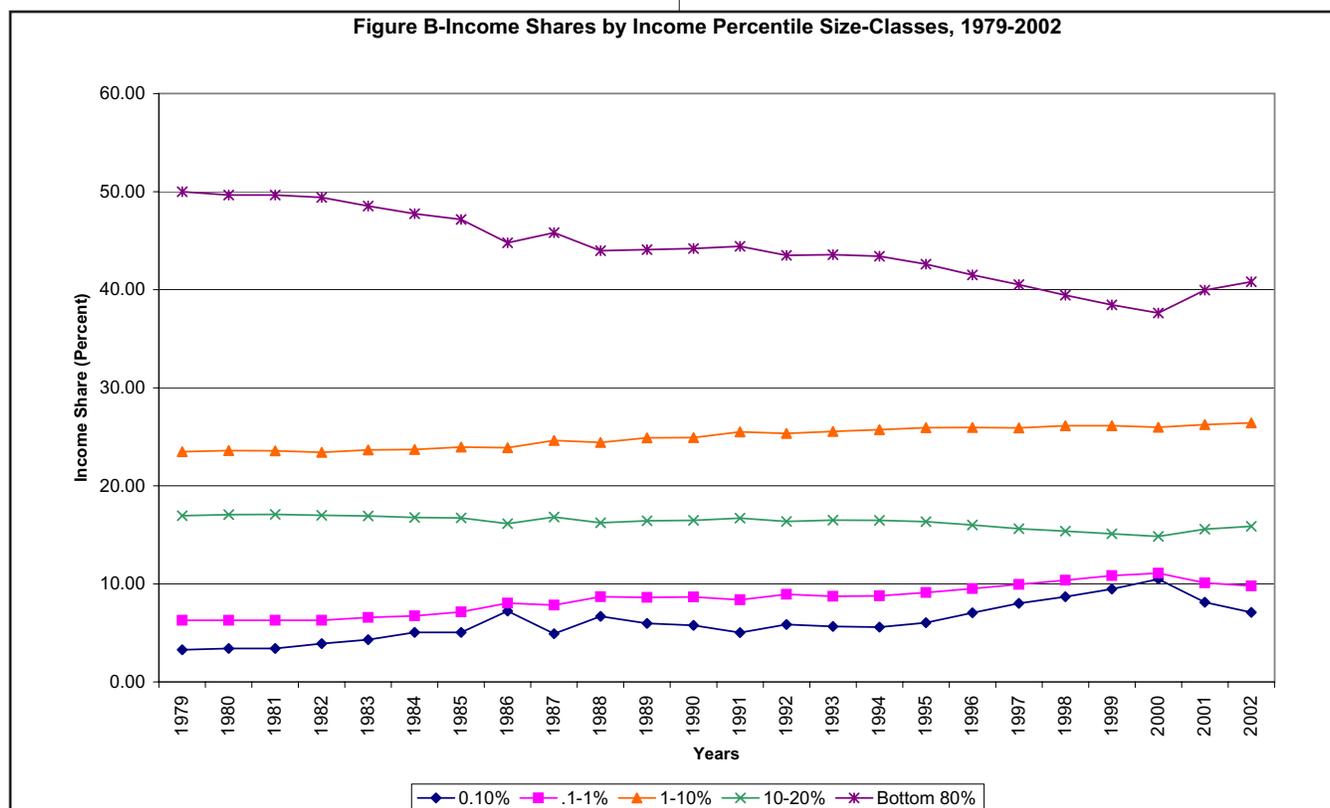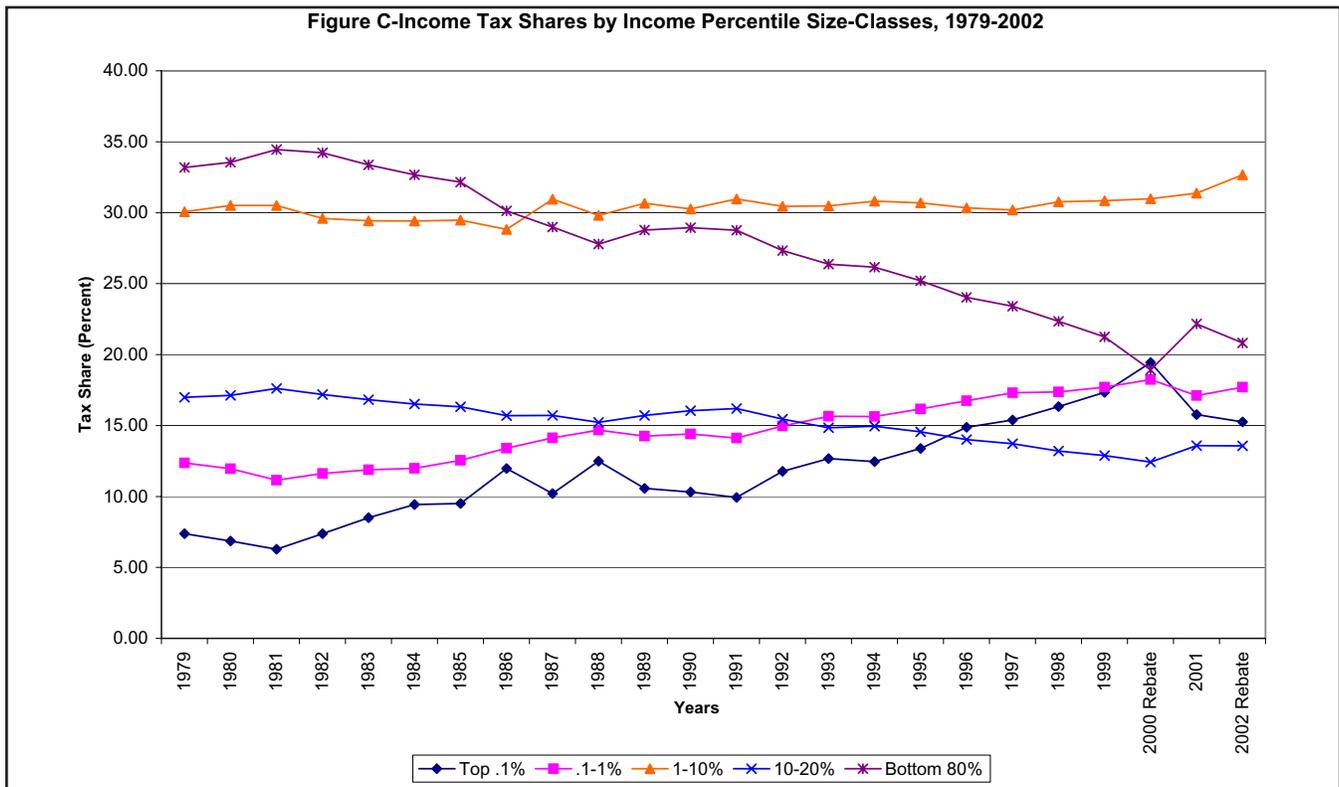
for the top 0.1 percent) for 2000 (Figure C). The corresponding percentages for 2000 for the 1-percent and 0.1-percent groups are 37.68 percent and 19.44 percent, respectively, accounting for the 2000 tax rebate, which is discussed below. For the recession year of 2001 and the subsequent year (2002) with its large decline in net gains from the sale of capital assets, these shares declined to 32.53 percent for the top 1 percent and 15.06 percent (15.25 percent including the rebate of the child tax credit) for the top 0.1-percent group (32.95 percent and 15.25 percent, respectively, including a rebate of a portion of the child tax credit). 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.[8,9]

The 1-to-5 percent size class exhibited relatively modest change in its share of taxes, increasing from 17.53 percent to 20.29 percent (20.52 percent including the rebate for the child tax credit) 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(see Figure D).[14] 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 B). For 2002, including the child tax credit rebate, the average tax rate fell to 12.56 percent, close to the lowest rate over the 24 years of this study of 12.53 percent for 1991.

In examining the average tax data by income size, four distinct periods emerge. First, the average tax rates

Figure D-Average Income Tax Rates by Size-Classes, 1979-2002
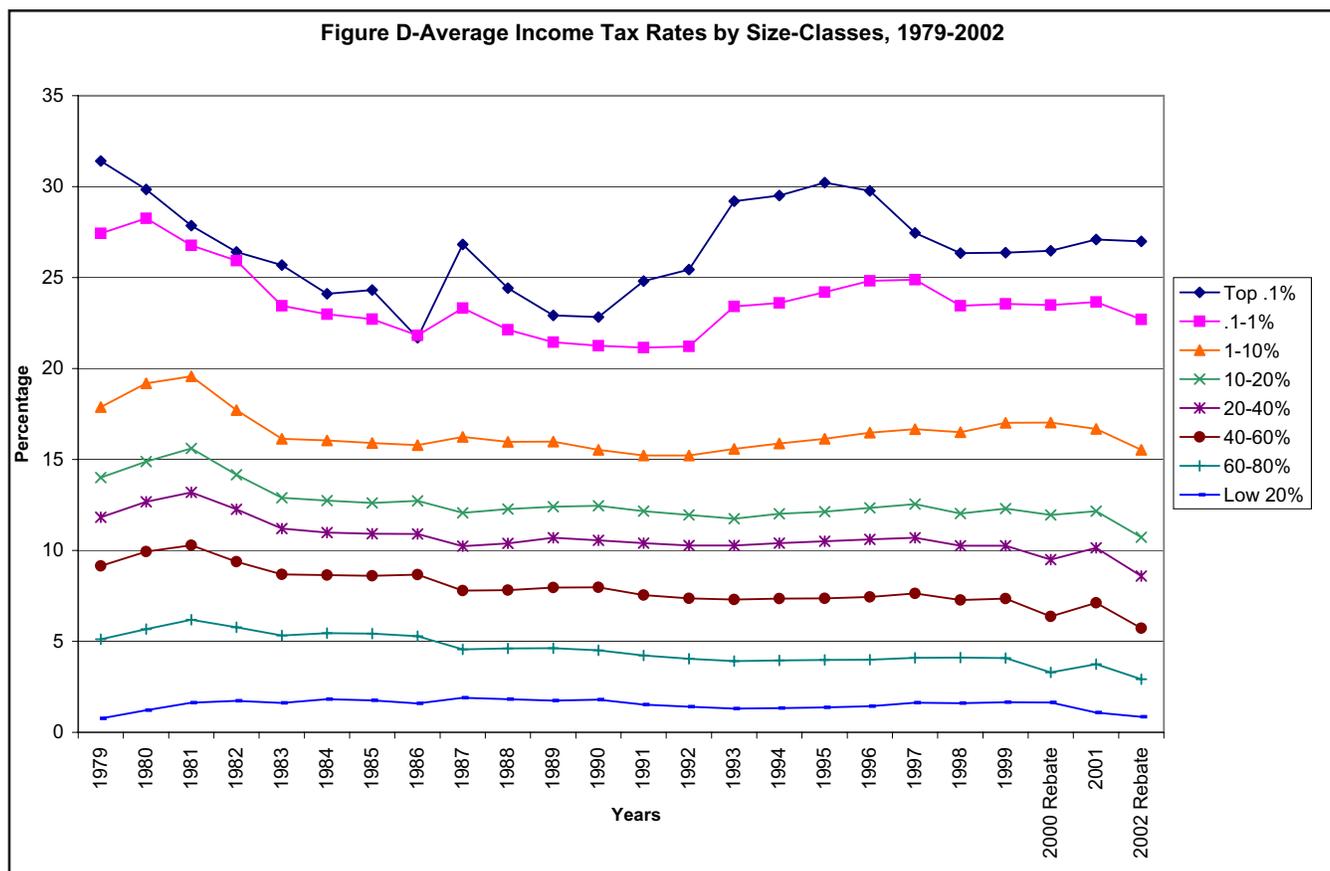
were generally climbing up to the implementation of the Economic Recovery Tax Act (ERTA) effective for 1982. This was an inflationary period, and prior to indexing of personal exemptions, the standard deduction, and tax brackets, which caused many taxpayers to face higher tax rates. (Indexing became a permanent part of the tax law for Tax Year 1985.[7]) 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.

Figure E-Tax Shares (Including Social Security Taxes) by Percentile Size-Classes, 1979-1999										
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

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.

Next, we consider if those years affected by the Taxpayer Relief Act of 1997 (1997 through 2000), 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 above 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-to-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 the 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 (estimated at $37.9 billion) is taken into account, the average rates for 2000 decreased for all groups, except for the top 0.1 percent and the 1-to-5 percent, reversing the prerebate 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 percent and the 5-to-10 percent paid lower average taxes along with the bottom 20-percent group.

For 2002, when the 10-percent rate applied to all returns and all rates above 15 percent were reduced by one-half of 1 percentage point, the average tax rate fell for every group. Further, as the economy stagnated, another rebate of $400 per child was sent to individuals who received a child tax credit for that year. This was in lieu of receiving the additional amount for 2003 as part of the increased child tax credit provided by the Jobs and Growth Tax Relief Reconciliation Act of 2003 (JG-TRRA). Simulating this on Tax Year 2002, we estimated that $14.2 billion were sent to taxpayers further reducing average taxes for 2002. The individuals who gained the most from this rebate were in the 5-to-10 percent group through the 40-to-60 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.[15] 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 JGTRRA, on the 1999 data.[16]

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 percent 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 percent, 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 it (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 it, 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, about 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 that 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, respectively.

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 that pays 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

Figure F-Average Tax Rates (Including Social Security Taxes) by Percentile Classes, 1979-1999										
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

percent income group, 2.91 percentage points less than the 5-to-10 percent income group, and 1.24 percentage points less than individuals in the 10-to-20 percent group. In fact, under the provisions of EGTRRA/JGTRRA, individuals in the 0.1-percent group wind up paying less than 1-percentage point (0.99) more than the 20-to-40 percent income group. In contrast, the highest income group paid average combined taxes that were 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.

Using Panel Data

For 1979, 1989, and 1999, we used a panel of individual tax returns that were selected at a 1-in-5,000 return random sample embedded in each year's Individual Statistics of Income (SOI) sample. These returns were based on primary taxpayers having certain Social Security number endings and being part of Social Security's Continuous Work History Sample (CWHS). The reason for studying a panel of returns is to obtain a more well-rounded approach to analyzing tax returns over time. While "the rich" may appear to be getting greater concentrations of income over time, the composition of who "the rich" are may also be changing over time. By looking at the panel, we defined income groups from the combined data (indexed for inflation) over the 1979, 1989, and 1999 period. In order to have a better income concept over time, we altered retrospective income by including total Social Security benefits. Since this was not on a tax return for 1979, in that sense, income would be understated for that year (SSB for 1979 was estimated at $29 billion). Then, we analyzed how income and taxes changed in each of these years, classifying each year's returns in quintile classes.

In analyzing this panel over time, we classified returns into quintile classes for each of the 3 years, 1979, 1989, and 1999. We started with 90.6 million returns filed for 1979 and followed these returns. We looked at movement of returns between quintile classes over time (see Figure G). In order to not include small changes in income causing returns to change classes, we only showed movement of more than one quintile. As can be seen, movement increased greatly the lower the quintile for the first year (1979). While 70.7 percent of the highest income individuals remained high-income in 1999, just 12.4 percent of the lowest quintile remained low-income in 1999. Also, the percent of returns dropping out of the panel decreased consistently with the size of 1979 income.

In further analyzing this panel over time, we only included returns that were filed for each of the 3 years, 1979, 1989, and 1999. This left us with 58.8 million returns out of the 90.6 million returns filed for 1979. Using inflation-indexed income, we then combined the income and taxes over time to create a "combined income and tax" for each of the tax returns. We then reclassified each return into percentile classes, with the 5-percent income class being the highest class analyzed (due to the high sampling variability at levels above this). Looking at average taxes for the combined income groups, the 1979 and 1999 data look progressive, similar to our analysis above in looking at cross-sectional income and Social Security taxes (Figure H). For 1989, the combined 5-percentile class paid lower average taxes than the 5-to-10 percent combined income group. Again, this regressivity is similar to what we found previously using the annual cross-section data. Comparing tax shares for the combined panel in comparison to the cross-section, we found that the trends are the same for the top

Figure G-Movement From 1979 To 1999 of More Than One Class

1979 Percent Class	1999 Percentile Class					Dropouts
	TOP 20	20 to 40	40 to 60	60 to 80	BOTTOM 20	
TOP 20	70.7%		5.9%	2.1%	0.5%	20.8%
20 to 40		64.8%		4.8%	1.4%	29.0%
40 to 60	14.7%		41.8%		2.4%	41.1%
60 to 80	8.4%	13.0%		28.8%		49.8%
BOTTOM 20	7.6%	13.7%	14.2%		12.4%	52.10%

10 percent and the top 10-to-20 percent classes, but the high-income panel returns paid a lower share for each year (Figure H1). The trend was also the same for the bottom 80 percent of returns, but, in this case, the panel returns paid a consistently higher share of taxes.

▶ 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 24 years, both before- and after-taxes.

Once the Lorenz curves were estimated for all years, Gini coefficients were calculated for all 24 years for before- and after-tax and are presented in Figure I. The Gini coefficient, which is a measure of the degree of inequality, generally increased throughout the 24-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 economic downturn in 2001 and 2002 actually decreased the levels of inequality to 0.555 (pre-tax) and 0.525 (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.[17] 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 Figure I, 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

Figure H-Combined Panel 'P' : Average Tax Rates(Including Social Security Taxes) by Size-Classes, 1979-1999							
Year	Top 5% 'P'	5-10% 'P'	10-20% 'P'	20-40% 'P'	40-60% 'P'	60-80% 'P'	Low 80% 'P'
1979	26.98	23.27	21.24	20.42	19.37	18.08	14.4
1989	23.52	23.87	22.98	22.18	20.69	18.88	15.25
1999	25.67	24.46	23.04	21.14	19.32	17.95	12.67

Figure H1-Comparison of Combined Panel ('P') vs. Cross-Sectional Data: Tax Shares (Including Social Security Taxes) by Size-Classes, 1979-1999						
Year	Top 10%	Top 10% 'P'	10-20%	10-20% 'P'	Bottom 80%	Bottom 80% 'P'
1979	40.59	29.39	17.70	15.07	41.72	55.54
1989	43.66	35.78	17.63	15.45	38.71	48.77
1999	52.19	47.38	15.98	14.04	31.83	38.58

Figure I-Gini Coefficients for Retrospective Income, Before and After Taxes, 1979 – 2002

Year	Gini Before Tax	Gini After Tax	Difference	Percent Difference
1979	0.469	0.439	0.030	6.325
1980	0.471	0.441	0.031	6.477
1981	0.471	0.442	0.029	6.233
1982	0.474	0.447	0.027	5.731
1983	0.482	0.458	0.025	5.132
1984	0.490	0.466	0.024	4.933
1985	0.496	0.471	0.024	4.860
1986	0.520	0.496	0.024	4.573
1987	0.511	0.485	0.026	5.101
1988	0.530	0.505	0.026	4.817
1989	0.528	0.504	0.024	4.592
1990	0.527	0.503	0.024	4.498
1991	0.523	0.499	0.024	4.582
1992	0.532	0.507	0.025	4.709
1993	0.531	0.503	0.028	5.207
1994	0.532	0.503	0.028	5.292
1995	0.540	0.510	0.029	5.404
1996	0.551	0.521	0.030	5.496
1997	0.560	0.530	0.030	5.368
1998	0.570	0.541	0.029	5.136
1999	0.580	0.550	0.030	5.185
2000	0.588	0.558	0.031	5.222
2000 Rebate	0.588	0.557	0.032	5.417
2001	0.564	0.534	0.030	5.352
2002	0.555	0.525	0.030	5.339
2002 Rebate	0.555	0.525	0.030	5.334

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 38.6 percent to 35 percent for 2003. The effects of the new tax laws (EGTRRA/JGTRRA) can be seen in Figure J. This 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 I. 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 J 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 and 1999. 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. Further, when Gini coefficients were calculated for these 3 years using the combined panel data, the trends over time were almost exactly the same. However, these coefficients were consistently lower for the panel, showing that there is less inequality than what is suggested by looking at cross-sectional data only.

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.

Figure J-Gini Coefficients for Retrospective Income (Including Social Security Taxes), Before and After Taxes, 1979 - 1999

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

Figure J1-Gini Coefficients for Retrospective Income (Including Social Security Taxes), Before and After Taxes Using All CWHS, Combined 1979 - 1999

Year	Gini Before Tax	Gini After Tax	Difference	Percent Difference
1979	0.311	0.291	0.019	6.272
1989	0.416	0.403	0.014	3.247
1999	0.498	0.477	0.021	4.217
Combined	0.447	0.428	0.019	4.199

Analysis of panel data shows that these trends are not quite as great as seen by looking at annual cross-section data, but the trends cited above are still apparent.

▶ Notes

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

[2] Petska, Tom; Strudler, Mike; and Petska, Ryan, 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*, 2000.

[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 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, 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,* 1999.

[5] Strudler, Michael and Petska, Tom, An Analysis of the Distribution of Individual Income and Taxes, 1979-2001, *2001 Proceedings of the American Statistical Association, Social Statistics Section*, 2002.

[6] Nelson, Susan, 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, 1987.

[7] Hostetter, Susan, 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*, 1988.

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

[9] 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.

[10] For the years 1979 through 1992, the percentile threshold size classes were estimated by osculatory interpolation as described in Oh and Oh and Scheuren.[11,12] 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.

[11] Oh, H. Lock, Osculatory Interpolation with a Monotonicity Constraint, *1977 Proceedings of the American Statistical Association, Statistical Computing Section,* 1978.

[12] Oh, H. Lock and Scheuren, Fritz, Osculatory Interpolation Revisited, *1987 Proceedings of the American Statistical Association, Statistical Computing Section,* 1988.

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

[14] 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.

[15] Internal Revenue Service, *1999 Data Book,* 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.

[16] 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.

[17] 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.

Use of Individual Retirement Arrangements To Save for Retirement--Results From a Matched File of Tax Returns and Information Documents for Tax Year 2001

Peter Sailer, Internal Revenue Service, and Sarah Holden, Investment Company Institute

With $2.6 trillion in assets at yearend 2001, Individual Retirement Arrangements (IRA's) had grown to represent nearly one-quarter of the $11.2 trillion in the U.S. retirement market (Figure 1). Defined contribution plan assets had risen to $2.7 trillion, with 401(k) plans holding an estimated $1.7 trillion, or 15 percent of the total retirement market, at yearend 2001. This year, the Employee Retirement Income Security Act (ERISA) turns 30-years-old. In 1974, when ERISA was passed, the total U.S. retirement market was only $367.5 billion in assets.[1] Thirty years later, U.S. households hold more than $12.1 trillion in retirement assets, with IRA assets exceeding $3.0 trillion.[2] This paper will focus mainly on one of these retirement vehicles--the one for which IRS files contain the most data. This retirement vehicle is the Individual Retirement Arrangement (or IRA).

By combining tax returns and information returns in one database, the Statistics of Income (SOI) Division has made it possible to study trends in contributions to IRA's, as well as the participation in other types of retirement plans, by individual taxpayers.[3] This paper will analyze the detailed SOI data for Tax Year 2001, paying particular attention to comparing taxpayers with IRA activity to the population of taxpayers who were eligible to participate in that tax year. In addition, this paper will show the interaction of IRA activity with employer-provided retirement plans.

▶ All Taxpayers with IRA's

While SOI has collected traditional IRA deductible contribution information for every tax year starting in 1975 (Figure 2), those contributions only tell a very small part of the IRA story. Detailed SOI data from the information Form 5498 reveal a more complete picture. For example, in 2001, while deductible contributions to all IRA's totaled $13.2 billion (including deductible contributions to traditional IRA's of $7.4 billion as shown in Figure 2), an additional $23.4 billion were contributed to IRA's on an after-tax (nondeductible) basis (Figure 3, column 4 minus column 6). More importantly, rollovers, primarily from qualified retirement plans increased IRA holdings by $187.1 billion in 2001. Pulling IRA assets down in 2001 were withdrawals and poor equity market returns, so that, by yearend 2001, total IRA assets had edged down slightly to $2,619.4 billion. While much of this drop can be attributed to reduced returns on capital, it is also true that the level of IRA contributions rose by an anemic one-tenth of 1 percent compared with Tax Year 2000, with contributions to traditional and Roth IRA's actually dropping.[4]

When both traditional and Roth IRA's are considered, any individual with compensation under the age of 70 ½ could make a contribution to an IRA, up to a maximum of $2,000 (or total compensation, if less than $2,000) for Tax Year 2001. In making this computation, non-working married persons could count their spouses' earned incomes as their own for the purpose of making an IRA contribution.

Individuals age 70 ½ or older with earned incomes could not contribute to traditional IRA's, but they could still make payments to Roth IRA's, as long as they had incomes under $110,000 for single people (including unmarried heads of households); under $160,000 for married persons filing jointly (including recently widowed spouses with children); or under $10,000 for married persons filing separately.

The income concept used to determine eligibility was "modified adjusted gross income." This is basically adjusted gross income (or AGI)--the bottom line of page 1 of Form 1040, with a few items added back:

- Deductible IRA contribution(s);

- Student loan interest excluded from AGI;

- Excluded foreign earned income;

- Excluded foreign housing allowances;

- Excluded bond interest;

- Employer-paid adoption expenses.

Overall, only 9.4 percent of those taxpayers eligible to make IRA contributions did so in 2001. When eligible taxpayers are classified by size of adjusted gross income (Figure 4), it turned out that less than 4 percent of eligible taxpayers with incomes under $25,000 actually made contributions. Participation rates gradually rose through the $200,000 under $500,000 class, where about 21 percent of eligible taxpayers contributed, and then declined again for the highest income classes.

When eligible taxpayers are classified by age group (Figure 5), the highest participation rate (over 14 percent) occurred for the 55- to 64-year-old group. Apparently, many taxpayers wait until a fairly advanced age to start making IRA contributions. Participation rates were much lower for taxpayers under 45 and over 70, the latter likely influenced by the age limitation on traditional IRA contributions.

In conclusion, when all types of IRA plans were considered, participation rates tended to rise as income levels rose. However, tax return information repeatedly shows that all income groups take advantage of deductible IRA contributions. Among tax returns with deductible traditional IRA contributions in 2001, about 17.8 percent had AGI of less than $25,000; 32.4 percent had AGI between $25,000 and $50,000; 19.9 percent had AGI between $50,000 and $75,000; and 29.9 percent had AGI of $75,000 or more.[5]

▶ Taxpayers with Deductible Traditional IRA Contributions

The deductible traditional IRA allows eligible taxpayers to deduct the IRA contribution (up to $2,000), and exempts all proceeds from taxation until the money is withdrawn. The eligibility requirements for deductible IRA's are more stringent than those for nondeductible traditional IRA's or Roth IRA's.[6] For Tax Year 2001, the taxpayer eligible for deductible contributions

- Had to have compensation;

- Had to be under age 70 ½;

- Could not be taking the full $2,000 Roth IRA contribution;

- If covered by an employer-provided pension plan, had to have modified AGI of less than:

  ➢ $43,000 if single or unmarried head of household;

  ➢ $63,000 if married filing jointly or a surviving spouse;

  ➢ $10,000 if married filing separately.

There was no income limit for taxpayers who were not covered by employer-provided pension plans, with one exception: if a married person filing jointly was not covered by a pension plan, but his or her spouse was, the noncovered spouse could not make a deductible IRA contribution if the couple's modified AGI was $160,000 or more.

In the charts showing taxpayer participation in deductible traditional IRA plans as a percentage of eligible taxpayers, data are shown separately for covered and noncovered taxpayers, since different rules apply to the two groups. Coverage by an employer-provided plan was determined either by the presence of contributions to a SEP or SIMPLE IRA on Form 5498, or a checkmark in the "Retirement Plan" box of Form W-2.

Overall, only 3.0 percent of eligible taxpayers took a traditional IRA deduction. When taxpayers were classified by coverage/noncoverage by an employer-provided pension, 2.4 percent of the covered and 3.2 percent of the noncovered taxpayers took the deduction. As shown in Figure 6, participation in this program varied considerably over various income levels, with 13 percent of taxpayers in the $200,000 under $2,000,000 class taking the IRA deduction. (The reason such a large income interval was chosen is that there was remarkably little difference in participation rates over this income range.) Obviously, at these income levels, only noncovered employees were eligible to take the IRA deduction.

As shown in Figure 7, the highest participation in deductible traditional IRA's is among those approaching retirement age. In the 55- to 64-year-old age group, 6.5 percent of eligible covered taxpayers took the deduction, as did 8.2 percent of eligible noncovered taxpayers. The

highest age class ends at 70 ½ years, the maximum age at which one could qualify for the deductible traditional IRA contribution.

Figure 8 divides the taxpayer population as a whole (not just the eligible population) into six groups, based on participation in deductible traditional IRA plans. Only 2 percent of the entire population took the deduction for Tax Year 2001. Fully 65 percent of all taxpayers were eligible to invest in deductible IRA's, but did not. Ineligible taxpayers included those with no compensation (12 percent of the population), covered taxpayers above the income limit (17 percent), those over age 70 ½ (2 percent), and those electing to make a full $2,000 Roth IRA contribution instead of a deductible IRA contribution (2 percent).

▶ All Taxpayers with Retirement Plan Accumulations

Taxpayers may accumulate assets for retirement through a variety of tax-advantaged programs. Figure 9 shows that fully 26 percent of the taxpayer population had assets invested in nonemployer-sponsored IRA's. These assets (shown as traditional IRA or Roth IRA fair market value on Form 5498) were accumulated either through contributions to these plans, or through rollovers on job change or retirement from employer-sponsored plans, such as those set up under Section 401(k) of the Internal Revenue Code. Among these 26 percent, 10 percent of the taxpayer population not only had assets invested in nonemployer-sponsored IRA's, but were also participating in employer-sponsored plans, as evidenced by the presence of SEP or SIMPLE IRA contributions on Form 5498, or participation in employer-sponsored plans indicated on Form W-2.

Unfortunately, individuals' assets accumulated in employer-sponsored plans (such as 401(k)s) are not available from any documents in the Internal Revenue Service's record system. However, it seems safe to assume that the 9 percent of the population who had no IRA assets or current employer-sponsored plan coverage, but reported taxable pension income on their Forms 1040, had assets (or at least obligations) from employer-sponsored plans.

All told, IRS tax return and information forms show that, in 2001, about 60 percent of taxpayers had assets in

and/or income from IRA's and/or employer-sponsored plans. Figure 9 shows 40 percent of the population neither receiving nor accumulating retirement assets. Of course, this number refers only to assets officially designated as retirement plans. Many of these individuals may be accumulating interest-bearing or dividend-paying assets, or other assets that can be sold at a future date to fund retirement.

▶ Footnotes

[1] See Federal Reserve Board, *Flow of Funds Accounts*, table L.225, June 10, 2004 release.

[2] See Investment Company Institute (June 2004).

[3] See Sailer, Weber, and Gurka (2003).

[4] For Tax Year 2000 contribution details, see Sailer and Nutter (Spring 2004).

[5] See Campbell and Parisi (Fall 2003).

[6] See Internal Revenue Service (2001) for details.

▶ References

Campbell, David and Parisi, Michael, "Individual Income Tax Returns, 2001," *SOI Bulletin*, Fall 2003, pp. 8–45.

Federal Reserve Board, *Flow of Funds Accounts*, Board of Governors of the Federal Reserve System, June 10, 2004.

Hrung, Warren B., "Information, the Introduction of Roths, and IRA Participation," *Contributions to Economic Analysis & Policy*, Volume 3, Issue 1, Article 6, 2004.

Internal Revenue Service, *Individual Retirement Arrangements (IRA's), Publication 590*, for use in preparing 2001 returns.

Investment Company Institute, "Mutual Funds and the U.S. Retirement Market in 2003," *Fundamentals*, Volume 13, Number 2, Investment Company Institute, June 2004.

Sailer, Peter J. and Nutter, Sarah E., "Accumulation and Distribution of Individual Retirement Arrangements, 2000," *SOI Bulletin*, Spring 2004, pp.121–134.

Sailer, Peter J.; Weber, Michael E.; and 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," *National Tax Association Proceedings, Ninety-Fifth Annual Conference, 2002,* Orlando, FL, pp. 364–369.

Sailer, Peter; Gurka, Kurt; and Holden, Sarah, "Accumulation And Distributions Of Retirement Assets, 1996–2000--Results From a Matched File of Tax Returns and Information Returns," *2003 Proceedings of the American Statistical Association*, Alexandria, VA.

Smith, Paul A., "Retirement Saving over the Long Term: Evidence from a Panel of Taxpayers," *Working Paper*, Department of Treasury and Federal Reserve Board, Washington, DC, December 2002.

▶ 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
U.S. Retirement Market, 2001

Annuities* ($0.9 trillion) 8.5%

IRAs ($2.6 trillion) 23.4%

Federal, State, and Local Government Pension Plans ($3.1 trillion) 27.6%

401(k) Plans ($1.7 trillion) 15.4%

Private Defined Benefit Plans ($1.8 trillion) 16.3%

Other Defined Contribution Plans ($1.0 trillion) 8.8%

Total: $11.2 trillion
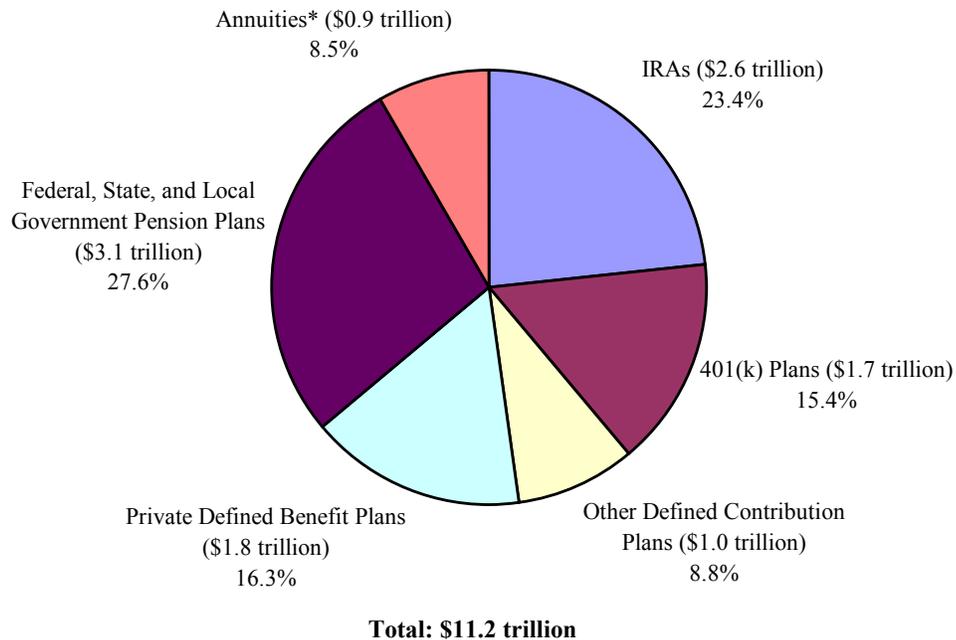
*Does not include annuities held in IRAs, 403(b) plans, 457 plans, or private pension plans.

Sources: Investment Company Institute, Internal Revenue Service, Statistics of Income Division, and Federal Reserve Board

Figure 2
Deductible IRA Contributions to Traditional IRAs,* 1975–2001
(billions of dollars)

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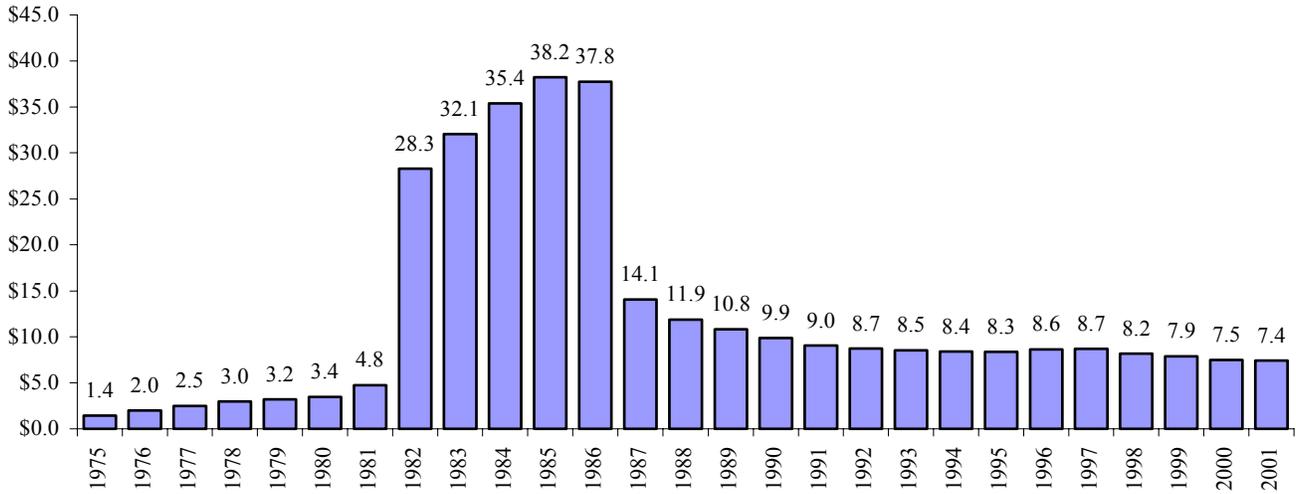
Source: IRS, Statistics of Income Division, *Individual Income Tax Returns, Publication 1304* , various years, and *SOI Bulletin.*

Figure 3. Individual Retirement Arrangement (IRA) Plans by type, Tax Year 2001

Type of plan	Beginning of year FMV		Total contributions		Deductible on Form 1040		Rollovers	
	Number of Taxpayers	Amount	Number of Taxpayers	Amount	Number of Taxpayers	Amount	Number of Taxpayers	Amount
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total	46,270,141	2,629,309,067	15,987,806	36,524,664	4,504,937	13,167,381	3,602,806	187,080,603
Traditional IRA Plans	38,076,500	2,407,022,354	5,583,757	9,825,898	3,718,917	7,406,866	3,602,806	187,080,603
SEP Plans	3,313,204	134,047,902	1,786,931	10,071,870	642,053	4,991,601	n/a	n/a
SIMPLE Plans	1,568,426	10,351,751	1,728,736	5,468,896	143,966	768,913	n/a	n/a
Roth IRA Plans	9,485,189	77,579,420	6,806,294	11,116,124	n/a	n/a	n/a	n/a
Education IRA Plans 3/	241,238	307,640	82,088	41,876	n/a	n/a	n/a	n/a

Type of plan	Roth conversions		Withdrawals 1/		Other changes 2/ Amount	End of year FMV	
	Number of Taxpayers	Amount	Number of Taxpayers	Amount		Number of Taxpayers	Amount
	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Total	0	0	9,185,958	104,527,365	-129,010,549	48,404,401	2,619,376,420
Traditional IRA Plans	255,062	-3,052,037	8,553,004	98,690,314	-107,320,567	39,283,457	2,394,865,938
SEP Plans	n/a	n/a	342,199	4,452,660	-8,305,687	3,523,805	131,361,424
SIMPLE Plans	n/a	n/a	98,049	471,710	-1,756,655	1,959,748	13,592,282
Roth IRA Plans	255,062	3,052,037	370,077	875,818	-5,874,730	11,026,390	79,349,804
Education IRA Plans 3/	n/a	n/a	73,919	36,863	-105,681	206,655	206,972

Note: Except as noted, all data are from matched Forms 1040 and 5498; all figures are estimates based on samples--amounts in thousands of dollars.

1/ Withdrawals are reported on Form 1099-R; excludes withdrawals for the purpose of rollovers to other IRA accounts, or Roth IRA conversions.

2/ Residual of change in fair market value minus all the enumerated changes.

3/ Education IRAs were renamed Coverdell Education Savings Accounts (ESAs) in July 2001; excludes Ed-IRAs owned by non-filing dependents.

Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Figure 4: Percent of Eligible Taxpayers Contributing to Any Type of IRA Plan by Size of AGI, Tax Year 2001

Percent

- All types of IRAs

AGI	Percent
Under $25,000	3.8%
$25,000 under $50,000	8.7%
$50,000 under $75,000	11.5%
$75,000 under $100,000	14.6%
$100,000 under $200,000	20.7%
$200,000 under $500,000	21.1%
$500,000 under $1,000,000	20.5%
$1,000,000 or more	16.9%

Adjusted Gross Income (AGI) on Tax Return
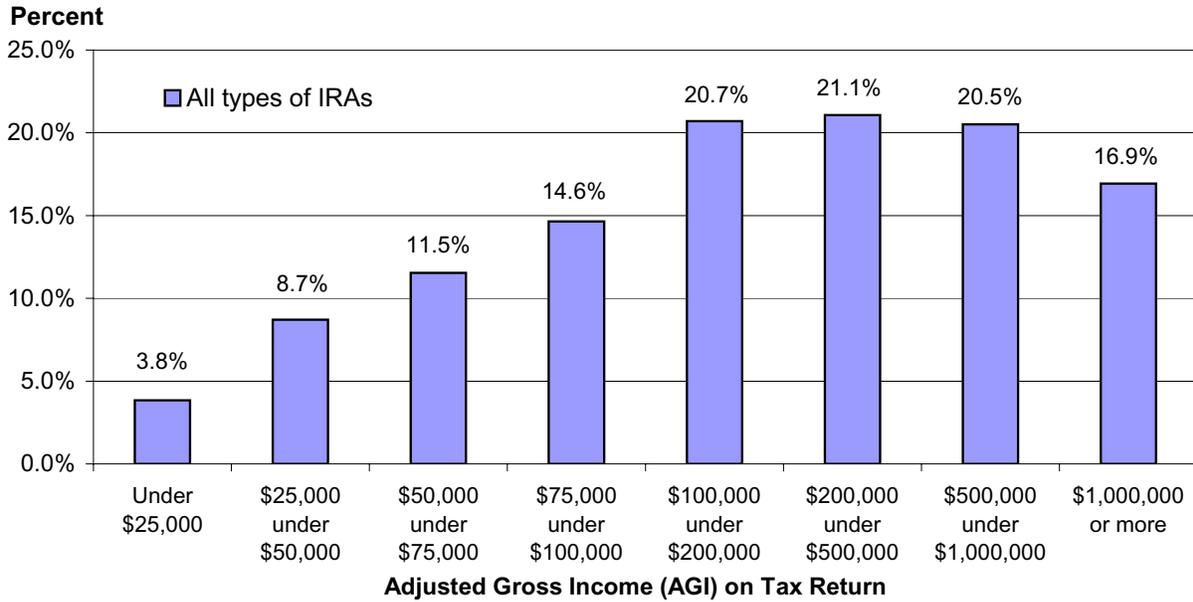
Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Figure 5: Percent of Eligible Taxpayers Making IRA Contributions by Age, Tax Year 2001

Percent

- All types of IRAs

Age of Taxpayer	Percent
< 20	2.6%
20 to 24	4.3%
25 to 34	8.0%
35 to 44	9.6%
45 to 54	12.2%
55 to 64	14.6%
65 to 70	11.8%
70 & over	2.3%

Age of Taxpayer

Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Figure 6: Percent of Eligible Taxpayers Taking Traditional IRA Deduction by AGI and Employer-Provided Retirement Plan Coverage, Tax Year 2001
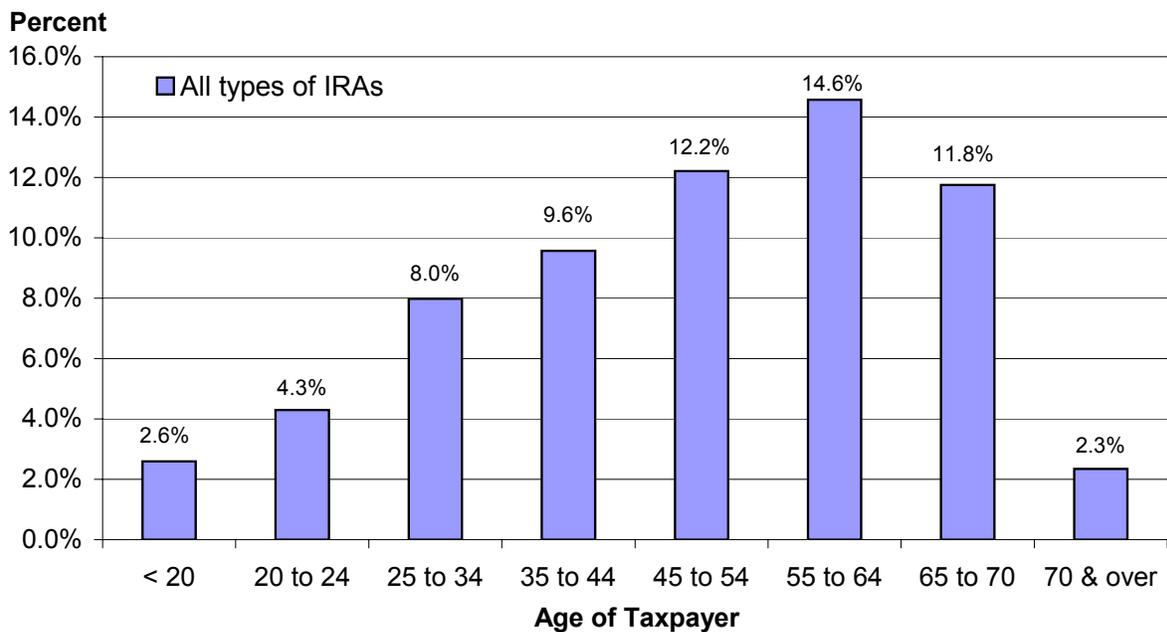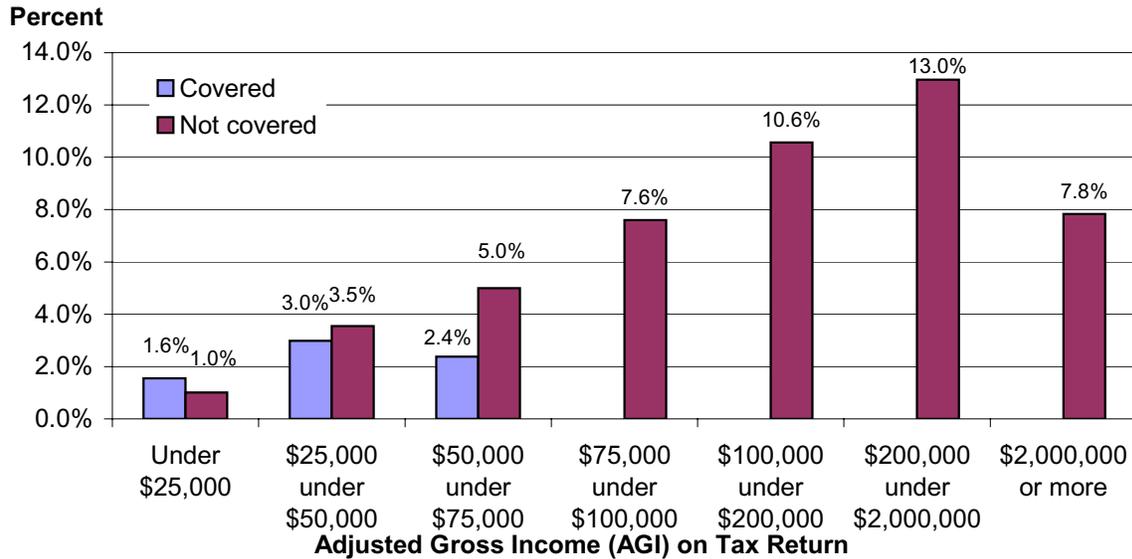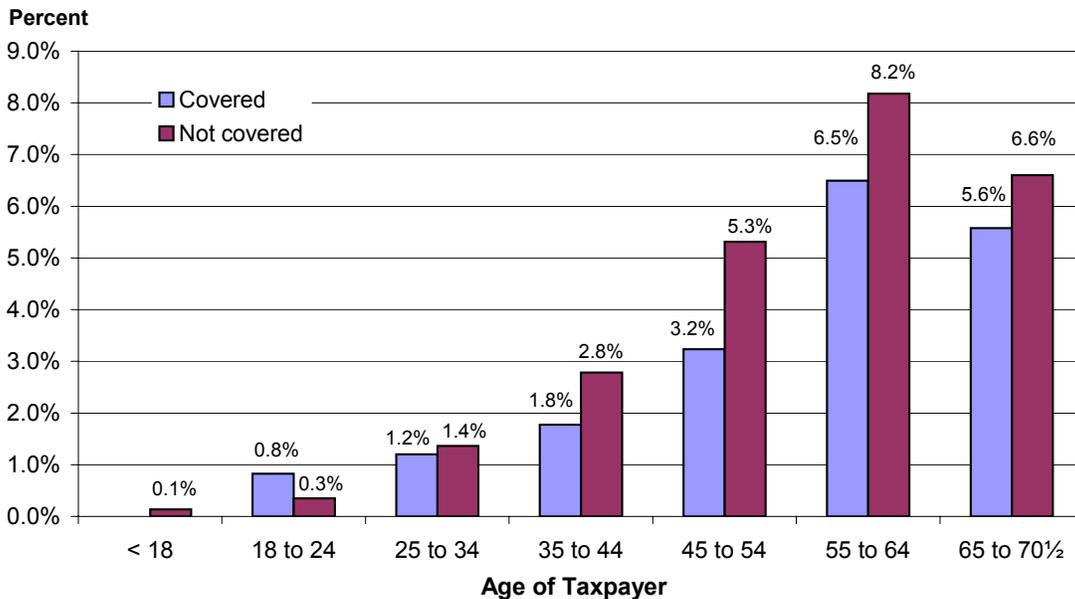
Percent

Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Figure 7: Percent of Eligible Taxpayers Taking Traditional IRA Deductions by Age and Employer-Provided Retirement Plan Coverage, Tax Year 2001

Percent

Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Figure 8: Percent of All Taxpayers by Eligibility for IRA Deductions, Tax Year 2001

☐ Eligible, taking deduction

▦ Eligible, not taking deduction

▨ No taxable compensation

☐ Covered and above income limit

▤ Older than 70½

▥ Full Roth Contribution Instead

2% 2% 2%

17%

12%

65%
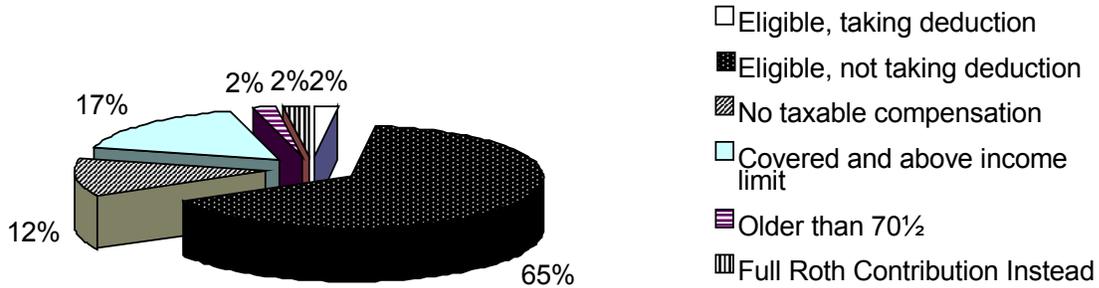
Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Figure 9: Percent of All Taxpayers by Type of Retirement Plan Participation, Tax Year 2001

▤ % with IRA FMV only (Trad. and Roth)

☐ % with employer coverage only

▨ % with both

☐ % with neither, but with taxable pension income on tax return

■ % with no coverage, FMV, or pension income

16%

40%

25%

9%

10%
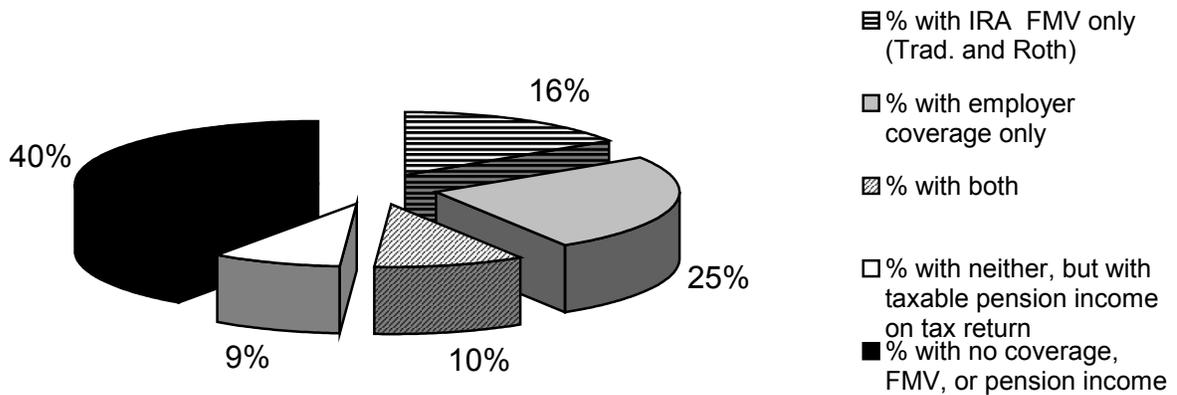
Source: Matched file of income tax returns, Forms 5498, and 1099-R for Tax Year 2001

Customer Satisfaction Initiatives Within the Statistics of Income Division of the Internal Revenue Service

Kevin Cecco, Internal Revenue Service

The Internal Revenue Service's (IRS) Statistics of Income (SOI) Division has a long history of collecting and disseminating critical tax statistics. The SOI function goes back to the enactment of the modern income tax in 1913. It was documented 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." Today, SOI conducts tax studies on the operations of the tax laws with respect to individuals, corporations, partnerships, sole proprietorships, estates, nonprofit organizations, and trusts, as well as inbound and outbound international activities.

Measuring customer satisfaction continues to play an important role for many of the Federal statistical agencies. SOI relies on results from a number of satisfaction surveys to assess its communication and feedback, as well as evaluate the information and services provided to its customers. SOI has made a commitment to administer satisfaction surveys for its primary customers at the Office of Tax Analysis (OTA), the Joint Committee on Taxation (JCT), and the Bureau of Economic Analysis (BEA), as well as for internal employees and customers within the IRS. As a critical source of valuable information, the surveys allow SOI to tailor data collection, analysis, and dissemination efforts more effectively. This paper will focus on providing an historical perspective of conducting customer surveys, summarizing results from several customer surveys, and offering future plans for expanding customer satisfaction initiatives within SOI.

▶ Background Information

In an effort to emphasis the critical importance of collecting and evaluating customer feedback, SOI has made a commitment to collect customer satisfaction data on an annual basis. This commitment allows SOI the opportunity to assess the quality of service to its primary stakeholders. The commitment of conducting the surveys also illustrates SOI's seriousness in continually improving the level of service to its customers.

SOI's commitment to collecting customer feedback was reignited in 2000 when a decision was made to collect survey data in the Treasury Department at OTA. Additionally, the Statistical Support Section (SSS) of SOI, a group of mathematical statisticians providing statistical support to various organizations within the IRS, started collecting feedback from its customers in 2000. In 2002, SOI expanded the customer satisfaction survey process to include customers at JCT and BEA.

In an effort to widening the scope of gathering customer feedback even further, the customer satisfaction initiative was expanded in 2003 to individuals contacting SOI's Statistical Information Services (SIS) office. The SIS office was established approximately 15 years ago to facilitate the dissemination of SOI data and reports and respond to all data information requests. Since its origin, the SIS has established a reputation for always providing an answer or referral to the many challenging tax statistics questions of the general public.

▶ Capturing Critical Information from the Customer Satisfaction Surveys

The various SOI customer satisfaction surveys are designed to be relatively brief and visually engaging, encouraging respondents to participate in the survey process. The surveys have two goals. First, the surveys are designed to collect critical information about the services that staff provides to its primary customers. One way SOI can strive to improve satisfaction is by collecting customer feedback, identifying customer needs, and determining how well it is meeting their needs. Second, the surveys are designed to measure the overall customer satisfaction with SOI's products, services, and personnel.

The OTA, JCT, and BEA surveys are composed of five sections--four of which include customer contact, staff characteristics, product opinion, and overall satisfaction. Each question is designed to obtain feedback on specific indicators in the different areas. The survey also includes two open-ended questions that elicit miscellaneous comments regarding relationships between SOI and its customers.

The Statistical Support Section of the SOI also surveys its primary customers in the IRS. The SSS works as consultants, providing statistical assistance to various organizations outside of SOI, but within the IRS. Similar to the OTA/JCT/BEA objectives, the SSS survey goals are to measure the level of service provided by SSS statisticians, as well as gauge overall customer satisfaction. Further, the SSS survey includes questions measuring the extent of customer contact, staff characteristics, product opinion, and service improvements.

The Statistical Information Services (SIS) started surveying its customers during the spring of 2003. SIS is responsible for fielding inquires regarding data produced and published by SOI and other organizations in the IRS. It receives inquiries most often by telephone, e-mail, and fax, but is also contacted via mail or through face-to-face visits. The goals of the SIS survey are to measure the level of customer satisfaction concerning services provided to customers during the most recent inquiry, identify problems that customers encounter when contacting SIS, and improve the tools and products customers access while searching for IRS data.

▶ Methods of Data Collection

In 2003, the mode of collecting OTA responses was using a paper survey. Hard copies of the survey, along with a cover letter explaining the importance of the data collection, have been handcarried from SOI to OTA. Respondents completed and returned the surveys in interoffice envelopes. In 2004, the OTA and JCT surveys were administered electronically. Respondents downloaded and completed the survey, then e-mailed the completed survey back to SOI.

The BEA survey was administered in an electronic fashion over the past 2 years. An electronic cover let-

ter, explaining the intent of the survey, was e-mailed to the respective individual with the survey electronically attached. Like the OTA and JCT surveys, the BEA respondents downloaded and completed the survey, then e-mailed the completed survey back to SOI.

The SIS survey was also electronically administered last year. The survey was e-mailed to all individuals who contacted SIS between March and June 2004. The respondents were requested to download the survey, complete the various questions, and return the survey by attaching it to an e-mail message.

The SSS survey has been administered in a Web-based environment over the past several years. Since all of the SSS customers are internal IRS employees, the SSS has the opportunity to survey its customers using Web-based technology. The respondents are requested to click on an attached Web link that leads them to the SSS survey. Respondents simply point and click through a series of questions on the Web-based survey. Once the survey is completed, responses are electronically submitted to a desired database.

▶ Summary of Results from 2003 and 2004 OTA, JCT, and BEA surveys

Table 1 highlights the number of surveys distributed to customers at OTA, JCT, and BEA, as well as the response rates for the 2003 and 2004 customer surveys. Although the number of respondents at JCT and BEA is small, collecting and assessing data from these organizations are of critical importance to SOI. Their response rates vary from a low of 38 percent on the 2004 OTA survey to a high of 93 percent on the 2003 BEA survey.

Table 1--Response Rates for OTA, JCT, and BEA Customer Surveys

	Surveys Distributed		Response Rate	
	2003	2004	2003	2004
OTA	47	47	55%	38%
JCT	15	14	87%	79%
BEA	14	15	93%	87%

Discussions are under way to address and remedy the precipitous drop in the response rate of the OTA survey. In addition, discussions have been ongoing to determine the appropriate universe of individuals who should receive customer satisfaction surveys from SOI. The lack of frequent contact with SOI products and staff has been correlated with lower response rates.

Table 2 highlights the usefulness of SOI's data and products. Results from this survey question are included in SOI's scorecard of performance indicators. Specifically, the usefulness question elicits the extent of agreement with SOI products and services meeting customer needs. In all three surveys, the extent of agreement (combination of agree and strongly agree) with SOI's products and services meeting the needs of OTA, JCT, and BEA was over 80 percent.

Table 2--Usefulness of SOI's Data and Products

Percent within Office Surveyed		Office Surveyed		
		OTA	JCT	BEA
		2003 / 2004	2003 / 2004	2003 / 2004
Product Met Customer Needs	Strongly Agree	54% / 23%	39% / 9%	46% / 39%
	Agree	42% / 61%	46% / 81%	46% / 46%
	Not Sure	0% / 6%	8% / 0%	0% / 0%
	Disagree	4% / 0%	0% / 9%	0% / 8%
	Strongly Disagree	0% / 0%	8% / 0%	8% / 0%

Table 3 highlights the overall customer satisfaction rates from OTA, JCT, and BEA for the years of 2003 and 2004. As the data reveal, all three customers provided very positive opinions regarding overall satisfaction with SOI. Interestingly, the customer satisfaction rates have remained fairly constant over the past several 2 years.

Table 3--Overall Satisfaction with SOI

Percent within Office Surveyed		OTA	JCT	BEA
		2003 / 2004	2003 / 2004	2003 / 2004
Overall Satisfaction	Totally Satisfied	65% / 56%	54% / 55%	54% / 69%
	Somewhat Satisfied	23% / 28%	31% / 45%	39% / 31%
	Neither	4% / 6%	8% / 0%	8% / 0%
	Somewhat Dissatisfied	4% / 0%	0% / 0%	0% / 0%
	Totally Dissatisfied	4% / 0%	0% / 0%	0% / 0%
	No Response	0% / 0%	8% / 0%	0% / 0%

The survey concluded with several open-ended questions, seeking recommendations and suggestions for providing outstanding service to the customer base. Verbatim responses from the three organizations covered a wide array of concerns and were not terribly specific in nature. Therefore, it became challenging to synthesize these responses into themes where improvements could be easily made. In time, the verbatims were grouped into broad categories. The most recurring themes focused on finding the right balance between quality and timeliness, improving communication when changes in data structure or timing are necessary, and developing additional documentation for data products and services.

▶ **Results from the 2003 and 2004 Statistical Information Services (SIS) Survey**

In order to gauge the level of customer satisfaction with the services provided to its customers, the SIS Office administered a survey that collected information regarding the selected customer's most recent inquiry.

Specifically, the SIS office surveyed customers who either called or e-mailed for some type of assistance. After assisting the customer with the inquiry, SIS e-mailed a survey to the particular customer and asked that it be completed and returned. Gathering this information provided an opportunity for SIS to assess the level of customer satisfaction with products and services offered by its staff. Additionally, the results provided insight as to what SIS could do to improve its products and services.

SIS receives data inquiries, along with other statistical and tax-related questions, from a wide variety of customers. Most of the questions received by SIS come in the form of phone calls or e-mails. In 2003, a sample of customers was offered an opportunity to complete the survey. In 2004, all eligible customers contacting SIS between March and June were offered an opportunity to complete the survey. Table 4 highlights response rates from the 2003 and 2004 SIS survey.

Table 4--Survey Response Rates for the SIS Customer Survey

	Surveys Distributed	Response Rate
2003	259	55%
2004	425	43%

As shown above, the overall response rate dropped between 2003 and 2004. Declines in response rates remain a challenging problem with many Federal agencies. SOI needs to explore all viable options at its disposal to reverse this discouraging, downward trend in response rate. In fact, several steps are currently under way to address the drop in participation in the survey. Possible changes being considered for the upcoming 2005 survey include developing a multimode survey and updating the format of the survey.

Table 5 compares the usefulness of SIS's data and products between 2003 and 2004. Specifically, the usefulness question asked if products and services produced by SOI met the needs of the customer. As previously mentioned, results from this question are incorporated into balanced measures that SOI collects and disseminates on a quarterly basis.

Table 5--Usefulness of SIS's Data and Products

Product Met Customer Needs	2003	2004
Strongly Agree	52%	43%
Agree	30%	33%
Not Sure	5%	10%
Disagree	8%	7%
Strongly Disagree	5%	7%

Table 5 shows a slight decline in results between 2003 and 2004. Overall, the percentage of customers either agreeing or strongly agreeing that SOI's products and services met their needs decreased from 82 percent in 2003 to 76 percent in 2004.

Table 6 highlights the overall satisfaction with SIS. The table provides satisfaction rates for 2003 compared to 2004. As the table shows, customer satisfaction rates, defined as either totally or somewhat satisfied, remained constant over the past 2 years. The overall rating of customer satisfaction was 85 percent in 2003, compared with a satisfaction rating of 86 percent in 2004.

Table 6--Overall Satisfaction with SIS

Percent within Office Surveyed		2003	2004
Overall Satisfaction	Totally Satisfied	54%	44%
	Somewhat Satisfied	31%	42%
	Neither	8%	10%
	Somewhat Dissatisfied	0%	3%
	Totally Dissatisfied	0%	1%
	No Response	8%	0%

▶ **Results from the 2003 and 2004 SSS Survey**

In September of 2003 and 2004, the Statistical Support Section (SSS) distributed a survey to its customers throughout the various organizations of the Service. The survey was administered using Web-based technology, meaning customers opened a Web link attached to an e-mail message. The respondent simply pointed and clicked through the survey, then submitted responses to an SOI server. A quick summary of results from the surveys is provided below.

Table 7 highlights response rates from the 2003 and 2004 SSS surveys. Not surprisingly, the response rate for the 2003 and 2004 surveys are nearly identical.

Table 7--Response Rates for the Statistical Support Section Customer Survey

	Surveys Distributed	Response Rate
2003	90	74%
2004	103	75%

Table 8 highlights the usefulness of data and products produced by the Statistical Support Section. As mentioned earlier, the SSS works as consultants, providing statistical assistance to various organizations outside of SOI. These consultants provide guidance and expertise related to sampling, questionnaire design, cognitive research, and other analytical services. Obtaining a rating of the usefulness of the products and services provided by this group is of utmost importance to the group's manager. Comparing results between 2003 and 2004 reveals a slight decline in the overall usefulness rating of products and services. The rating was 97 percent in 2003, compared to 95 percent in 2004.

Table 8--Usefulness of SSS's Data and Products

Percent of respondents indicating…		Statistical Support Survey	
		2003	2004
Product Met Customer Needs	Strongly Agree	73%	70%
	Agree	25%	25%
	Not Sure	2%	3%
	Disagree	0%	1%
	Strongly Disagree	0%	1%

Table 9 highlights the overall satisfaction with SSS. The table provides customer satisfaction rates for 2003 and 2004. For the most part, customer satisfaction rates remained exceptionally high over both years. The overall rating of satisfaction was 99 percent in 2003, compared with a satisfaction rating of 98 percent in 2004.

Table 9--Overall Satisfaction with SOI

Percent within Office Surveyed		Statistical Support Survey	
		2003	2004
Overall Satisfaction	Totally Satisfied	87%	91%
	Somewhat Satisfied	12%	7%
	Neither	0%	2%
	Somewhat Dissatisfied	0%	0%
	Totally Dissatisfied	1%	0%
	No Response	0%	0%

▶ Future of Collecting Customer Feedback Within Statistics of Income

SOI has recently expanded the survey satisfaction initiative to readers and users of the *SOI Bulletin*. The 2004 Summer *SOI Bulletin* includes a short, one-page customer satisfaction survey. Similar to the other surveys mentioned in this article, the *SOI Bulletin* survey collects feedback from customers who receive the *SOI Bulletin* on a regular basis. Respondents are asked to tear out, complete, and return the perforated survey. Results from the survey will be used to make necessary improvements to the *SOI Bulletin*. The *SOI Bulletin* survey is also being distributed to a select group of advisors who provide valuable opinions and advice to SOI. Results from the *SOI Bulletin* survey will be summarized in early 2005.

A commitment has also been made in gathering customer feedback regarding internal SOI Technical LAN and computer support. Recently, SOI distributed an electronic customer satisfaction survey to its employees. The purpose of the survey is to gather data on the quality and level of service by SOI's Technical team. Results from the survey are currently being tabulated and analyzed. Final results should be available by December 2004.

Finally, SOI is working toward surveying customers visiting the TaxStats Web site. In the fall of 1996, a select group of IRS products became available to the public on TaxStats. Over the years, TaxStats has grown dramatically--now an integral part of the IRS. Capturing opinions and perceptions from TaxStats users is the next logical area for SOI's customer satisfaction focus. Specifically, plans are under way to develop an automated Web-based customer satisfaction survey for TaxStats customers. If all goes as planned, SOI may be able to conduct a Web-based survey by the end of 2005.

▶ Conclusions

Measuring customer satisfaction will continue to be a major priority for SOI. A commitment of collecting and evaluating customer satisfaction data will ensure that SOI does not lose focus on critical issues that impact its primary customers. Furthermore, an emphasis on collecting customer satisfaction data will reinforce the SOI culture of providing outstanding service to customers. As is evident from the data presented in this paper, SOI has done a very good job of exceeding the expectations of its customers. However, SOI should not rest on its successes, but rather work even harder to ensure that it meets or exceeds the many expectations of its customers.

2

Quality Assessment of Administrative Records Data

McMahon
Ludlum
Johnson ◆ Jacobson
Henry ◆ Ahmed ◆ Legel
Rosage

Assessing Industry Codes on the IRS Business Master File

Paul B. McMahon, Internal Revenue Service

An early process in the development of any business survey is the construction of a sampling frame, and a list of establishments is usually the preferred frame. The most favored sources for such a frame are records systems with lots of auxiliary information, which permit stratification, probability proportional to size sampling, calibration estimation, and other options. The Internal Revenue Service's Business Master File System is one such source.

The records on that system are not available to any who would survey this population, but the laws do provide that certain agencies do have access. Limited data are available to the Census Bureau, for example. However, the Service's Master File Systems are designed with accounting and administration in mind, not survey sampling. Thus, there are a number of conventions that, if not understood, could degrade the usefulness of records from that system.

These issues were addressed in past papers, most recently in the areas of processing conventions (McMahon, 1999), delayed filing effects (McMahon, 2002), and regulatory exemptions (McMahon, 2003). Another issue is the quality of the data on that system when the information is not directly connected to matters of tax collection, but is of considerable interest for a sampling frame. One such variable is the industry code.

We examine this code using records processed during Calendar Year 2003 both because it is the latest full year available and because it shows the effects of the latest revisions to the North American Industry Classification System (NAICS). Since Corporation data for Tax Year 2002 are not available as of this writing, we confined this review to businesses organized as partnerships.

▶ Sources of the Data

The records that the Service provides for use in sampling frames arise from the filing of tax forms. In this particular case, we are concerned with the annual records filed on Form 1065, *Partnership Return on Income*. The entities providing these forms are businesses that have two or more owners and are not incorporated, though there are a small number of exceptions.

The exceptions involve some legal forms of business permitted by some States, like "Publicly Traded Partnerships" and "Limited Liability Companies." The existence of these variations on the partnership theme arises from the power of the States under the constitution, which means that the Federal Government must deal with the consequences, in this case by having these hybrid organizations file the partnership form.

That form has four pages, although attachment pages, such as Schedule K-1, *Partner's Shares of Income, Credits, Deductions, Etc.* (one for each partner), and depreciation forms are usually present as well. The associated instructions for the basic form are 34 pages in length, including the mailing instructions and industry classification rules. Contrast this with the 42 pages devoted to the short title list in the 2002 manual for NAICS. In the full classification system, there are 1,179 separate industries, which are far too many to expect the taxpayer to search through [1] and would cost too much to mail to each requestor. As a result, the Service reduced this list to 427 six-digit industry codes that list in just three pages of the instructions.

The industry codes used by the Service differ only by combining industries into more general categories. That is, the Service did not create any special group from a subset of one of the NAICS codes. Moreover, with the exception of the sole proprietorships, the Service uses the same codes across the various types of businesses.

Businesses, however, do change their focus from time to time, and this might result in a change of industry. For example, a company might build residences, rent models, and sell completed units. Depending on

the circumstances, then, it could be in one of three industries. The IRS instructions set the rule that the code to be assigned depends on the activity that provides the greatest share of a firm's total receipts.

Total receipts, however, appear nowhere on the tax form. Instead, a detailed computation is required that requires 17 amounts from three schedules, which in turn reference still other forms and schedules [2].

Taken together, the long list of codes and the complicated process of deciding the industry, as well as the taxpayer's time, make it very likely that the code used in a previous year will simply be copied onto the current version of the tax form. This is a process quite like that used by the various Individual Income Tax softwares, which, while consistent over the years, may not reflect the current status. This situation may well explain why roughly 4,000 partnership returns were received during 2003 with industry codes that were based on the obsolete Standard Industrial Classification (SIC) codes (see Table 1, below).

Table 1: Tax Year 2003 Partnerships: Transaction Records Validity

	Number	Proportion
Valid NAICS	2,297,000	95.9%
Valid SIC	3,700	0.2
Invalid NAICS	95,000	4.0
Invalid SIC	600	--
(*Proportions do not add to 100% due to rounding*)		

Although only a small proportion of the partnership returns are filed electronically, in order to use the data effectively in a sampling frame, the data must be accessible in that format. This means that the paper returns must be transcribed, at least in part. In practice, and as we have noted elsewhere, only a relatively small number of items are abstracted, but the industry code is one of them.

Sometimes, the respondent's handwriting is illegible, or they have provided clearly incorrect values. Those cases are directed to a reviewer for correction, though that may result in assigning a code "999000" for

"unknown." This may occur more frequently during periods where large numbers of records must be processed, but we have not examined this possibility.

For administrative reasons, the electronically-filed returns are automatically edited to include the same data items as those abstracted from the paper returns. The resulting records are known as "Transaction Records," following the usage in accounting practice.

The validity code on which Table 1 depends is the result of a simple test of whether a given industry code entry is on a list, and does not mean that the code is appropriate for the firm in question. Ascertaining the verity of a code for any particular record would require a separate source of that information.

Fortunately, there are other sources for an industry code available on the sampling frame. Once a partnership transaction record is complete and passed a series of perfunctory tests, it is ready for a process called "Posting." This process involves matching a transaction to a Business Master File Account based on the Employer Identification Number and selected other data, updating that account, and transferring some information to the transaction. We are interested here in the "Entity" part of the data, which includes such items as the name and address for contacting the firm, and an industry code. (We will, henceforth, refer to this code as the "Entity NAICS" code to distinguish it from the code on the Return Transaction.)

Table 2. Tax Year 2003 Partnerships: Entity Industry Sources

	Number	Proportion
NAICS-Based Codes		
Transaction	2,157,000	90.0%
Social Security	219,000	9.1
Exam	4,900	0.2
Other	30	--
SIC-Based Codes		
Transaction	6,000	0.3
Social Security	600	--
Code Not Available	8,800	0.4

The information from the Social Security Administration is introduced at the time a firm receives an Employer Identification Number. Part of the processing of an application at Social Security involves assignment of a NAICS code, which is then passed to the Service along with other data needed to initiate an account.

Revisions to industry codes can arise as part of those administrative actions where agents contact the businesses, and these are grouped under the title "Exam" in Table 2. The other sources are really too small to detail, though they can include information about exempt organizations (since there are no constraints on the nature of an owner of a partnership).

The nearly 9,000 records with an industry code "Not Available" might be those with NAICS codes not on the Service's list. We tested this hypothesis by matching a copy of the 2002 version of these codes to those records. There were no matched records. A manual review of a handful suggests that data from an adjacent area of the return had been erroneously entered as the industry.

While most of the Entity NAICS entries arise from returns, via transactions, the codes are not necessarily from the current tax year. Almost 3 percent of such transactions had either invalid transaction NAICS codes or some SIC-based entry. We know these data must be from another source due to the rules on updating the Master File Accounts.

Those rules for updating the industry on the Master File accounts start with permitting only valid codes to be considered. Next, NAICS-based codes have higher priority than the SIC-based versions. And then, the source matters too: data from Exempt Organizations, over Social Security, over IRS's Examination, over the return transaction, over the occasional information from Collections, in that order. Finally, the posting program selects the code that has the greater specificity if all other factors are equal. (This routine applies to all records that are posted to the Business Master File, not just partnership records.)

In short, the process favors new over old, for greater source reliability (at least in the opinion of those designing the system), and for greater detail over lesser.

Given the strong reliance on information from the tax returns, we would expect significant agreement between the Entity NAICS and the transaction's code. Overall agreement, however, may hide real problems in some sectors.

For the balance of this review, we will confine our attention to the sectors, based on the first two digits of the NAICS Code. In part, this is due to space constraints for this article; but mostly, it is due to concerns about disclosure and the distribution of the Statistics of Income Partnership sample.

Analysis of the Frame

The data in Table 3 are from the sampling frame (not a sample), using the Entity NAICS as the source for the sector, and with records excluded where the industry code is based on the Standard Industrial Classification (SIC) or is invalid. The rate of agreement between the two industry codes is almost 96 percent, which is not too surprising given the source for most of the codes. Over 90 percent of the codes arise from a Return Transaction, though some will be from prior-year records instead of the current tax year. The agreement rate for those records with the industry code arising from the transaction is, unsurprisingly, over 99.9 percent.

The agreement rate for records where the Entity NAICS did not arise from the transaction was 67.4 percent.

Sixteen of the 21 categories shown in Table 3 have agreement rates greater than 90 percent, with 7 higher-than 95 percent. Most of the other groups have rates in the 80-to-90-percent range, and these sectors are among those with the fewest firms. Indeed, the smallest, Public Administration, has the lowest rate of agreement between the two NAICS codes.

This sector, though, would seem to be out of scope for a business survey. It may be that these organizations are charities forming some sorts of joint operations; we cannot tell from the data available, which are too sparse to begin with.

The other "sector" that is out of place is the group of "Unknown" firms. Since these comprise about 4.4

Table 3: Tax Year 2002 Partnerships Sector-Level Agreement Between Industry Codes

2002 North American Industry Code System (NAICS) Title	NAICS	Records With	Entity NAICS from Transaction		Entity and Transaction Sectors Agree	
	Sector	NAICS	Number	Percent	Number	Percent
Agriculture, Forestry, Fishing, and Hunting	11	125,763	119,463	95.0%	123,276	98.0%
Mining	21	26,046	23,700	91.0%	25,530	98.0%
Utilities	22	2,528	2,213	87.5%	2,326	92.0%
Construction	23	133,448	106,613	79.9%	123,180	92.3%
Manufacturing	31-33	40,263	35,101	87.2%	37,427	93.0%
Wholesale Trade	42	35,776	28,013	78.3%	31,310	87.5%
Retail Trade	44-45	124,100	107,755	86.8%	115,394	93.0%
Transportation and Warehousing	48-49	27,922	25,082	89.8%	26,234	94.0%
Information	51	25,585	20,458	80.0%	23,112	90.3%
Finance and Insurance	52	281,027	225,095	80.1%	266,524	94.8%
Real Estate and Rental and Leasing	53	1,008,948	976,126	96.7%	986,818	97.8%
Professional, Scientific, and Technical Services	54	157,084	138,160	88.0%	148,020	94.2%
Management of Companies and Enterprises	55	18,353	15,889	86.6%	15,866	86.4%
Administrative and Support and Waste Management and Remediation Services	56	37,691	26,842	71.2%	30,331	80.5%
Educational Services	61	6,141	4,158	67.7%	5,027	81.9%
Health Care and Social Assistance	62	47,350	40,861	86.3%	45,154	95.4%
Arts, Entertainment, and Recreation	71	33,951	27,696	81.6%	31,598	93.1%
Accommodation and Food Services	72	73,359	67,112	91.5%	70,769	96.5%
Other Services (except Public Administration)	81	70,881	62,192	87.7%	68,148	96.1%
Public Administration	92	48	32	66.7%	30	62.5%
Unknown	99	104,499	104,494	100.0%	103,981	99.5%
Total		2,380,763	2,157,055	90.6%	2,280,055	95.8%

percent of the population, larger than most sectors, the characteristics of this group are of immediate interest. Three main variables are of particular interest: Net Income or Loss, Total Assets, and Total Receipts, because they indicate the size and activity of a firm.

The data in Table 4 depend on the transaction records, and, thus, the monetary variables do have some limitations. For example, some items that would belong in an economic definition of Total Receipts or Net Income/Loss are not available from those records. Still, the main contributing items are present, such as gross receipts and net rent from real estate.

The firms that have an unknown industry have a disproportionate number showing no net income or loss among the items available on the frame. Not only do nearly 85 percent show zero for that amount, but that group provides more than half of the firms without net income or loss during 2002. Even when we exclude those with a zero for that amount, the distribution of net income or loss drops off much more rapidly, at roughly thrice the pace, than for firms with reported industries.

The picture for Total Assets is less clear, but this is due in large part to a regulation that permits firms with less than $250,000 in total receipts and less than

Table 4: Tax Year 2002 Partnerships--Distributions of Firms by Selected Variables					
	All	Valid NAICS		Unknown Industry	
Net Income/Loss		Number	Percent	Number	Percent
-1,000,000 or More	24,094	24,044	1.1%	50	0.0%
-250,000 Under -1,000,000	54,924	54,792	2.4%	132	0.1%
-1 Under -250,000	828,178	821,171	36.1%	7,007	6.7%
0 or Not Reported	*173,815*	*85,554*	*3.8%*	*88,261*	*84.5%*
1 Under 250,000	1,141,527	1,132,816	49.8%	8,711	8.3%
250,000 Under 1,000,000	112,347	112,086	4.9%	261	0.2%
1,000,000 or More	45,878	45,801	2.0%	77	0.1%
Total	2,380,763	2,276,264		104,499	
Total Assets					
0 or Not Reported	*679,896*	*582,588*	*25.6%*	*97,308*	*93.1%*
1 Under 250,000	792,447	787,636	34.6%	4,811	4.6%
250,000 Under 1,000,000	437,614	436,231	19.2%	1,383	1.3%
1,000,000 Under 25,000,000	439,259	438,307	19.3%	952	0.9%
25,000,000 or More	31,547	31,502	1.4%	45	0.0%
Total	2,380,763	2,276,264		104,499	
Total Receipts					
0 or Not Reported	*373,559*	*283,159*	*12.4%*	*90,400*	*86.5%*
1 Under 250,000	1,450,103	1,437,916	63.2%	12,187	11.7%
250,000 Under 1,000,000	347,008	345,586	15.2%	1,422	1.4%
1,000,000 Under 25,000,000	198,720	198,248	8.7%	472	0.5%
25,000,000 or More	11,373	11,355	0.5%	18	0.0%
Total	2,380,763	2,276,264		104,499	

$600,000 in total assets to withhold that information from their filings. The dropoff is not as steep as it is for Net Income, but the effect is still there.

This pattern of concentration at zero with attenuated tails of the distributions continues for Total Receipts. Actually, all but a few hundred of the records that reported no net income or loss also had zeros for amounts of total assets and total receipts.

This raises the question of what industry these firms actually belong in. Remembering that the instructions for filing asks the respondent to use total receipts as the basis, if that amount is in fact zero, then should not the response be "unknown?"

These firms may be characterized as inactive, with the filings being in response to the form the Service mailed. In fact, using the Statistics of Income Partnership Study, we estimate that there are about 137,000 such firms, nearly 27,000 more than the frame counts. The difference is likely due to the variations between the tax law definitions and those based on economic concepts used for the SOI study.

▶ Partnership Sample

Thus far, the discussion has focused on the data from the administrative systems only. If we assume that agreement between the Transaction Record and the Entity NAICS implies validity, then we see that the proportion of partnership records with "valid" industry sectors is about 95.8 percent. Removing those records where the industry is "unknown" only drops this figure to 95.6 percent.

These conclusions rest, however, on a simple list matching, not on inspection of source records. Fortunately, the Statistics of Income Partnership Study for Tax Year 2002 included a significant effort to verify the NAICS codes (though without contacting the respondents). This effort included researching publicly available published and Internet data.

Of the 34,800 records selected for this sample, 33,600 were considered "in scope" and received the extra attention. In the end, only 17 records could not be assigned a NAICS code. The corresponding estimated population for the "unknown industry" is about 2,700, or slightly over 0.1 percent. The coding used the Service's version of NAICS, not the full set of codes.

Note that matching the full NAICS list's 6-digit codes against those assigned to the sample results in about 16,400 records, almost half, being identified as having invalid codes. That is, if the full population were treated as the sample was, about a third (761,000) would not have valid codes under the naïve assumption.

The sample was drawn from the frame, described in the previous section, as the records were filed during 2003. Strata were defined by size of total assets, net income (or loss) or receipts, industry, and select other characteristics of special importance to our sponsors.

We included industry in the design because division level estimates were deemed important. With the real estate leasing businesses comprising over a third of all partnerships, a proportionate distribution of the sample over all the groups would have left several sparsely sampled. Hence, we reduced the sample in real estate

and increased the sample for other industry divisions, and particularly those with few firms. This resulted in a sample with sufficient records at the sector level to assess the accuracy of the NAICS codes, at that level of aggregation, on partnership transaction records.

We compare, in Table 5, the estimated distribution across industry (for active partnerships) using the Entity NAICS codes, and the codes assigned during the data abstraction. The frequencies are quite similar. Most of the estimates using the validated codes are a bit higher than those based on the Entity NAICS, with the greatest proportionate differences in the less populous sectors.

Some difference is expected, of course, because there was a recoding of most of the nearly 40,000 records without a NAICS code. There was also a large movement from "Other Services," which may be what the respondents decided to use when they could not easily find an answer.

However, the similarity of the distributions masks a greater disagreement between the two sets of codes. The overall accuracy drops to 92.5 percent from over 95 percent, but even this needs to be qualified. "Real Estate Rental and Leasing," which contains almost 45 percent of the population, has an error rate of only 1.9 percent. This low error rate is undoubtedly due to the ease that the original coding clerks for the transaction records have in determining an industry: these returns all have Form 8825, *Rental Real Estate Income and Expenses of a Partnership or an S Corporation*, attached.

On the other hand, we should also consider that the category "Other Services" is the equivalent of "miscellaneous." That list of codes is rather long, at three pages; so, having a large number of records from that category being reassigned is to be expected.

Removing those sectors from consideration reduces the overall agreement to only slightly more than 89 percent. "Educational Services" has a small sample, and only a dozen or so were reassigned to other sectors. "Wholesale Trade," however, presents quite a puzzle, with over 100 records reclassified, and only about a third into "Retail Trade" where we might expect them.

Table 5: Tax Year 2002 Partnerships--Sample Estimates of Industry Distribution

2002 NAICS Title	Sector	Entity NAICS	Edited NAICS	Entity & Sample Agree	Error Rate
Agriculture, Forestry, Fishing, and Hunting	11	117,048	117,667	110,941	5.2%
Mining	21	28,095	29,549	27,896	0.7%
Utilities	22	2,331	2,507	2,019	13.4%
Construction	23	126,423	134,114	115,173	8.9%
Manufacturing	31-33	36,787	38,364	33,185	9.8%
Wholesale Trade	42	37,240	37,800	30,470	18.2%
Retail Trade	44-45	118,595	122,013	109,400	7.8%
Transportation and Warehousing	48-49	26,573	26,007	23,569	11.3%
Information	51	23,613	28,580	21,334	9.7%
Finance and Insurance	52	256,820	263,024	248,520	3.2%
Real Estate and Rental and Leasing	53	985,603	999,786	966,940	1.9%
Professional, Scientific, and Technical Services	54	155,372	145,612	133,832	13.9%
Management of Companies and Enterprises	55	17,896	18,773	15,450	13.7%
Administrative and Support and Waste Management and Remediation Services	56	37,794	44,405	30,337	4.1%
Educational Services	61	5,569	6,269	4,575	17.9%
Health Care and Social Assistance	62	46,321	47,468	44,411	4.1%
Arts, Entertainment, and Recreation	71	39,227	42,691	35,859	8.6%
Accommodation and Food Services	72	73,881	77,698	71,099	3.8%
Other Services (except Public Administration)	81	67,177	57,121	49,332	26.6%
Unknown or SIC-Based Code	Unknown	39,804	2,724	2,053	94.8%
Total	All	2,242,169	2,242,169	2,074,342	7.5%

▶ Conclusion

A major reason for this review was to ascertain whether the industry codes on the IRS's Business Master File system for partnerships is sufficiently reliable for stratification purposes. With respect to real estate firms, the quality is quite sufficient, at least for the Entity NAICS. The picture is less clear with respect to those sectors with small populations, where, in some cases, the proportion reclassified is modest, while, in others, the error rates are quite high.

We cannot, of course, generalize to other types of administrative records maintained on the Business Master File, such as Corporation Income Tax Returns, though we note that they appear to have a similar situation with respect to having clearly invalid codes. That investigation will have to be the subject of another paper.

Nor can we attribute the error to any source. The nature of the data before us does not allow us to distinguish between errors by the respondent or the reviewer, except, of course, where the form contains an old SIC-based industry code. This is, however, only a small piece of the non-NAICS coded records.

The sample was too small for more detailed analysis, but it is certain that the finer the coding, the more relative

error we can expect. It is also clear that the methods employed to refine the sample cannot be used on the entire population with any hope of success.

► Notes

[1] North American Industry Classification System, United States (2002), Executive Office of the President, Office of Management and Budget, Introduction, page 16.

[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.

► References

McMahon, Paul (1999), "Administrative Records, Regulations, and Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, Paul (2002), "Proxies in Administrative Records Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, Paul (2003), "Regulatory Exemptions and Item Nonresponse," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Data Interpretation Across Sources: A Study of Form 990-PF Information Collected From Multiple Databases

Melissa Ludlum, Internal Revenue Service

Private foundations contribute billions of dollars each year to charitable initiatives directed toward such issues as environmental protection, health and human services, promotion of the arts and humanities, and educational outreach and opportunities. With several hundred billion dollars in asset holdings, private foundations constitute a substantial segment of the nonprofit sector. Unlike public charities, which are often funded by, and therefore directly accountable to, the public, private foundations generally receive funding from a limited number of sources. Furthermore, an individual or small group typically controls the majority of a foundation's activities. Due to this narrow base of support and control, detailed financial information on private foundations is often more difficult to obtain than similar information for other charitable organizations. In many cases, data collected from tax return records and disseminated by the Internal Revenue Service (IRS) provide the most comprehensive information available on the financial composition and charitable giving habits of private foundations. Statistics derived from these sources can provide a window into the charitable activities of these organizations. Additionally, the information supplied to IRS provides insight into both the investment portfolios of private foundations and into the nature and amount of their charitable and noncharitable expenditures. These data can also reveal emerging trends and developments in the private foundation segment of the nonprofit sector. Analyses conducted using such data provide a framework for the development of tax policy related to private foundations and assist practitioners and foundation staffs in the establishment of key self-governance principles.

Unlike the majority of taxpayers, who report information to IRS on "tax returns" designed to assist in the calculation and payment of income taxes, private foundations complete "information returns" designed to collect a wide range of information. Because of their primarily charitable missions, private foundations receive exemption from Federal income taxes; they are, however, subject to an array of stringent legal requirements. Un-

der regulation, they are required to distribute a certain percentage of their asset holdings to charitable activities each year. Secondly, although private foundations are exempt from *income* tax, they are required to pay an *excise* tax on their investment incomes. In addition, unlike corporate or individual taxpayers, private foundations are subject to public inspection requirements. This means they are responsible for ensuring that their annual information returns, known as Forms 990-PF, are widely available to the public. Each year, private foundations file the extensive, twelve-page return with IRS, reporting standard income statement and balance sheet items, as well as additional information on charitable distributions, compliance with rules that govern private foundations, involvement in various types of activities, and certain employment information.

The public inspection requirement promotes increased data availability and thus provides a wide range of analysis opportunities for interested researchers. Users can obtain microlevel data from Forms 990-PF from a number of sources. For example, independent organizations such as the Foundation Center and GuideStar obtain Forms 990-PF from IRS and post them to the Internet on a continuing basis. Another organization, the National Center for Charitable Statistics (NCCS), makes an annual file of return data from the IRS Returns Transaction File (RTF) available to researchers wishing to obtain data for large numbers of organizations. This file, which the IRS provides to the NCCS annually, includes limited data for the population of Form 990-PF filers. The Statistics of Income (SOI) file provides yet another resource for private foundation data. This file includes error-corrected data items for a sample of Forms 990-PF.

While the numerous available data sources enhance research options, reconciling them to one another can be a difficult experience for data users. Measuring data quality and discrepancies among them is a formidable, but necessary, challenge. Before conducting analysis, researchers should be aware of the range of available

data sources, as well as the limitations and advantages that characterize the data sets obtained from these sources. Such information is especially important when supplementing data from any one source with information obtained from another. Understanding the unique characteristics of data obtained from each source also helps to explain, and reduce, statistical variation between them. Additionally, assessment of these data sources allows opportunities to combine information from them, possibly reducing data collection costs and expediting processes. This paper will discuss two IRS-derived data sources, the IRS Returns Transaction File and the SOI File, and determine the various quality and consistency issues associated with each source. It will describe the various administrative data sources from which private foundation data may be obtained, outline the methodology for identifying comparable tax returns to create a standardized dataset, examine the results of preliminary analysis conducted on aggregate and microlevel statistics from the datasets, and present conclusions and future applications derived from the research conducted.

▶ Data Sources Overview

When IRS receives a Form 990-PF, a limited number of data items are key-entered as the return is processed and posted to what is known as the RTF. IRS creates an annual RTF extract, which includes information from all returns received by IRS during a given "processing," or calendar, year. The extract includes approximately 100 money amounts, or financial items, with an additional

85 fields of codes and other nonfinancial information. When working with RTF-derived data, it is important that users are aware that the file may include a number of superfluous records, such as duplicate or incorrectly filed returns. Under most circumstances, data users should remove such records before conducting most analyses.

When using RTF data, several important factors should be taken into account, particularly if the data are used in conjunction with data from other sources. First, the timeframe that a set of returns represents must be considered. An extract for a given calendar year should include the "population" of Forms 990-PF filed with IRS during that year. However, organizations file Form 990-PF based on reporting year, which corresponds to the year actually printed on the return. As illustrated by Figure A, which shows examples of accounting periods that can be present in a typical Reporting Year, an organization determines its reporting year based on its accounting period, specifically, based on the month in which its accounting period begins. Thus, an organization would file a Reporting Year 2000 return if its fiscal year accounting period *began* in any month of Calendar Year 2000 [1]. However, many Reporting Year 2000 returns, such as those with accounting periods that began in December 2000 and ended in November 2001, would not have posted to the RTF until Calendar Year 2002. When conducting time-series analysis, or analysis among multiple data sources, it is important to understand the relationship between accounting periods, calendar or processing years, and reporting years in order to achieve the most consistent dataset possible.

Figure A: Examples of Filing Periods Represented by Returns Filed for Reporting Year 2000
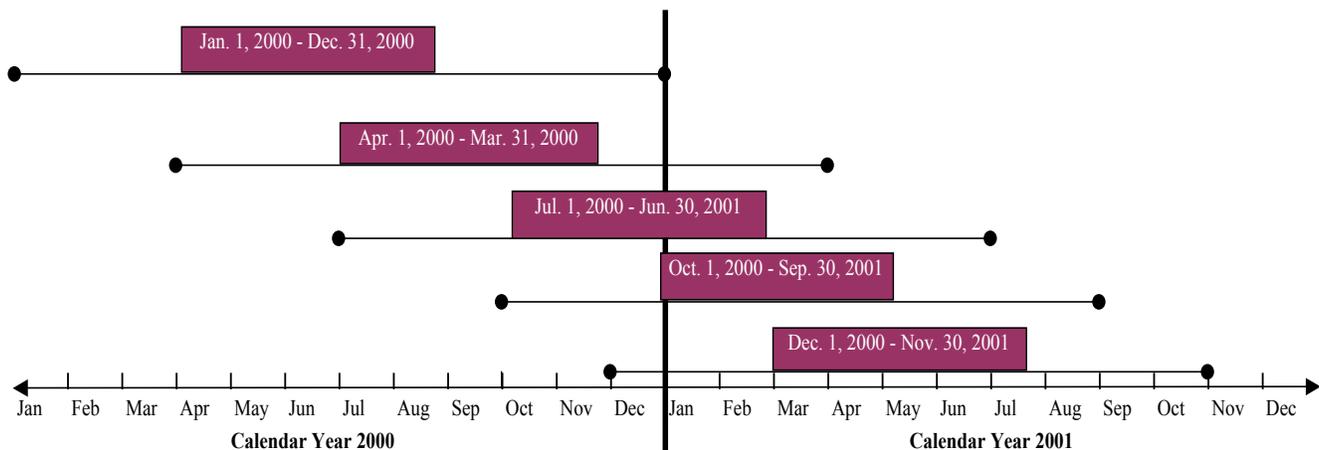
Secondly, although different types of organizations file the same return, they may not necessarily be subject to the same tax treatment. Both tax-exempt private foundations and nonexempt charitable trusts are subject to the private foundation rules and are thus required to file Form 990-PF. However, in some cases, nonexempt charitable trusts may also be responsible for paying income tax, reported on a separate, additional return. Such a distinction could easily affect the behaviors of these organizations. Therefore, these segments of filers should be identified and treated as distinct types of entities, thus allowing the opportunity to examine these data in both separate and aggregate frameworks. If an RTF data user is aware of this distinction, he or she can easily identify nonexempt charitable trusts and private foundations based on their assigned subsection codes.

Based on postings to the RTF, SOI samples approximately 10 percent of all Forms 990-PF filed for a given reporting year. The SOI file contains more than 200 financial items, with 75 fields dedicated to codes or nonfinancial information. The SOI staff enters data into an online system, which identifies taxpayer and other errors, which are corrected during the data entry process. Often, supplemental information is included with Forms 990-PF on schedules and other attachments. Where appropriate, information from these attachments is used to supplement or enhance data reported by the filer. A typical completed reporting year sample includes numerous allocations. For example, SOI made nearly 17,000 allocations for the Reporting Year 2000 sample.

Unlike the RTF extract, which includes all returns filed in a given calendar year, the SOI Reporting Year sample must be conducted over 2 calendar years. This method of data collection is used as it ensures almost complete coverage of a reporting year population, preventing organizations from being excluded from the sample in cases where their returns are filed outside of the anticipated calendar year. Like the RTF, the SOI file includes returns filed by nonexempt charitable trusts, but duplicate returns and returns with inconsistencies that cannot be resolved are removed before dissemination.

▶ Analysis Methodology

The first challenge in measuring consistency and quality issues between the two sources was to standardize and combine the data sources by creating a standardized dataset; the resulting dataset was designed to include data from a single reporting year and to be free of duplicate and extraneous records. To create the standardized dataset, a series of steps was taken to ensure that the highest possible level of consistency was achieved between RTF and SOI data.

The analysis includes returns filed for Reporting Year 2000, which IRS received over several calendar years [2]. To identify the appropriate returns, while still limiting the number of years of RTF data that were included in the analysis, the final dataset was limited to those extracts containing returns posted in Calendar Years 2001 and 2002. This timeframe coincides with the period in which data were collected for the SOI Reporting Year 2000 file.

In addition to including nearly the entire population of timely-filed Reporting Year 2000 Forms 990-PF, the combined extract also included returns filed for other reporting years between 1998 and 2001. Figure B shows the percentage of returns from each reporting year that appeared on the 2001 and 2002 combined RTF extract. Less than half of the returns on the extract represented Reporting Year 2000, and a substantial number were filed for Reporting Year 2001, with smaller but significant numbers representing other reporting years.

Figure C illustrates the number of returns associated with each year in the combined extract. The calendar year populations appear in the larger ovals, with the Reporting Year 2000 subset represented by the smaller ovals. Only the 72,559 returns filed for Reporting Year 2000, identified as the sum of the subsets of the 2 calendar years, and represented in the smaller ovals, were initially considered for inclusion in these analyses.

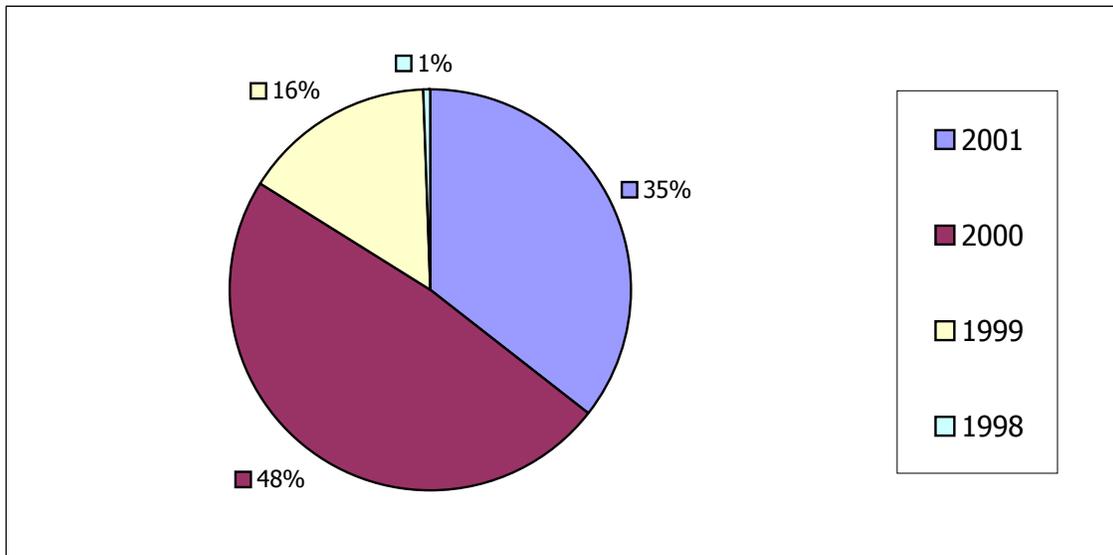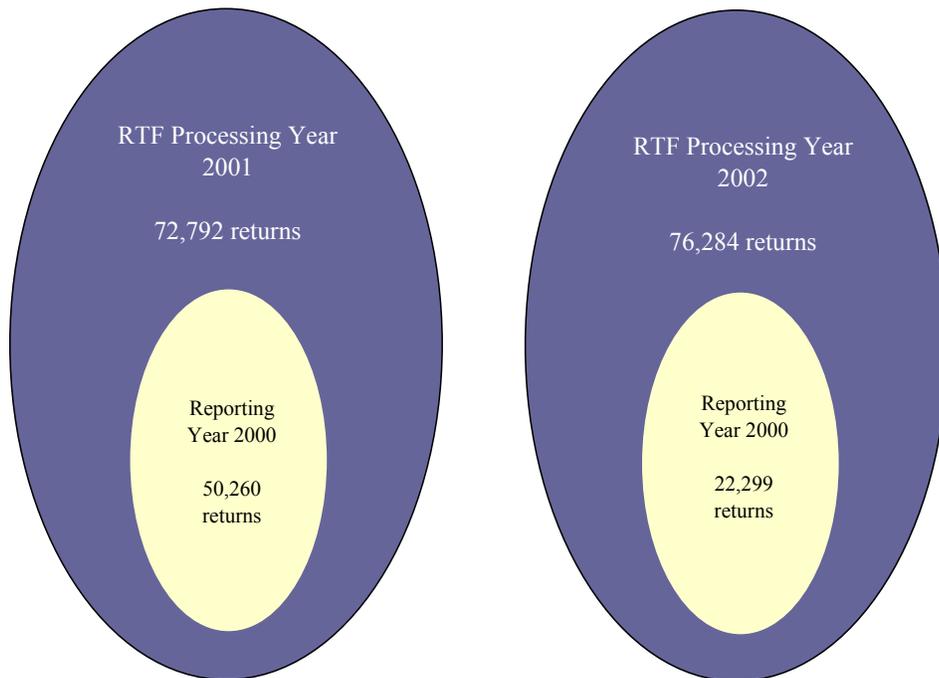## Figure B: Reporting Years Represented in the Combined Extract

- 1%
- 16%
- 35%
- 48%

Legend:
- 2001
- 2000
- 1999
- 1998

Figure C: Components of the Combined Returns Transaction File

RTF Processing Year 2001

72,792 returns

Reporting Year 2000

50,260 returns

RTF Processing Year 2002

76,284 returns

Reporting Year 2000

22,299 returns

Once the subset of included organizations was narrowed, based on reporting year, several additional steps were taken to arrive at a standardized dataset. Records were removed if their associated "status codes" indicated that the organizations were inactive or no longer exempt. In some cases, returns appeared more than once on the RTF. A series of procedures removed these duplicate returns from the standardized dataset. Finally, the completed dataset included only returns filed by private foundations, identified based on the assigned subsection code. Once concluded, these steps revealed an RTF population of 68,355 returns suitable for inclusion in the analysis.

For comparison purposes, the SOI file for Reporting Year 2000 was used for this analysis. The sampling period for the file began in January 2001 and continued through December 2002. The file is a random Bernoulli sample, based on organization type and asset size, using different parameters for private foundations than for charitable trusts. In addition to being subject to different tax treatment than private foundations, nonexempt charitable trusts are generally much smaller, in terms of asset size, than are their tax-exempt counterparts. Private foundations with $10 million or more in assets and nonexempt charitable trusts with $1 million or more in assets were selected at rates of 100 percent, with decreasing rates applied to smaller-sized organizations [3]. For the initial research, the SOI file remained largely intact, with one exception: all returns that were ultimately determined to be "charitable trusts" were removed from the data. While returns filed for charitable trusts were removed from the RTF based on subsection code, they were removed from the SOI file using a more perfected data field, which is not available on the RTF [4]. This field rectifies errors in organization type that are often present on the RTF at the time of sampling.

▶ Aggregate Analysis

After standardization of the data sets, aggregate RTF and SOI data were compared. For major data items, the two sources did not provide significantly different results. Figure D is a comparison between the coefficients of variation, used to estimate SOI sampling error, that were calculated for three major data items, and the percentage differences between estimates derived from the RTF and SOI data files. Note that, for two of the three categories, total revenue and total expenses, the percentage difference between the two datasets falls inside of the sampling error estimates. For one category, fair market value of total assets, the difference by which the RTF amount exceeds the SOI amount is somewhat larger than the sampling error. The larger difference may be attributed to a variety of differences in editing and error correction, which are driven by the purposes for which the data are collected. While RTF data entry operators often key data directly from the Form 990-PF for examination and tax collection purposes, SOI editors may substitute amounts from attachments in lieu of amounts reported on the return. These types of substitutions and corrections allow SOI to produce statistics that are more accurate and to provide additional data items for customers that use microdata files.

▶ Microdata Analysis

To analyze microdata fields between the two datasets, individual returns were linked from the SOI file back to the parent RTF, based on their unique Taxpayer Identification Numbers (TIN's). Returns were not linked unless they appeared on the RTF dataset that was used for aggregate analysis. Once linked, the files were compared for inconsistencies between major data items. The inconsistent fields were then weighted, using the SOI design-based weights, to determine the effects of the SOI correction processes on the overall population estimates. A field was identified as "inconsistent" if the amount transcribed to the SOI file differed by more than $25 from the amount that appeared on the RTF. While

Figure D: RTF and SOI File Comparison: Percentage Differences and Coefficients of Variation

Item	Coefficients of variation (percentages)	Difference RTF to SOI (percentages)
Total assets (fair market value)....................	0.66	4.83
Total revenue...	1.50	0.65
Total expenses..	2.84	2.19

corrections were made to many data items common to the two datasets, nine major fields appeared to be corrected by SOI editors most frequently.

The three balance sheet items that represented securities--corporate stock, corporate bonds, and Government obligations--were corrected most often and, based on the median values of these corrections, with the most magnitude. Figure E shows RTF fields to which SOI editors commonly made corrections. In most cases, these corrections probably resulted from procedural differences in data entry, rather than operator error. SOI data entry operators collect information from supplemental attachments and schedules, in addition to the data that appear

on the Form 990-PF, to enhance the quality and accuracy of the microdata. The maximum and minimum correction values exemplify the effects of large keying errors on the RTF. Weights associated with the returns identified as corrected were applied to estimate the effects of SOI data entry on the overall population of private foundations. The categories of stocks, bonds, and Government obligations remained the most-often corrected financial items, after the weights were applied. The category "undistributed income," a field that represents required charitable distributions that foundations did not make in Reporting Year 2000, represents a larger portion of the total weighted corrections made than in the unweighted total. This indicates that more changes to the field were

Figure E: Unweighted and Weighted Corrections, Amounts and Values

Data Item	Minimum value	Maximum value	Unweighted Corrections			Weighted Corrections		
			Number	Median value	Mean value	Number	Median value	Mean value
Corporate stock	-5,241,441,621	118,170,705	1,640	-986,193	-16,639,345	10,725	-70,299	-2,763,242
Corporate bonds	-186,930,409	441,778,508	657	-984,966	-3,264,243	3,974	-123,775	-715,434
Government obligations	-344,684,265	454,418,685	538	-425,973	-1,615,332	3,354	-58,743	-270,657
Total assets, book value	-19,021,602,054	2,276,122,860	367	-21,264	-54,123,269	2,311	-5,159	-8,614,235
Total assets, fair market value	-34,824,317	397,295,763	297	-30,001	1,015,237	2,701	-19,247	-1,101,468
Total expenses	-70,188,315	28,400,237	240	-13,103	-506,701	1,567	-2,679	-81,322
Total revenue	-70,188,315	117,315	237	-15,127	-941,870	1,241	-986	-185,424
Undistributed income	-28,751,786	11,363,248	222	3,664	134,594	4,009	-438	11,105
Other revenue	-70,188,315	291,274	206	-15,033	-593,730	1,106	-1,009	-114,774

Figure F: Percentage of Returns with at Least One Correction, by Asset-Size Category
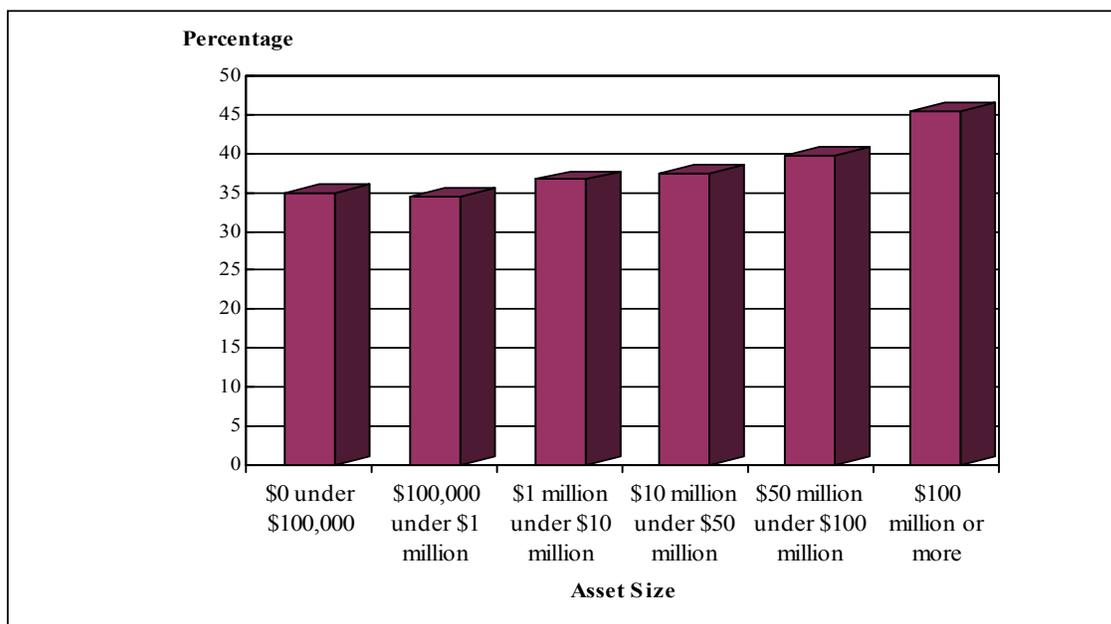
made to smaller, and therefore more heavily weighted, asset-size class returns. The relationship between foundation size and number of corrections was examined by arranging returns included in the microdata analysis into commonly used used "asset-size" categories. Figure F shows the percentage of returns with at least one correction to one of the nine data items examined, by asset size category. The proportion of corrections, generally, increased slightly with foundation size. More than 45 percent of the returns filed by the largest organizations, those with assets of $100 million or more, had a least one correction, indicating that the largest organizations are proportionally more often corrected than are their smaller counterparts. Overall, for the nine selected items, nearly 40 percent of the returns in the SOI sample had data inconsistent with those appearing on the RTF.

▶ Conclusions and Future Research

Based on this research, several important conclusions regarding data consistency, compatibility, and collection can be reached. In the past, SOI has been hesitant to supplement information unavailable on the SOI file with similar data from the RTF. However, it appears that these data can be used as complements, as long as the RTF data files are properly restricted to be consistent with the SOI file. While the SOI dataset is the only source for many data fields, in the future, the RTF may provide a valuable source for obscure, but sometimes necessary, data items. An important conclusion regarding data collection can also be reached based on this research. Currently, only a handful of items, none of which is financial, are incorporated directly from the RTF to the SOI transcription process. In many cases, however, some items that are available on the RTF 990-PF file remain largely unchanged during the SOI editing process. In the future, SOI may wish to build on this information and identify items that can be captured directly from the RTF to reduce the redundancy of operator transcription. SOI resources could then be directed toward transcribing additional data items, which may

not currently be available from any source.

Several future research options are available that could also help to illuminate data quality and collection issues. Currently, a sample of large-case returns that are included on both the RTF and SOI files is being transcribed based on information that appears directly on the Internet-posted, publicly available return. The data are being collected without additional information from attachments or schedules being transcribed. The information will provide insight into an avenue that researchers commonly use for information--the Internet--and will determine if the data posted by these organizations are consistent with those collected by IRS. Another valuable venture would involve comparing data from the SOI and RTF files for a number of years to ensure that that RTF data quality does not fluctuate between calendar years. This information could assist in determining definitive sources for specific data items. Ultimately, the results of this research may assist in improving resource allocation in the collection and dissemination of private foundation data.

▶ Notes and References

[1]   For example, a return that had an accounting period that began in January 2000 and ended in December 2000 was filed for Reporting Year 2000. This return would have likely been posted to the RTF in Calendar Year 2001, as the required filing data is 5 and ½ months after the end of the accounting period.

[2]   In some cases, a return that was filed late or by a taxpayer who received numerous extensions to file could have been received by IRS outside of the traditional, 2-calendar-year window.

[3]   The realized sampling rates for the Reporting Year 2000 SOI study of private foundations are shown below:

Fair Market Value of Total Assets	Realized Sampling Rate (percentage)
Private Foundations	
Under $125,000	0.3
$125,000 under $400,000	0.8
$400,000 under $1,000,000	1.9
$1,000,000 under $2,500,000	4.3
$2,500,000 under $10,000,000	21.0
$10,000,000 under $25,000,000	100.0
$25,000,000 or more	100.0
Charitable Trusts	
Under $100,000	1.2
$100,000 under $1,000,000	13.4
$1,000,000 or more	100.0

[4] Private foundations and charitable trusts were identified on the RTF based on their respective subsection codes. Private foundations are assigned a subsection code of "03," while nonexempt charitable trusts are assigned a subsection code of "92." Generally, organizations were also coded for the SOI File based on their subsection codes. However, in cases where subsection codes appeared to be incorrect or were not available, SOI staff conducted additional research to determine the proper subsection code for organizations on the SOI file.

The Effect of Content Errors on Bias and Nonsampling Variance in Estimates Derived From Samples of Administrative Records

Barry W. Johnson and Darien B. Jacobson, Internal Revenue Service

T he Statistics of Income Division (SOI) of the Internal Revenue Service (IRS) uses a number of methods for ensuring the quality and integrity of the data it produces for tax administration research. As a first line of quality assurance, codes and mathematically related data items are extensively tested as SOI employees enter them into computer databases. In addition, for a subsample of returns selected and processed in most studies, SOI assigns a second employee to reenter and edit the data. Values from the first and second edit are then computer-matched. A supervisor resolves discrepancies discovered during the match. The original value, second value, and correct values are all collected as a part of the quality review system, as are a set of codes that describe the cause of the error, in broad categories.

This paper will use quality review data from Federal estate tax returns (Form 706) selected into the Calendar Year 2002 SOI Estate Tax Study to estimate the effects of nonsampling error on estimates derived from the final data file.

▶ Background

The Federal estate tax is levied on estates for the right to transfer assets from a decedent's estate to its beneficiaries; it is not an inheritance tax. A Federal estate tax return must be filed for every U.S. decedent whose gross estate, valued on the date of death, combined with certain lifetime gifts made by the decedent, equals or exceeds the filing threshold applicable for the decedent's year of death. A decedent's estate must file a return within 9 months of a decedent's death, but a 6-month extension is usually granted.

All of a decedent's assets, as well as the decedent's share of jointly owned and community property assets, are included in the gross estate for tax purposes and reported on Form 706. Also reported are most life insurance proceeds, property over which the decedent possessed a general power of appointment, and certain transfers made during life.

Expenses and losses incurred in the administration of the estate, funeral costs, and the decedent's debts are allowed as deductions against the estate for the purpose of calculating the tax liability. A deduction is allowed for the full value of bequests to the surviving spouse. Bequests to qualified charities are also fully deductible.

▶ Data Description

The 2002 SOI Estate Tax Study was a stratified, random sample of returns filed in Calendar Year 2002 and was the second year in a 3-year study of Federal estate tax returns filed 2001-2003. The sample was designed for use in both estimating tax revenues in all 3 calendar years and personal wealth holdings for 2001 decedents. The 3-year sample period was devised to ensure that nearly all returns filed for 2001 decedents would be subjected to sampling, since a return could be filed up to 15 months after the decedent's death. The design had three stratification variables: size of total gross estate plus the value of most taxable gifts made during the decedent's life, age at death, and year of death. The year-of-death variable was separated into two categories, 2001 year of death and non-2001 year of death, in order to facilitate studies of 2001 decedents. Returns were chosen before audit examination and selected using a stratified random probability sampling method. A portion of the sample was selected because the ending digits of the decedents' Social Security Numbers (SSN) corresponded with those in the 1-percent Social Security Administration Continuous Work History Sample. However, the majority of returns were selected on a flow basis using the Bernoulli sampling method.

The sampling mechanism was a permanent random number based on an encryption of the decedent's SSN. Sample rates were preset based on the desired sample size and an estimate of the population. Sampling rates ranged from 3 percent to 100 percent, with more than half of the strata selected with certainty.

Data collection for the 2002 Estate Tax Study was

conducted at the IRS Cincinnati Submission Processing Center. Employees entered the data from the estate tax return into a database using a Graphical User Interface (GUI) data entry system. Nearly 100 distinct data items were captured, with some balance sheet assets recurring hundreds, even thousands, of times, as assets were allocated to 32 different categories, such as stocks, bonds, and real estate. Tax returns ranged in size from a dozen to many thousands of pages, including appraisals, investment account listings, and legal documents. Tests embedded in the data entry system were used to validate entries and to ensure that mathematical relationships among variables were correctly preserved. There were more than 200 validation tests performed on each tax return included in the 2002 study.

While embedded testing can assure that codes are correct within a given range of values and that fields are mathematically consistent, many of the decisions that employees make when transforming tax return information into statistically usable data are not easily tested. For example, while several codes may be valid, determining the best code to describe a particular taxpayer's behavior or characteristics cannot always be automated. To address this problem, SOI developed a double entry quality review system. This system is a valuable tool for measuring both individual employee performance and overall data quality.

▶ Quality Review System

A subsample of returns in the 2002 Estate Tax Study was subjected to additional review for quality assurance purposes. Returns were included in the quality review (QR) subsample through two different mechanisms, 100-percent review and product review. The 100-percent review consisted of all returns that were edited while an employee was in training. Product review was selected after the training period had been completed, and it comprised a 10-percent random sample of each employee's work. The product review sample was selected on a flow basis method using a pseudorandom number called the Transform Taxpayer Identification Number, or TTIN. The TTIN is a unique random number that is generated by mathematically transforming selected digits of the decedent's Social Security Number. The TTIN was then compared to the sample number, which represented the

sample rate, in this case, 10 percent. If the TTIN was less than the sample number, then the return was selected for product review.

Under the double-entry quality review system, one return was entered into the computer system twice by two different employees. The first employee did not know that a return was selected for review until after the first edit was complete, and the second employee was not allowed to see the first employee's entries. Therefore, each return had two versions in the database, the first edit and the second edit, and each was entered independently of the other.

When both employees finished editing a return, the computer compared the values from the original and QR versions. In some cases, the two versions matched perfectly; so, the return was released from the system, and the first edit data was treated as final and stored for later analysis. However, if mismatches between the two versions occurred, the discrepancies were stored in a separate data table to be reviewed by a supervisor.

The supervisor reviewed the discrepancies and charged the errors, assigning two codes to each discrepancy--one to identify the incorrect value and the other to describe the cause of the error. A discrepancy code was assigned to the error to explain which version was considered incorrect. Discrepancy codes were assigned to one of the following: the first version, the second version, both versions, or neither version. An error was assigned to both versions if both of the employees entered or interpreted the information from the return incorrectly. In this case, the supervisor was also required to supply the correct data value. In some cases an error was not assigned to either version, usually when the discrepancy was the result of a data processing peculiarity and not a true database error. After the error was assigned a discrepancy code, a numeric error resolution code was assigned to describe why the entry was incorrect. Error resolution codes indicate situations such as spelling errors, incorrect money amounts, or incorrectly assigned codes.

Once the supervisor reviewed all the discrepancies, each employee was given a list of the discrepancies, along with the discrepancy and error resolution codes,

so that any first edit errors detected during quality review could be corrected prior to considering return processing complete. The feedback from the review also enabled employees to learn from their mistakes on each return and carry this knowledge into the editing of other returns. In the end, there is a database consisting of a table that includes all the values from the second edit of the return as entered, a quality review table containing a record of each discrepancy between the first and second edits (along with codes indicating who made the error and why), and a final data table containing the correct version of the return data that will ultimately be sent to customers.

For this paper, only a portion of the quality review data was used for analysis. First, data that were collected during periods of training, 100 percent review, were excluded. Second, only errors that were charged to the first edit or to both edits, meaning that the error required a correction to the final data set, were retained. This was done because these errors are more representative of errors that remain in the roughly 90 percent of the 2002 estate tax sample that was not selected for quality review. Third, errors that reflected idiosyncrasies related to the edit process itself, and not true data errors, were eliminated.

▶ Empirical Results

Quarterly accuracy rates for each employee who worked on the Estate Tax Study for 2002 were generated using the product review data (see Figure 1). These rates were calculated using the number of returns that had at least one error charged to the first edit divided by the total number of returns that had been selected for quality review. The accuracy rates for all of the employees are not very high. However, these rates are a return level measure; any return with one or more errors is considered incorrect. The Form 706 includes an average of 150 data entry fields, while complex returns can have more than a thousand entries; so, the probability of making just one mistake is very high. In fact, the average number of errors for each return is only 6.3.

Traditionally, supervisors have focused quality improvement efforts on those fields that are in error most frequently. By looking at the occurrence of variables *ex-*

Figure 1: Employee Accuracy Rates

| Employee	Quarter 1	Quarter 2	Quarter 3	Quarter 4
17000	46.3%	23.9%	41.7%	21.7%
17100	25.0%	0.0%	0.0%	0.0%
17200	29.2%	30.8%	31.9%	40.0%
17300	57.1%	100.0%	91.7%	33.3%
17400	52.1%	28.6%	50.0%	37.9%
17500	44.4%	24.1%	54.8%	0.0%
17600	42.2%	51.9%	33.9%	46.2%
17700	41.9%	28.6%	39.3%	34.5%
17800	49.1%	25.0%	58.5%	45.6%
17900	52.3%	34.3%	59.0%	50.0%
17001	23.1%	34.2%	18.6%	44.7%
17002	39.2%	33.3%	36.2%	45.0%
17003	22.9%	20.7%	37.8%	29.1%
17004	34.2%	31.6%	22.0%	72.7%
17005	30.8%	0.0%	0.0%	37.9%
17006	26.5%	27.7%	41.4%	42.9%

ante, using the first edit data, and *ex-post*, using the final corrected data file, it is possible to identify the frequency of original edit errors in the quality review sample. Figure 2 shows the percent changes in frequencies for variables on the file; each diamond represents a different variable. Frequencies change because many variables on the file represent balance sheet items, assets like stocks, bonds, mutual funds, and various types of real estate, which are not necessarily present in each decedent's portfolio. When an asset is incorrectly classified, not only does it change the dollar value of estimate, it also changes the frequency of occurrence of that particular attribute or asset type in the population estimates. This can be particularly problematic if the asset is of special interest to researchers. For example, there has been much discussion in the press about providing estate tax relief to small business owners. Errors that either under- or overcount the number of estates that have small

Figure 2: Percent Change in Frequencies, Original and Final Edits
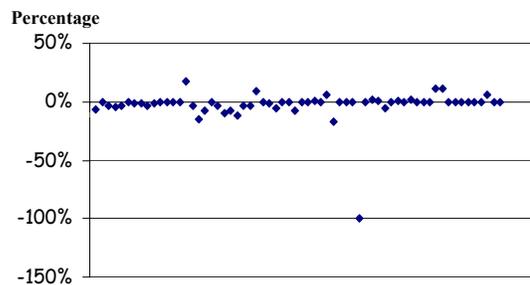
businesses could have an impact on this debate. The percentages shown on the graph represent the aggregate correct frequency in the overall quality review sample, less the aggregate number originally reported, divided by the correct number. Negative percentages indicate cases where an asset was incorrectly included on the first edit. For example, the first employee may have incorrectly classified a balance sheet entry as a publicly traded stock, while the second employee may have correctly classified it as a mutual fund invested in a mix of financial assets. The percent changes in frequencies are generally close to zero, but there are some notable outliers.

Figure 3 shows percentage changes in dollar amounts between first and second edits overlaid on the frequency differences shown in Figure 2. Each point represents a single variable on the file. While the pattern for the dollar differences is similar to that of the frequencies, with many differences close to zero, the magnitude of the dollar differences is larger for several variables. There are two variables for which the original entries resulted in aggregate dollar values that were overstated by roughly 150 percent. This highlights the potentially large effects on final estimates that can arise from even one large dollar value error, especially for variables that are not widely distributed in the overall population. Thus, it is important to monitor both the size and frequency of data entry errors.

Figure 3: Percent Change in Dollar and Frequency Values, Original and Final Edits
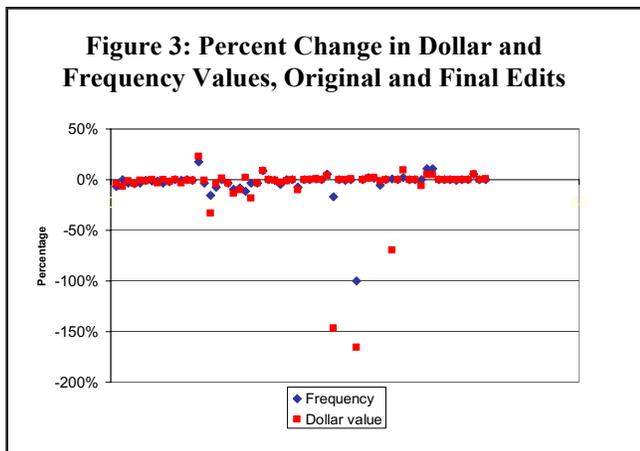
Unweighted error statistics are clearly useful for monitoring data quality and assessing opportunities for operational improvements during a study period. However, since the SOI study of Federal estate tax returns is based on a stratified random sample of the filing popu-

lation, the effect of data entry error on final population estimates derived from this sample will vary inversely with the selection rate associated with each return. Using appropriate sample weights, it is possible to use the 10-percent QR sample to estimate the effects of data entry errors on population estimates derived from the remaining 90 percent of the returns in the final SOI data file that were not subjected to double-entry quality review. Weighted estimates provide a different perspective on the effects of nonsampling error due to the nature of the underlying estate study sample and the fact that the financial characteristics of estate tax decedents vary greatly among age and wealth classes. For example, younger decedents and those with large estates are selected into the estate tax sample with certainty and comprise more than 40 percent of the total sample file. Both groups of decedents are more likely to have had portfolios that are more complex and, thus, more subject to data entry errors than their either less wealthy, or older, cohorts. This is because many older wealth holders convert their portfolios to assets that produce tax-preferred income, usually resulting in returns that contain fewer business arrangements, which are more difficult to classify than market assets. Because the quality review sample is not stratified, weighted estimates will provide a more balanced measure of the overall effects of data entry errors on final estimates. Weighted estimates for the quality review sample were generated by using the design-based weight from the stratified estate study sample (W_s), multiplied by a quality review weight (W_q). The quality review weight itself was developed by first post-stratifying the quality review samples within the original selection strata as indicated below [1]:

$$\text{Final Weight} = \quad W_s*W_q$$
$$\text{Where } W_s = N_i/n_i$$

Post-Stratification: $W_q = n_{if}/n_{qif}$

For some strata, the quality review sample was either zero or too small to create a post-strata cell. For these cases, strata were collapsed across age categories so that estate size classes were preserved.

Figure 4 shows full population dollar value estimates from the quality review data using the post-stratified quality review weight and compares them to population

estimates using the full weighted estate study sample. Each pair of data points represents a different variable on the file. The quality review data estimates for each variable are denoted by the gray squares, and the full sample estimates are denoted by the black diamonds. For most variables, the QR sample estimates are larger than the population estimates from the full estate sample, indicating that the QR sample introduces a positive bias. This bias arises because the QR sample is a simple random sample of a stratified sample that favors large dollar value returns. In such cases, ratio raking can often be employed to decrease the bias; however, in this case, the QR sample size was insufficient in the lower gross estate size classes.

Figure 4: Full Sample vs. QR Sample Estimates
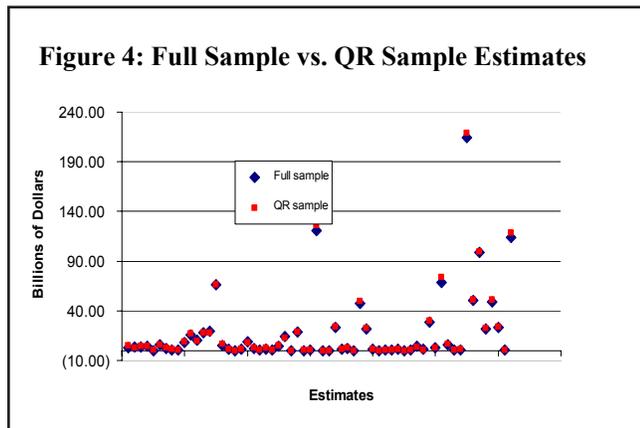
While the weighted QR data estimates are somewhat biased due to the design of the sample, they still provide an important indication of the effects of data entry errors on final estate tax sample estimates. Figure 5 shows weighted and unweighted estimates of aggregate differences between original and final values of both frequency and dollar value estimates for selected variables. A negative value means that a variable was overrepresented in the original, uncorrected data, and a positive value means it was originally underrepresented. Weighted results rank errors differently for some of the variables. For example, errors in classifying noncorporate business assets had a much greater impact on final weighted estimates than would have been evident had the analysis been limited to examining the unweighted QR data. Conversely, the unweighted QR data implied that the effects of errors on estimates of farm real estate were greater than they are in the final, weighted estimates. Clearly, using weighted estimates, along with

the unweighted quality review data, provides a more balanced method of assessing where to focus data quality improvement efforts.

Figure 5: Differences Between First and Final Edit

Data Element	Frequency	Dollar Value
Non-corporate businesses	-11.00%	-5.79%
	-5.29%	*-3.55%*
Publicly traded stock	15.02%	20.00%
	15.38%	*23.40%*
Closely held stock	-3.06%	-1.01%
	-3.42%	*-.71%*
Real estate	6.70%	7.34%
	6.82%	*6.17%*
Farm land	-.91%	-1.09%
	-1.95%	*-3.66%*
Funeral expenses	.25%	.15%
	.09%	*.04%*

Values in *italics* are unweighted

Figure 6 compares the weighted percent differences between original edit estimates and final, corrected estimates with coefficients of variation (C.V.) from the full estate tax study sample in order to relate the sampling and nonsampling variances associated with selected fields. For some estimates, such as the values for noncorporate businesses and publicly traded corporations, the nonsampling error attributable to data entry is much greater than the sampling variance. For others, such as estimates of stock in closely held or untraded corporations and farm land, the sampling error, represented by the C.V., is actually greater than the nonsampling error attributable to data entry errors, indicating that data entry errors are not a significant cause of additional variance in the estimates. Fields for which nonsampling error

Figure 6: Data Entry Error vs. Sample Variance

Data Element	Frequency		Money Amount	
	% diff	C.V.	% diff	C.V.
Non-corporate businesses	-11.00%	4.45%	-5.79%	3.89%
Publicly traded stock	15.02%	.78%	20.00%	1.17%
Closely held stock	-3.06%	3.47%	-1.01%	2.18%
Real estate	6.70%	1.92%	7.34%	2.19%
Farm land	-.91%	4.34%	-1.09%	4.68%
Funeral expenses	.25%	.57%	.15%	1.19%
Spousal trusts	4.25%	2.97%	1.29%	1.58%

is relatively large provide opportunities for future data quality improvement efforts.

▶ Conclusion

There is much to be learned through careful analysis of the data generated by SOI's double-entry quality review systems. The results of these analyses can be used to improve data collection systems and enhance worker training. Information on nonsampling error should also be useful to data users who could use data quality metrics to more accurately interpret economic modeling results and to ultimately build models that are more robust.

This analysis, however, revealed that the database format and the type of data that are collected from the quality review samples make certain types of analysis difficult, if not impossible. While a complete copy of the second edit is saved for all QR returns, the original, uncorrected first edit values are not saved when first edit errors require corrections. Information on discrepancies is kept in all cases, but, because corrections can involve changing any number of related fields, it is difficult to reconstruct exactly the first employee's original entries. If more sophisticated analysis is desired, including the study of secondary errors that arise as a result of a primary data entry error, archiving a complete copy of the first edit, along with associated error reason and discrepancy codes, should be considered.

It is also important that supervisors apply error reason and discrepancy codes consistently. All too often, discrepancies are resolved by several different supervisors. Some, especially those serving in a temporary capacity, may feel a great deal of peer pressure to avoid assigning errors to individual employees, even in cases where the assignment of an error would not directly impact employee performance appraisals, such as when an error is attributable to lack of clarity in editing instructions. This inconsistency makes it difficult to measure the extent to which errors exist and to learn of ways to avoid them in the future.

Related to this problem is that the measure of employee performance currently in place is not adequate. It is simply unfair to use a return level measure of accuracy when the difficulty of the work is so variable across returns. A more balanced measure would relate the number of individual errors an employee makes to the number of fields he or she actually edited, thus giving full consideration to the number of edit decisions that were made on each return.

Finally, there are sample design issues that became apparent from this analysis. The QR sample is biased and could be improved by taking into consideration the underlying structure of the estate tax study sample design. Even this would not provide coverage of variables that are relatively rare, but perhaps important, in policy debates. To address this problem, samples could either be increased or targeted to include more returns with important characteristics, such as those filed for small business owners, or returns that, because of the types of entries made during first edit, are more likely to contain significant problems. Samples could also vary with worker skill levels. One possibility would be to develop a system that sets a weekly QR sample rate for each individual employee based on individual rolling average accuracy rates. Sample rates could be set automatically based on preset performance standards. Automating the process would avoid putting supervisors in the awkward position of having to 'punish' poor performers with additional oversight, making it easier to match feedback and training efforts to performance levels.

▶ Footnote

[1] The subscript "if" signifies that certain reject returns were removed from the estate study sample prior to post-stratifying.

Editor Judgment Effect: Modeling a Key Component of Nonsampling Error in Administrative Data

Kimberly Henry, Yahia Ahmed, and Ellen Legel, Internal Revenue Service

This paper is a modest attempt to model a key component of nonsampling error in administrative data, particularly tax data. Tax data items present obstacles for statistical uses that are far outweighed by the fact that responses on tax returns are likely to be more accurate than financial-related responses to general surveys. These obstacles lead to a kind of nonsampling error that we refer to as editor judgment error. The Statistics of Income (SOI) Division of the IRS developed a processing procedure called statistical editing to abstract tax return data for statistical purposes. Statistical editing helps overcome limitations inherent in tax return statistics and achieves certain statistical definitions desired by data users. Statistical editing involves adjusting certain taxpayer entries based on supplemental information reported elsewhere on the tax return (such as attached schedules that support a reported total). It is a major factor in producing SOI's corporation income tax return statistics.

In the next section, we describe the SOI corporate sample design, identify sources of nonsampling error, and define the term "editor judgment error." We then describe current SOI editing and quality review processes; outline the purpose of our study and its limitations; discuss bias and variance component models, which were adapted from simple response error measurement models; and summarize results and conclusions.

▶ Sample Design Description and Nonsampling Error Sources

The data for this study are from the 2001 SOI Corporate sample, which consisted of corporations that filed income tax returns with accounting periods ending between July 1, 2001, and June 30, 2002. The realized 2001 sample contained 147,093 returns (including inactive corporations and noneligible returns) selected from a population of 5,563,663. The sample is a stratified random sample, where stratification is based on 1120 form type. Within form type, further stratification is achieved by use of either size of assets alone, or both size

of assets and a measure of income. A Bernoulli sample is selected independently from each stratum, with rates ranging from 0.25 percent to 100 percent. The sample is selected weekly as the Form 1120 returns are posted to the IRS Business Master File. It takes 2 years to select the sample due to the combination of noncalendar year filing and the 6-month extension options.

Sampling errors arise from using a sample instead of a census, and SOI publishes them in the form of Coefficients of Variation (IRS, 2001, pp. 29-36). Nonsampling errors include all others, such as coverage, nonresponse, measurement, and processing errors.

Coverage errors, when a unit is not available on the sampling frame, can occur if a corporation files an extension. Imputation procedures using adjusted prior-year data are used to correct for coverage errors in large companies.

Missing data, or nonresponse errors, occur when other IRS functions have returns selected for the sample, rendering them unavailable for SOI processing. Imputation procedures and weighting adjustments are used to adjust for missing large and small companies, respectively. Noncoverage imputation and missing returns represented 0.03 percent and 0.22 percent of the 2001 sample, respectively (IRS, 2001, pp. 7-14).

Measurement errors occur when a taxpayer enters an incorrect value, for various reasons. SOI does not sample amended returns or contact taxpayers.

Finally, processing errors occur while abstracting, transcribing, and cleaning the data. Since the editors abstract administrative data from tax returns and enter them into SOI database systems for statistical purposes, editor judgment error falls into this nonsampling error category. However, it is more than transcription error because certain judgments are required from the editors due to a combination of transcribing data collected for tax liability, which is subject to different corporate accounting practices, and study standards created for

statistical purposes.

▶ Current SOI Editing and Quality Review Processes

Fifty-nine editors at two IRS Service Centers abstracted approximately 1,400 corporate tax return items for the 2001 sample. This data abstraction process was complicated due to many factors, for example:

□ The extracted items from any given return often require totals to be constructed from various other items on other parts of the return.

□ There are currently ten form types, with different layouts, schedules, and attachments; so, data extraction is not uniform across form type.

□ There is no legal requirement that a corporation meet its tax return filing requirements by filling out, line by line, the entire U.S. tax return form. Some returns are also exempt from filling out entire sections; for example, currently, Form 1120 returns with total assets and total receipts below $250,000 do not have to report their balance sheet items.

□ There is no standardized accepted method of corporate accounting used throughout the country. For example, different companies may report the same data item, (such as deposits, a subset of other current liabilities), on different lines of the tax form.

Despite complexities such as those listed above, study standards place SOI's editors in a position to make judgments during data abstraction. Errors in these judgments are the largest source of editor error in the corporate sample.

To assist the editors, SOI's National Office (NO) staff in Washington, DC, implement many procedures that attempt to make the editing process consistent with the 1120 study standards and reduce editor effect. This is similar to the concept of standardized interviewing used in other survey organizations. For example:

□ Detailed editing instructions are prepared every year--the 2001 manual contained more than 900 pages.

□ Over 700 computerized tests are performed on abstracted data to ensure certain accounting conditions are satisfied, such as balanced totals or absence of consistent amounts between front-page items and attached schedules. All tests are reviewed and tested by NO staff the year prior to data abstraction in a process called Systems Acceptability Testing.

□ The staff build utilities into the edit computer system that offers industry-specific suggestions, guidelines, and requirements for particular sections of the form.

□ They review and monitor the sample throughout the program year for unusual accounting conditions and codes. During the last 4 months, the largest corporations within each industry are reviewed as well as the largest industry differences across asset classes.

□ The NO staff conduct extensive edit training and review all items on all returns edited during certain periods of the program year to overcome inexperience due to new tax laws, edit instructions, codes, or even an entirely new program. For example, editors improving throughout the year are given more complicated returns, the first of which were completely reviewed with their supervisors.

While complete review was an excellent training tool, the editors knew in advance which returns were going to be reviewed. For the purposes of our study, the returns may have been biased; so, they were omitted from analysis.

During data editing, approximately fifty returns were randomly selected for each editor for quality review. Once an editor's return was selected for review, another editor on the same team independently re-edited it. After the returns were compared item by item and discrepancies were stored in SOI databases, the

editors' supervisor determined the correct value (either the first editor's value, the second's, both, or neither). Any amounts that differed by less than $10, along with character, display, and generated item mismatches were omitted from quality review. We used only the first editor values because they are the final file values and the second editor knew which returns were for review. Assuming that a taxpayer is correct, the errors described in Table 1 are used to determine service center accuracy ratings, and we included all of them:

Table 1: Types of Errors

Type of Error	Description
Amount	An incorrect amount was entered in an item.
Omitted Entry	A zero or blank item that should have a code/amount present.
Extra entry	An item with a code/amount in it should have been blank or zero.
Entry on omitted form	An item was not edited because the form or schedule was not edited.
Improper allocation	An amount that should have been allocated to another item was not moved or was moved incorrectly.

Improper allocations were the most frequent errors; so, this type of error is illustrated in Table 2.

Table 2: Improper Allocation Example

Item	Edited Amount	Correct Amount	Error
A	1,000.00	0.00	1,000.00
B	0.00	1,000.00	-1,000.00
C	2,000.00	2,000.00	0.00
Total	3,000.00	3,000.00	0.00

Here, for three hypothetical items A, B, and C (which may not be located on the same page, form, or attachment), both totals match; the system will not catch the error despite errors in two of three items. An important aspect of improper allocation errors is that they often result in net error effects of zero: here, errors in items A and B cancel each other out. This is important when calculating national-level estimates for totals, but a concern for estimates of A or B.

▶ Study Purpose and Limitations

The quality review system was developed to check edit manuals, measure training effectiveness, and evaluate the editors. As previously mentioned, approximately fifty returns were randomly selected for each of the fifty-nine editors for quality review. Given this pre-existing quality review system, our goal was to develop quality performance statistics and quantify the editor effect.

Table 3: Errors and Error Rates, Quality Review Study vs. Our Study

Item	QR Study	Our Study
# returns	3,080	373
# errors	9,229	760
# errors possible	33,880	4,103
error rate	.272	.185

As shown in Table 3, data used for our study were a subsample of 373 returns from the 3,080 quality review returns. All 3,080 returns were not included because returns with assets more than $250 million were only edited by a group of the most experienced editors, then reviewed by NO staff. In order to compare across all form types, service centers, teams within service, and editors within teams, we selected this subsample, which consists of all Form 1120 and Form 1120 Regulated Investment Company returns with total assets less than $250 million. Most importantly, all editors edit these returns during the program year, regardless of their experience. There were 73,115 of these returns in the corporate sample, for which NO staff relied on the editors' judgments for most of them because they were reviewed only under special circumstances. Our subsample is small compared to the SOI sample (about 0.51 percent); so, the results from this relatively small sample were analyzed assuming the observations were from independent, identically distributed random variables and sample weights were not used (Brick et al., 1996).

We selected eleven variables from the balance sheet and income statement sections of the returns in our study that were of interest to our subject-matter specialist; it is obvious from their names that many are ambiguous. Table 4 displays the number of errors and error rates for the eleven selected variables.

Table 4: Number of Errors and Error Rate, by Item

Item	# Errors	Error Rate
Gross Receipts	58	0.014
Other Assets	68	0.017
Other Costs	72	0.018
Other Current Assets	57	0.014
Other Current Liabilities	58	0.014
Other Deductions	110	0.027
Other Income	81	0.020
Other Investments	76	0.019
Total Deductions	62	0.015
Total Income	63	0.015
Trade Notes/Accounts Receivable	55	0.013

Error rate is equal to number of errors out of the 4,103 errors possible. Other Deductions has the highest error rate of 2.7 percent because Deduction item editing tasks are more complicated due to complex and varying accounting rules.

▶ Bias Estimation and Variance Decomposition

Measurement error modeling was first proposed by Hansen et al. (1952) and Seth and Sukhatme (1952). Their model specified that a single observation y_i from a randomly selected respondent i is the sum of two terms: a true value, μ_i, and an error term, ε_i. Mathematically, this is written as

$$y_i = \mu_i + \varepsilon_i \qquad (1)$$

While we did not measure response error, we adopted these models to our data to measure editor judgment error. In model (1), μ_i, the true value, is a random variable whose distribution depends on the sample design. The distribution of the editor error variable ε_i

is conceptual; it could be viewed as sampling from a hypothetical population of errors. Thus, the assumptions for model (1) are

$$E[\varepsilon_i | i] = B_i \neq 0$$
$$Var[\varepsilon_i | i] = \sigma_i^2$$
$$E[\sigma_i^2] = \sigma^2$$
$$Cov[\varepsilon_i \varepsilon_j] = 0, \; i \neq j$$

In words, a systematic bias exists because the mean of the errors is not zero and the variances are not equal. Also, errors are uncorrelated: the errors for a first or second edited return do not affect other returns in the same edit period, and errors across edit periods for the same return are uncorrelated.

Assuming unrestricted simple random sampling,

$$E[\mu_i] = \overline{\mu}$$
$$V[\mu_i] = \sigma_\mu^2$$
$$Cov[\mu_i, \mu_j] = 0, \; i \neq j$$

In our study, the observed value is the first editor's value in the sample, while the true value is either the first or second editor's value (whichever was determined to be correct by the supervisor), and i denotes unit. It deserves mention that model (1) has potential weaknesses, particularly if the first and second editor's values are correlated, but it can provide a useful approximation for the editor's contribution of error. The model also allows for calculating statistics to measure editor accuracy further than number of errors out of number of errors possible.

Under model (1), we assume that the first editor's error term no longer averages to zero, possibly due to editor bias, defined as

$$B = \sum_{i=1}^{N} (y_i - \mu_i) \qquad (2)$$

The bias can be estimated by the *Net Difference Rate* (NDR), which is given by

$$NDR = \bar{y} - \bar{\mu} \qquad (3)$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$, $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i$, and n is the sample size.

It can be shown that if μ_i is the true value, then the expected value of the NDR is the bias, and its variance exists (Biemer and Atkinson, 1992). Table 5 shows the estimated NDR and *student's t statistcs* for the eleven items. Negative bias values should be interpreted as editors underestimating variables, and positive NDR estimates indicate overestimates.

Table 5: Net Difference Rate, by Item

Item	NDR	*t*
Gross Receipts	-749,441	0.16809
Other Assets	293,125	0.23662
Other Costs	7,847	0.00683
Other Current Assets	361,062	0.19090
Other Current Liabilities	1,989,871	0.26820
Other Deductions	-958,930	0.26017
Other Income	-662,720	0.27392
Other Investments	-59,372	0.03116
Total Deductions	543,972	0.21601
Total Income	500,441	0.16296
Trade Notes	32,635	0.01395

At first, the NDR estimates look very large in both directions. Since most errors are improper allocations, an entire amount is determined to be in error. Since the *t statistics* are all less than 1.96, we can conclude that editor judgment error appears to be a random error, not a systematic error as first assumed. We can assume that $E\left[\varepsilon_i | i\right] = \beta_i = 0$, i.e., the editor error averages to zero because it is a random error.

Since simple random sampling is assumed and the bias is zero, it can be shown that the variance of a mean over all possible editing review samples and all possible editing trials can be decomposed into

$$Var[\bar{y}] = Var[\bar{\mu}] + \frac{\sigma^2}{n}$$
$$= SV + EV \qquad (4)$$

The sampling variance, SV, is the ordinary variance with no editor error. The editor variance, EV, is the variability of returns averaged over conceptual repetitions of the editing under the same conditions.

Hansen et al. (1964) define the Index of Inconsistency (IOI) as

$$IOI = \frac{EV}{SV + EV} \qquad (5)$$

which we use to estimate the proportion of random errors associated with editor judgment error in total variance. Estimated IOI values are shown in Table 6.

Table 6: Index of Inconsistency, by Item

Item	IOI
Gross Receipts	0.0155
Other Assets	0.3084
Other Costs	0.0140
Other Current Assets	0.1526
Other Current Liabilities	0.1829
Other Deductions	0.2091
Other Income	0.1365
Other Investments	0.0464
Total Deductions	0.0247
Total Income	0.0336
Trade Notes	0.0370

Other Assets (0.3084) and Other Deductions (0.2091) are the items with the greatest proportion of editor judgment error. All other IOI estimates were less than 0.2, which is a small proportion compared to other surveys (Lessler and Kalsbeek, Chapter 11).

► Conclusions

To summarize, despite large NDR values in both directions due to editor judgment errors, particularly

improper allocations, the expected value of the bias for all items is zero. Further analysis of the NDR yielded different results by edit team. Internal examinations of NDR comparison graphs by team, item, and editor were useful in identifying strengths and areas of editing improvement that can be addressed through training. Third, the t statistics are also small, so editor judgment error for these returns is a variable error, not a systematic error. Variable errors tend to cancel each other out. Variance decomposition for our eleven items showed editor variance is a small component of total variance. Overall, our measures demonstrate high quality editing; so, reliance on their judgment is justified when every possible error scenario cannot be programmed, foreseen, or identified by National Office Staff.

This study is a first attempt, and a modest one, to quantify the effect of SOI's editors on data quality. Our encouraging results are a strong argument for the necessity of more research. We examined the simplest tax returns in order to compare the editors, returns whose errors have the smallest impact on overall quality of national estimates. The largest errors associated with the largest tax returns require a separate error measurement study because they are sampled with certainty and therefore do not contribute to sampling error. Further, the validity of taxpayer values, which are assumed to be correct when corporate returns reach SOI, is another area deserving examination.

► References

Biemer, P. and Atkinson, D., "Estimation of Measurement Bias Using a Model Prediction Approach," *1992 Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 64-73.

Biemer, P.P. and Fesco, R.S. (1995), "Evaluating and Controlling Measurement Error in Business Surveys," *Business Survey Methods,* John Wiley & Sons, New York, pp. 257-281.

Biemer, P. and Stokes, L. (1991), *Measurement Errors in Surveys,* John Wiley & Sons, Chapter 24, pp. 487-516.

Brick, M.; Kim, K.; Nolin, M.J.; and Collins, M. (1996), "Estimation of Response Bias in the NHES:95 Adult Education Survey," *Working Paper Series,* National Center for Education Statistics, Washington, DC.

Hansen, M.H.; Hurwitz, W.N.; and Madow, W.G. (1952), *Sample Survey Methods and Theory,* John Wiley & Sons, New York, Volume II.

Hansen, M.H.; Hurwitz, W.N; and Pritzker, L. (1964), "The estimation and interpenetration of gross differences and the simple response variance," *Contribution to statistics,* in C.R. Rao (editor) Pergamon Press, Oxford, and Statistical Publishing Society, Calcutta, pp. 111-136.

Internal Revenue Service, Statistics of Income--2001, *Corporation Income Tax Returns*, Washington, DC.

Lessler, J.T. and Kalsbeek, W.D. (1992), *Nonsampling Errors in Surveys*. John Wiley & Sons, New York

Sukhatme, P.V. and Seth, G.R. (1952), "Non-sampling errors in surveys," *Journal of Indian Society of Agricultural Statistics*, pp. 5-51.

The Evolution of IRS Telephone Quality Measures

Laura Rosage, Internal Revenue Service

The Internal Revenue Service (IRS), tasked with collecting taxes from this country's citizens, deals with more Americans than any other public institution. Unfortunately, over the years, the tax law has increased in complexity, and the myriad of forms has become confusing. In an effort to assist taxpayers comply with the law, IRS established a toll-free telephone service with Customer Service Representatives (CSRs) ready to help taxpayers with their tax-related questions. Since 1965, IRS has offered this free telephone assistance to millions of taxpayers. IRS assisters handled over 55 million telephone calls in Fiscal Year 2003 (October 2002 through September 2003) and 15.8 million calls during filing season 2004 (January through mid-April 2004) alone. With this large volume of inquiries handled by telephone assisters, the accuracy of the information provided has a potentially large impact and is of interest to stakeholders both within and outside of IRS. The way the accuracy of telephone assistance is measured has evolved over the years from test calls, to live monitoring of telephone calls, and soon to contact recording. This paper details the evolution of how IRS monitors calls and the latest move from a pass/fail method of measurement to a defects-per-opportunity methodology. It discusses the strengths and limitations of each method, the overall impact on quality rates, and future plans for improvement to the measurement process.

▶ Monitoring Taxpayer Calls for Quality Purposes

One of IRSs major goals is to make its telephone operations a world-class customer service organization. To do this, it is necessary to track the accuracy, efficiency, and quality of the service provided by those answering the telephones. However, when first implemented, the focus of telephone service review was only to evaluate employee performance. Managers reviewed the work of CSRs and used the data gathered as feedback for employees' performance appraisals and to identify training needs. It was not until the late 1980s that IRS attempted to measure the overall quality of the service provided to taxpayers.

Managerial Review

IRS's first iteration of quality measurement consisted of managers reviewing their employees at each call site. The manager would sit with an individual CSR and listen in on selected telephone calls. When the call was complete, the manager was able to provide immediate feedback to the CSR on any errors made or on issues of timeliness or professionalism. While better than no review at all, there were some drawbacks to this system. First, the data gathered during telephone call monitoring were really intended for CSR performance review rather than a site or national measurement of quality. Also, because managers monitored their own employees, complete impartiality of the reviews could not be guaranteed. Additionally, since managers sat with the CSRs, the assisters were aware they were being monitored. They could alter their behaviors during calls that were reviewed: Responding in a more professional manner, researching the tax issue more thoroughly, or adhering to their manual guidelines more fully. These changes could affect the quality of the call, giving IRS a skewed view of the performance of the employee, as well as the overall quality of the service provided to taxpayers. Finally, because review of the CSRs was performed at each of the call sites, there were issues with the lack of consistency of reviews from manager to manager and site to site.

Attempting to get a clearer picture of the actual service taxpayers received, IRS implemented technology that allowed managers to review CSRs remotely. Without alerting the CSR, they could listen in on and review telephone calls from the privacy of their offices, rather than sitting beside the assister as he or she was on the phone. This transparency eliminated the problem of the CSRs knowing they were being monitored and modifying their behaviors accordingly, but the issues of

inconsistency of reviews from site to site still existed. Additionally, the accuracy results from this process were generally very high and at odds with the Government Accountability Office's (GAO) assessment of quality, further supporting the concern that the monitoring carried out at a local level was not impartial.

Integrated Test Call Survey System

In an effort to eliminate these concerns, the Service implemented a new program in addition to the managerial review, the Integrated Test Call Survey System (ITCSS). This system, established in 1988, was designed to produce a national estimate of quality rather than relying on managerial review of employees to establish the measurement and to provide timely feedback to call sites. The sites could then use the feedback to target specific areas for improvement, then assess the success of their efforts. Under ITCSS, a centralized group of quality reviewers called into the toll-free IRS tax law assistance line, posed mock questions to CSRs, and rated the quality of the responses given. The creation of this centralized review process, where independent reviewers received identical training and held regular meetings on how to rate calls consistently, reduced the inconsistency and impartiality of rating the quality of service provided to taxpayers at the local level. Of course, this method of measurement also introduced other issues. Although the universe of test calls was modeled closely after the volume and topic of taxpayer inquiries, this national sample was not a sample of the universe of actual taxpayer calls, but a review of responses to fabricated questions posed by persons other than real taxpayers. Any data gathered from this test was an artificial measurement of the accuracy of information IRS assisters provided to the public. Additionally, ITCSS measured tax law calls only. For most other types of calls coming in on the IRS toll-free lines, it is necessary to know the identity of the caller and access his or her tax records to completely and accurately respond to the inquiry. This would not be possible with test callers. Also, after a time, even though the test questions were changed periodically, the CSRs were often able to identify calls from quality reviewers.

Centralized Quality Review Site

Because of the limitations of ITCSS, IRS eventually moved away from the test call system and created the Centralized Quality Review Site (CQRS) in Philadelphia. This site was established to centralize the IRS telephone review process into one location; to sample real, live calls from the universe of actual taxpayer inquiries; and to establish an estimate of the true level of service being provided to taxpayers. They were also charged with standardizing the review process of telephone calls and centralizing IRS telephone quality review data into one database.

The site, established in 1997, initially began with eight reviewers measuring the quality of tax law calls only. Over time, further types of calls were added. The CQRS now has over 50 full-time reviewers who monitor tax law calls, taxpayer account-related calls, collections calls, calls from the tax-practitioner priority line, calls from U.S. taxpayers overseas, calls from employers seeking business taxpayer identification numbers, and all Spanish-language calls, as well as requests for IRS tax forms.

The CQRS was able to establish an impartial measurement of quality for each call site and type of call by utilizing technology that enabled them to remotely monitor live taxpayer telephone calls coming into any IRS call site across the country. They reduced inconsistencies in the review process through holding regular consistency training, as well as utilizing a standard data collection instrument that gathered the same data elements for all calls and stored the information in a central database. And because they were monitoring real, live calls, the quality measurement generated from the review data produced the clearest picture of the level of service provided to taxpayers since IRS implemented quality review.

This standardization of the review process and improvement in the consistency of reviews were a major step toward accurately measuring the quality of service IRS provided to callers. Also, with remote monitoring, neither the caller nor the CSR was aware if that particular call was selected for review. This transparency meant that the monitored response was real, typical of the type of taxpayer/CSR interaction, and not altered in any way. Despite these advances, many new issues were introduced as IRS moved to monitoring live taxpayer

telephone calls. Because the telephone calls were live, reviewers were required to monitor the calls real-time. While this sounds innocent enough, real-time monitoring had a tremendous impact on reviewer resources.

Monitoring taxpayer telephone calls for quality in real-time consumes a considerable amount of resources. Initially, the largest depletion was due to dead air. A reviewer would dial into a site to monitor calls. If there were no available calls at that site (no taxpayers calling in or an unscheduled site closing), the reviewer would not know until listening to several minutes of silence or dead air. With the acquisition of software called Custom View, which allows the reviewers to see call traffic in the sites (real-time, less a 6-second delay), this problem was virtually eliminated. However, that was not the only problem with real-time monitoring.

To select a call for review, the quality reviewer at CQRS dials into a designated telephone number for a given site and type of call and is then attached to the next incoming call. The reviewer stays with the call, as long as it is in that particular site, and is able to hear the complete CSR/taxpayer interaction. Because it is a live telephone call, the reviewer experiences what the taxpayer experiences, including time on hold or waiting for the next available assister. Any hold or wait time is wasted time for a reviewer and can dramatically impact the number of telephone calls that he or she can monitor. Unfortunately, there is no way to eliminate these phenomena when monitoring live telephone calls.

Additionally, in order to sample from the entire universe of calls when monitoring in real time, the CQRS must have reviewers scheduled during all times of day that IRS call centers are open. Unfortunately, there are not enough review resources to cover all hours of operation, which are 7 a.m.-2 a.m. Eastern Time, meaning some calls are not subject to quality review. The CQRS does have staff monitoring phone lines from 7 a.m.-12 a.m.; so, only those calls received during the very early morning hours of 12 a.m.-2 a.m., less than 3 percent of the total universe of taxpayer calls, are not subjected to sampling for quality review.

Another minor issue associated with real-time monitoring is that the only record of the content of the call is the reviewer's notes. If the reviewer is unable to catch something that is said during the call, it can never be reheard or recaptured. A reviewer's determination of the accuracy of the call is dependent on what he or she is able to hear and jot down during the call. This can become an issue if a call site objects to the reviewer's evaluation of a call. Formal rebuttals from sites, requesting a re-evaluation of monitored calls, are frequently sent to CQRS for response. Unfortunately, since the disputed calls cannot be replayed, reviewer notes are the only evidence of what occurred during the call, leaving some room for continued disagreement.

Independent of the problems associated with real-time monitoring is the issue of call transfers. Using the current communications technology available at IRS, reviewers are unable to follow a call if it is transferred outside of the original site receiving the call. If customer service representatives receive a call that they are unable to answer, they must transfer that call to another assister. If that call is then routed to another call site, in the current telecommunications environment, the CQRS reviewer is unable to follow the selected call. Therefore, the reviewer cannot determine if the taxpayer received the correct answer to the inquiry. This situation is becoming increasingly more common given the current operational push for call site specialization, where assisters at a given call site are trained to answer only specific types of calls. Whether the taxpayer selected the wrong option from the automated menu or because the initial CSR who screened the call misunderstood the taxpayer's question, calls that are misrouted would have to be transferred to another site rather than to another assister within the same site. This increase in call transfers would result in an increased number of calls selected for review that CQRS reviewers would not be able to follow to completion.

▶ Recording Taxpayer Calls for Quality Purposes

Until recently, the recording of taxpayer telephone calls, while legal if the act of recording is disclosed to callers, was not permitted based on guidance from IRS Counsel. Call recording was viewed as an invasion of taxpayer privacy. However, since call recording has become standard in the customer service arena, IRS

has revisited the issue and approved call recording for quality purposes only. With the aid of a vendor, IRS is now testing and piloting call recording in select call sites. Call recording is scheduled for complete installation and implementation in all IRS call sites by FY 2006. Telecommunications technology being implemented will enable IRS to record 100 percent of all incoming calls, then systematically select calls for quality review. The selected recordings would then be reviewed by CQRS reviewers and entered into the standardized database.

Because review will still occur at CQRS, all the advantages of this consistent third-party review remain. However, call recording also brings a number of additional benefits. Primarily, call recording eliminates many of the drawbacks of the real-time monitoring of telephone calls. Once reviewers are able to listen to a recording of the taxpayer/CSR interaction, there will be no more listening to dead air and no waiting on hold. They will be able to fast-forward through any wait time while the CSR is researching the taxpayer's issue. Reviewers will also be able to rewind the recording and relisten to portions of the call, or the entire call if necessary, to more accurately assess what occurred during the call. Additionally, if the topic of the call is beyond the scope of a reviewer's training, he or she will be able to flag the call for evaluation by another reviewer with more technical expertise. Reviewers will also be able to listen to a recording at any time of day, regardless of when the call was placed, eliminating the need and added cost of an evening shift. Furthermore, the late night calls occurring between the hours of 12 a.m. and 2 a.m., not previously subject to review, will be available for quality review sampling under call recording. All of these advances secured through the implementation of call recording allow for cost savings, resource savings, and improvements to the quality sampling and review process.

An additional and unexpected resource savings is the reduction in the number of rebuttals from sites that feel the evaluation of a call by the CQRS was incorrect. During the pilot process, those sites with call recording have been able to listen to any call received at their sites, including those CQRS reviewers may have evaluated for quality purposes. Because call site managers are able to access and listen to the real CSR/taxpayer interaction

rather than rely on reviewers' notes, they can immediately eliminate rebuttals for calls they believe were evaluated correctly without any CQRS involvement. Now, only those calls where there is a legitimate disagreement in the call evaluation, are forwarded to CQRS for further action, resulting in an overall decrease in the amount of resources spent on rebuttals and re-evaluations.

Another advantage of call recording is that recorded calls, once "sanitized" to remove any taxpayer-identifying information, may be used for training purposes. Recordings of real taxpayer/CSR interactions will allow IRS to train CSRs how best to respond to taxpayer issues. By listening to the same call, assisters can be taught to respond to many different situations in a consistent way. In the same manner, the recordings can also be used by CQRS managers to train reviewers to consistently evaluate the quality of the service provided to taxpayers calling IRS.

Call recording also provides a major advancement in the way IRS is able to measure quality. Because reviewers are currently unable to follow a call if it is transferred from one site to another, it is impossible to capture everything that individual taxpayers experienced from the start of the call to the very end of the call. With the implementation of call recording at all IRS call sites, all incoming toll-free telephone calls will be recorded at each site. Therefore, if a call is transferred from one site to another, the portion of the call after being transferred will be captured and recorded at the destination site. Because all calls will be tagged with a unique identifier as they enter IRS, multiple segments of a single call can be combined after-the-fact. Thus, for the first time since quality review began, IRS will be able to capture the complete taxpayer experience for any call in the universe.

Because call recording is still in the pilot phase and the necessary hardware and software have not yet been installed in all sites, it is difficult to identify any disadvantages or problems with using this methodology to capture and review calls for quality measures. Certainly, initial startup costs are extremely high, but the resource savings and added benefits of call recording should eventually outweigh those one-time costs.

▶ IRS Toll-Free Telephone Assistance Quality Measures

Once IRS monitors or records a call, how is the quality of that call measured? The methodology behind the measurement of quality has also evolved over the years.

Past Methodology of Telephone Quality Measurement

Before FY 2004, there was a single measure for the quality of telephone calls coming into the IRS toll-free telephone assistance service. This measure was reported internally to IRS executives, and externally to Congress, GAO, and the Office of Management and Budget (OMB). Quality for a call was measured as pass or fail, where if one element, or "attribute," of a call was incorrect, the entire call was counted as incorrect. An attribute is any individual element of the call that is rated for accuracy by IRS reviewers. Depending on the taxpayer's issue, some attributes of a call may be: Did the assister greet the taxpayer courteously and professionally? Did the assister verify the taxpayer's Social Security number? Did the assister give the taxpayer the correct answer to their question? Did the assister provide his or her identification number? Certain attributes, while required elements during a call, may not affect the correctness of the answer provided to the taxpayer. Also, not all attributes are applicable to every type of call; so, one particular call may have very few attributes, while another may have several. Regardless of the number of attributes for any given call, a single call could only have the possibility of scoring 0 percent or 100 percent. Under the pass/fail methodology previously, if an assister answered the taxpayer's question correctly, but forgot to provide his or her identification number at the start of the call, as required under IRS procedures, the call was scored as 0 percent for quality measurement purposes. While this practice encouraged attention to details on the part of the telephone assisters, it presented an unclear measure of the quality of service provided to taxpayers, especially to external users of the data.

Current Methodology of Telephone Quality Measurement

In an attempt to construct a more accurate picture of the quality of the service provided to taxpayers, the pass/fail methodology was retired, and a new measurement system was implemented for FY 2004. This new method of measurement, defects-per-opportunity, was designed to distinguish between wrong answers and procedural errors that do not affect the accuracy of the answer provided to the taxpayer. IRSs single measurement for quality was separated into five individual measures:

- Customer Accuracy--Did the assister give the taxpayer the right answer?

- Regulator Accuracy--Did the assister follow all IRS regulations according to the tax code?

- Procedural Accuracy--Did the assister follow all internal IRS procedures for this type of call?

- Timeliness--Did the assister respond to the taxpayer in a timely manner?

- Professionalism--Did the assister respond to the taxpayer in a courteous and professional manner?

Given the nature of the measures, Customer Accuracy, Timeliness, and Professionalism are reported externally; whereas, Regulatory and Procedural Accuracy are measures intended for IRS use only. The five measures are each calculated as a percentage: the number of correct attributes divided by the total number of applicable attributes. Because Customer Accuracy has only one applicable attribute for any type of call--Did the taxpayer receive the correct answer?--a single call still only has the possibility of scoring 0 percent or 100 percent. However, with the elimination of all other non-applicable attributes, this measure of accuracy is now a very clear representation of the quality of the service

provided to taxpayers. Each of the other four measures generally has multiple applicable attributes for each call; thus, a single call can now score 0 percent, 100 percent, or anywhere in between.

Using FY 2003 data, Customer Accuracy was calculated using both methods:

FY 2003 data	Pass/Fail	Defects-per-Opportunity
Tax Law	80.10% (+/- 0.66%)	81.97% (+/- 0.63%)
Accounts	68.43% (+/- 0.44%)	88.15% (+/- 0.30%)

There was little effect on Tax Law calls, but the difference in the accuracy of Account calls is significant. This is due to the fact that, for Account calls, telephone assisters are generally required to perform many internal procedures where an error may occur that does not affect the accuracy of the answer provided to the taxpayer. Under the old pass/fail methodology, this would have caused the entire call to be counted as incorrect. Now, attributes relating to internal procedures are included in the measures of Procedural Accuracy and Regulatory Accuracy and are no longer included in Customer Accuracy, providing IRS executives, as well as Congress, GAO, and OMB, a clearer picture of the quality of service provided through the toll-free telephone assistance service--that assisters actually gave callers a correct answer approximately 88 percent of the time rather than the 68 percent previously reported.

▶ Future of IRS Telephone Quality Measures

Over the years, the way IRS monitors telephone calls and measures quality has undergone continuous

improvement. With the implementation of call recording, the Service will have taken the next step in the process. Beyond call recording, there are plans in motion to combine national quality review performed by CQRS and local managerial review into one standard database. This will be the last step in the standardization of the review process. Once completed, managers and quality reviewers will be reviewing calls using the same attributes and standards, and all quality data will be stored in a single place. This will provide individual call sites with additional data for error and trend analysis, allowing them to identify specific areas where additional training might improve quality.

In another move to improve the quality process, individual attributes from reviewed calls will soon be directly linked to telephone assisters' critical job elements (CJEs). CJEs are the specific items managers use to rate the performances of their employees. For example, call attributes regarding courtesy and professionalism will be linked with the professionalism CJEs for telephone assisters. The attribute for whether or not the taxpayer received the correct answer will be linked with the technical knowledge CJEs for assisters. Through this linkage, managers will be able to use their reviews to quantify the performances of their employees rather than relying solely on qualitative data and subjective judgments.

▶ Conclusion

A significant goal of IRS is to make its telephone operations a world-class customer service organization. By improving the way the level of service provided to taxpayers is measured, IRS can not only better determine how closely it has come to achieving that goal but can also identify areas for further improvement. With this continuous cycle of measurement and improvement, we hope this goal of providing world class customer service to taxpayers will soon become reality.

3

Research in Income and Wealth

Johnson ◆ Moore

Consider the Source: Differences in Estimates of Income and Wealth From Survey and Tax Data

Barry Johnson, Internal Revenue Service, and
Kevin Moore, Board of Governors of the Federal Reserve System

One implication of the decentralized nature of the statistical system in the United States of America, composed of over 70 Federal Government organizations, is that the data used by lawmakers and researchers to develop and evaluate Government policies come from a variety of sources. Survey and administrative data sources are frequently blended to create information systems capable of supporting a variety of research purposes. Because these two types of data are primarily designed for different purposes, one inherently created for research and the other for administration of Government programs, blending them generally poses serious challenges. This paper examines the comparability of administrative and survey data, focusing specifically on data from Federal income and estate tax returns collected by the Statistics of Income (SOI) Division of the U.S. Internal Revenue Service (IRS), and the Survey of Consumer Finances (SCF) sponsored by the Board of Governors of the Federal Reserve System. Through the use of two case studies, we detail key similarities and differences between these two data sources and demonstrate methods for reconciling estimates produced from them.

We then briefly discuss the Statistics of Income program and the Survey of Consumer Finances. We also discuss in detail differences between administrative and survey data, using administrative data from tax returns and SCF data to illustrate key points. We then present detailed comparisons of wealth estimates derived from U.S. estate tax returns and from the SCF, followed by a section comparing estimates from U.S. income tax returns and the SCF. The final section summarizes key points.

▶ The Statistics of Income Program

The Statistics of Income Division of the Internal Revenue Service was established almost immediately after the adoption of a Federal income tax in 1916 and was charged with the annual preparation of statistics with respect to the operation of the tax law. The first SOI report, based on income tax returns filed by individuals and corporations for Calendar Year 1916, was released in 1918. From the very beginning, SOI reports were almost entirely used for tax research and for estimating revenue, especially by officials in the Office of Tax Analysis of the Department of the Treasury and in the Congressional Joint Committee on Taxation. In the 1930's, a third major user of SOI data was added, the Bureau of Economic Analysis in the Department of Commerce, which uses SOI data extensively in constructing the National Income and Product Accounts. As the SOI program and products have expanded, users in other Government agencies, such as the Census Bureau, as well as many private and academic researchers, have come to rely on tax data produced by SOI for evaluating tax policy initiatives (see Wilson, 1988 for a complete history of the SOI program).

In order to fulfill its charge, SOI created a structured mechanism for transforming administrative data into statistical files, using its own data collection systems, completely autonomous of main IRS tax return processing. SOI currently conducts approximately 110 different projects involving data collection from returns and information documents; this paper will highlight two of these projects, the individual income tax file (ITF) and the estate tax data file (ETD). Data content is developed working closely with data users so as to ensure both continuity and usefulness. For most studies, data are extracted from stratified random samples of returns as they are filed to ensure timeliness. Specially trained employees located in IRS submissions processing centers collect the data under the supervision of subject matter experts from SOI headquarters. These specialists supply data editing instructions, conduct training classes, and review difficult cases. Data are entered into computer databases and checked using embedded tests that verify coded values and key mathematical relationships. In addition, subsamples of edited returns are subjected to field-by-field quality review. Finally, subject matter experts carefully review all files for accuracy before they releasing them to customers.

▶ Survey of Consumer Finances

The SCF is a survey of household balance sheets conducted by the Federal Reserve Board in cooperation with the SOI division of the IRS. Beginning with 1983, the survey has been conducted triennially, with data collected by the Survey Research Center at the University of Michigan in 1983, 1986, and 1989, and by NORC, a national organization for social science and survey research at the University of Chicago, from 1992 forward. Besides collecting information on assets and liabilities, the SCF collects information on household demographics, income, relationships with financial institutions, attitudes toward risk and credit, current and past employment, and pensions (for more details on the SCF, see Aizcorbe, Kennickell, and Moore, 2003).

The SCF uses a dual frame sample design to provide adequate representation of the financial behavior of all households in the United States. One part of the sample is a standard multistage national area probability sample (Tourangeau et al., 1993), while the list sample uses the IRS-SOI Individual Taxpayer File (ITF) to oversample wealthy households (Kennickell, 2001). This dual frame design provides the SCF with efficient representation of both assets widely held in the population, such as cars or houses, and assets more narrowly held by wealthy families, such as private businesses and bonds. Wealth data from the SCF are widely regarded as the most comprehensive data available for the United States.

Sample weights constructed for the SCF allow aggregation of estimates to the U.S. household population level in a given survey year (Kennickell and Woodburn, 1999; Kennickell, 1999). Missing values in the 1983 and 1986 SCF were imputed using a single imputation technique, while missing values in the subsequent 1989-2001 SCF were imputed using a multiple imputation technique (Kennickell, 1991, 1998b).

▶ Administrative Records and Survey Data

The American Statistical Association (1977) defines an administrative record as "[data] collected and maintained for the purpose of taking action on or controlling actions of an individual person or other entity." In the U.S. Government, administrative records have a long history of use in the production of Government statistics. In recent years, technological advances have made it easier for statistical agencies to process large datasets, encouraging even greater use of administrative records for research purposes. As a research tool, administrative records have many potential uses, including direct tabulation and indirect estimation of models or other statistics, as well as construction of survey frames and evaluation of survey results (Brackstone, 1987). In the best situations, administrative data may have several advantages over traditional survey data, including more complete coverage of a population (sufficient for regional statistics), low data collection costs, reduced respondent burdens, and better data quality. The potential problems with using administrative data for statistical purposes include the stability of a program over time, privacy concerns about nonadministrative use of data, conceptual issues relative to the population and items collected, and costs of transforming the data into a form useful for research purposes.

Surveys differ from administrative data in terms of their purposes, and such differences often have implications for their statistical structure, conceptual framework, and content. Almost all surveys are conducted to answer specific classes of research or public policy questions versus fulfilling an administrative function. This difference in purpose is reflected in the population frame, the unit of observation, the sample size, and the scope of the data. Some advantages of survey data over administrative data include the targeting of a specific population and variables of interest, the interaction with the respondent, and the ability to pledge that the data will be used solely for statistical (that is nonadministrative) purposes. Potential problems with survey data include difficulties in constructing a suitable frame, lack of legally mandated participation, high costs of increasing sample size, unit and item nonresponse, and measurement error. The following sections will examine all these issues in more detail.

Frame Issues

The population covered by a system of administrative records is defined through legislation, based on the scope of the program the records are intended to sup-

port. Often this population is truncated in some way, restricted based on specific demographic or economic characteristics. In some cases, individuals may have to take some action to become part of the administrative system (e.g., filing a tax return); so, it is important to consider what incentive there is for individual units to be registered. There may be perceived advantages for some individuals to evade registration, particularly if their circumstances place them at or near a threshold requiring mandatory participation. The populations of both Federal income and estate tax filers, for example, include only those U.S. citizens and resident aliens whose gross incomes, or gross estates, concepts defined by statute, were above specified thresholds. For each tax system, nonresident aliens are subject to different filing requirements, based on income earned or assets owned in the U.S. Income tax filers represent roughly 61 percent of the U.S. individual population, while estate tax filers have generally represented fewer than 5 percent of total annual U.S. deaths (see Sailer and Weber, 1999; Johnson and Mikow, 2002). Recent income tax filing gap estimates for Tax Year 2000 suggest that as many as 11 million taxpayers, or about 9 percent of the potential income tax filing population, either file returns late or not at all (see Brown and Mazur, 2003).

The Federal Committee on Statistical Methodology's (FCSM) *Statistical Policy Working Paper 6--Report on Statistical Uses of Administrative Records* points out that the unit of observation useful for statistical purposes often focuses on the attributes of groups of individual entities, while administrative records are often focused on identifying specific entities in order to take some sort of action based on their individual characteristics. Thus, the unit of observation available from administrative records may make certain research difficult or impossible. Records may contain information about individuals rather than families or households, or may be a mix of both individuals and households. In the case of Federal income taxes, married couples may file returns jointly, but they are also allowed to file separately in cases where marginal tax rates favor treating the two incomes separately. Dependent children and others living in a home may also be required to file separate returns to report both earned and unearned income. Differences in the economic unit reported on income tax returns limit the data's usefulness for some types of

research. Similarly, Federal estate tax returns represent only the decedent's wealth, including one-half the value of all community property [1] and property held as joint tenants [2]; assets owned independently by a surviving spouse are not reported.

The population targeted by a survey is determined by the purpose of the survey, the availability of a sampling frame, and the cost of the sample. The sampling frame for most surveys is derived from existing sources, such as geographically based population data, address listings, telephone directories, or administrative sources. Often, one of the most difficult issues with designing a survey is finding an appropriate frame (Lessler and Kalsbeek, 1992). Selecting the wrong sampling frame may lead to issues of undercoverage and may bias any results obtained from the survey data. A related problem arises if a survey targets a population that is difficult to locate or measure.

Directly related to the availability of a sampling frame is the potential cost of obtaining the frame information and the cost of interviewing a sample of the desired size. For target populations that are difficult to locate or appear infrequently in the frame, the cost of simply increasing the sample size to obtain better coverage can be prohibitive, although, sometimes, a frame contains information that may be used to target rare groups more efficiently. For example, one of the main goals of the Survey of Consumer Finances (SCF) is to measure the wealth of U.S. households. However, because wealth is highly concentrated in the population, sufficient coverage would require a very large area-probability sample. To this end, the SCF uses a dual-frame sample design in which an oversample of "wealthy" households is targeted using statistical records derived from tax returns provided by SOI [3]. Use of this sampling frame allows the SCF to collect data from wealthy households in a cost-effective and statistically efficient manner.

For survey data, the unit of observation is usually determined by the type of data required to answer certain research or policy questions. However, the choice of the unit of observation is also influenced by the type of sampling frame available to survey designers. In the SCF, the area-probability sample uses a sampling

frame in which the household is the unit of observation, but, for the list sample, the unit of observation is the tax-filing unit. Often, the tax-filing unit is analogous to the household, but, for certain households, such as households where a married couple files separately and households with multiple subhouseholds located within a household, there are differences. While there is the possibility of frame errors in the list sample, adjustments are made during the construction of the frame and during the sampling stage to limit the distortions (see Kennickell and McManus, 1993; Frankel and Kennickell, 1995; Kennickell, 1998a; and Kennickell, 2001).

Content Issues

The purpose for which administrative records were collected can have a profound effect on their usefulness for statistical purposes in terms of the amount of data available, data definitions, year-to-year consistency, and quality of the data. Many times, the usefulness of administrative record systems is limited because only those variables needed to administer the program are collected. These variables may be only a small fraction of the data reported on an administrative form.

In addition, because program requirements are established by legislation, data concepts and definitions used to meet program needs may not necessarily coincide with those required for social or economic analysis (Brackstone, 1987). For example, income for married couples is combined for joint filers of U.S. income tax returns; however, for some research purposes, it would be useful to know the amounts earned by each individual. When research and administrative needs differ, it can be very difficult to affect changes or improvements in content since statistical uses are often seen as secondary to an agency's primary purpose (FCSM Working Paper 6). This can pose serious limits on the overall usefulness of administrative data systems or require that the administrative agency undertake additional data collection and/or editing, incurring costs and delaying data availability.

Another consideration is that, while administrative records have much potential as a source of information on small geographic areas, to be useful, a precise geographic location code is needed. However, mailing addresses, frequently present on administrative records, may not always be the appropriate location, as when a post office box number is supplied rather than a street address. For Federal tax returns, addresses might be those of paid preparers rather than filers. In some instances, a filer may even own several residences.

An important aspect of data content is continuity over time, both in the items included and in the data definitions. Coverage and content in administrative records systems can be subject to discontinuities resulting from changes to laws, regulations, administrative practices, or program scope (Brackstone, 1987). For example, income tax law revisions in 1981, 1986, 1990, and 1993 all made significant changes to both the components of income subject to taxation and the allowable deductions from income that had significant impact on the statistical uses of tax return data (see Petska and Strudler, 1999). More recent changes in tax law will incrementally increase the filing threshold for estate tax return filers, from $675,000 in 2001 to $4,000,000 by 2009, and then abolish the tax entirely in 2010.

Data quality may also be a concern in administrative records systems. FCSM Working Paper 6 cautions that there can be considerable variation in quality across variables in an administrative records system. Information that may be statistically important, but only marginally relevant to administrative purposes, is often imperfectly reported, checked, and processed. Data items used primarily as background information may be of particularly low quality or even incomplete. This can also be the case for data collected specifically for statistical purposes using existing administrative channels. These items may be of lower quality if their priority is not very high to the administering authority or to the subject supplying the information (Jensen, 1987). Finally, data reliability may also be affected if the information respondents provide may be used to cause gains or losses to individuals or businesses. Underreporting on tax returns, for example, may have resulted in underpayment of as much as $120 billion in income taxes and $3.5 billion in estate taxes for Tax Year 1998 (Brown and Mazur, 2003).

FCSM Working Paper 6 suggested that administrative records sources are often a reliable source of timely data produced with predictable frequency. However,

since data collected and processed for administrative purposes are generally given priority over those required for statistical purposes, the amount of postprocessing required to render administrative data suitable for statistical purposes may affect data timeliness. In addition, the time and difficulty required to create desired statistics can vary considerably depending on a variety of factors. For example, for some research purposes, income data for households, rather than individuals, are required. To reconstruct households requires linking information documents with income tax returns filed by dependent filers and married couples who filed separately, using unique taxpayer identification numbers, all at the cost of significant resources (see Sailer and Weber, 1996).

Because surveys are freer than administrative systems to specify a conceptual framework, many issues related directly to the definition and scope of the data are less pressing. However, content and valuation issues of a different sort are present in survey data. One key issue is the voluntary nature of response to surveys versus the legally mandated participation in most administrative data programs. In most surveys, interviewers (either in person or via telephone) attempt to convince respondents to voluntarily donate time and information when there may be no direct benefit or punishment if a respondent refuses. Even if a respondent agrees to participate in the survey, it is still possible that the respondent will refuse to answer the questions truthfully and completely. Unit and item nonresponse are two important sources of non-sampling error in surveys; however, there are methods to help deal with both these issues, such as weighting and imputation.

For respondents who agree to participate and answer all the survey questions, measurement error is still a concern in survey data. Respondents may "guestimate" answers to questions; even if respondents' guesses overall are unbiased, such approximation reduces the estimation efficiency of the data. Respondents may also have difficulty recalling past events. Other typical measurement errors include rounding of dollar amounts, misunderstanding questions, and altering responses due to stigma or prestige attached to certain behaviors or a desire to protect privacy. A large volume of research exists on measurement error and its effects on survey data (see Lessler and Kalsbeek, 1992 and the references within).

While it is true that, for administrative data, unit and item nonresponse are usually not a problem on core items, it is not clear that administrative data are always more accurate than survey data. An example is the income values reported on IRS tax forms versus the income values reported in survey data; some individuals may intentionally misreport values on tax returns to reduce their tax liabilities. Those same individuals may report the true value in response to a survey question since there is no benefit to misreporting in the survey (via a lower tax liability).

Another content issue for survey data is the timeliness of the data. While many simple surveys are administered quite frequently, such as monthly, most of the more complex surveys occur yearly or even less frequently. Cost and other resource constraints are major factors in the timeliness of the survey data. For example, due to the high cost, complexity, significant data processing, and high respondent burden, the SCF is conducted on a triennial basis.

A final content issue for survey data is validation of the data. While it is sometimes possible to conduct validation studies after a survey is complete, these studies add additional cost to the survey. Validation of some items might require the cooperation of respondents, and requesting such cooperation may trigger suspicions in respondents that might lead to overall lower cooperation with a survey. Sometimes, selected data items are validated against external data sources, such as the Census or administrative data, but, often, no source for validation exists. This is in contrast to some administrative data, such as wages reported on tax forms, where amounts reported by filers are validated against amounts reported by their employers.

Privacy Issues

Any use of administrative records for research purposes must take account of laws protecting data privacy. In the U.S., privacy protections are either spelled out explicitly in agency-specific confidentiality statutes and regulations, or derived from Governmentwide statutes, such as the Privacy Act of 1974 (5 U.S.C. § 552a), and more recently, the Confidential Information Protection and Statistical Efficiency Act of 2002 (44 U.S.C. §

3501) (CIPSEA). In both instances, research uses of administrative data are often restricted to uses within the scope of an agency's mission and must be conducted by persons working for the agency as employees, contractors, or under the Government's Interagency Personnel Act (5 U.S.C. §§ 3371-3375) provisions that allow State government and nonprofit organization employees to work under the same provisions as employees as long as certain conditions are met. Other researchers are usually limited to public-use data sets or data tabulations, for which great care is taken to minimize the possibility of reidentifying data related to specific individuals. Public perceptions of privacy protection are vitally important to maintaining the goodwill required to sustain compliance levels, especially for agencies, like the IRS, which rely heavily on voluntary compliance for the success of their programs.

Government survey data are also often protected by the various privacy and confidentiality laws that apply to administrative data. The confidentiality of the respondent's data is of paramount importance to the current and future success of any survey. If respondents do not believe their data are sufficiently protected, both response rates and the overall data quality in the survey will suffer. Confidentiality and privacy laws provide important safeguards against potential abuse of respondent data by survey sponsors. In addition, surveys that produce publicly available data sets also must engage in a disclosure review to safeguard the identity of the respondents. The data collected during the SCF are protected by the Privacy Act of 1974, CIPSEA, and the Internal Revenue Code through an agreement with SOI. Information on the SCF disclosure review process is detailed in Fries (2003).

▶ **Wealth Data**

Both the SCF and Federal estate tax return data (ETD) provide important sources from which to study privately held wealth in the U.S. Both data sources collect extensive information on real estate, financial assets, businesses, tangible assets, and debts. The SCF also contains demographic information on household members, as well as extensive income and pension data. Federal estate tax returns provide a more limited demographic profile of the decedent, information on the

costs of administering the estate, and data on bequests to charities, the surviving spouse, and other living persons. Figure 1 provides a comparison of data available from both sources.

While there are many similarities between types of data available from the SCF and ETD, there are important structural differences. Some of the most significant include unit of observation, population coverage, and sample size. The SCF is a household survey which uses as its core unit of observation the "primary economic unit," which can consist of a number of different social arrangements, most commonly married or partnered pairs of individuals, and single persons, including those who were widowed, separated, divorced, or never married at the time of the survey, and all others in the household who are considered interdependent with them. Individuals living in institutions, such as nursing homes, are excluded from the area probability portion of the sample but may be in the list sample. All but the very wealthiest households, those with total assets of more than $600 million, are included in sample population [4]. The unit of observation in ETD is always an individual, and the population is limited to individuals with gross estates above the filing threshold applicable on the date of death, $675,000 for 2001 decedents [5].

One of the strengths of ETD is the large sample size. For example, the 2001 estate tax decedent file includes 17,376 records for individual decedents with total assets of at least $675,000. Of these, 9,322 were married, while 8,054 were widowed, single, divorced, or separated. The SCF includes 1,531 households with this level of wealth, only about 200 of which were either headed by widowed, single, divorced, or separated individuals. The large ETD sample size allows reasonably precise estimates for specific demographic groups, as well as geographic estimates by region or state.

While population estimates of wealth from both the SCF and ETD are based on weighted samples, there are significant differences in the method used to calculate the sample weights, which may have an impact on estimates derived from each source. Sample weights for the SCF are calculated using information from the sample design and are constrained using known population totals. Estimates of wealth from ETD rely on a multiplier which

Figure 1: Comparison of SCF and ETD File Content

Variable	Estate Tax Data	Survey of Consumer Finances
Demographic data:	Name, State of residence, year of birth, year of death, marital status, occupation, surviving spouse, (children, others if heirs) previously deceased spouse--year of death, name	State, year of birth, age, marital status, years married, previous marriage information, educational attainment, occupation, household characteristics including age of spouse, number of children, other dependents, age of parents
Real Estate:		
Personal residence	Single family, multiunit, ranch, mobile home; lot size; value (usually from real estate appraisal valued on date of death); mortgage amount	Single family, multiunit, ranch, mobile home; length of time living there; number of acres, value; mortgage type, amount, payment information; rent received
Rental property	Single family, multiunit, ranch, mobile home; lot size; value (usually from real estate appraisal valued on date of death); mortgage amount	Single family, multiunit, ranch, mobile home; length of time owning; value; rent received
Farm property	Value; acreage; mortgage amount	Value; acreage; mortgage type, amount, payment information
Financial Assets:		
Closely held stock	Name of corporation; number of shares; percentage ownership; market value; appraisal	Actively managed: number of businesses, for 3 largest: year formed, type, cost, method of financing, value, income received. For others: total value, cost, income. Nonactively managed: value, cost, type, income received
Publicly traded stock	Number of stocks, market value, name of corporation, brokerage account information	Number of stocks, market value, gain or loss, location (in the U.S. or not) employer stock (yes or no), brokerage account information
U.S. Government bonds	Market value	Face value, market value
Federal Savings bonds	Market value	Face value
Tax-exempt bonds	Market value	Face value, market value
Corporate bonds	Market value	Face value, market value
Mutual funds	Type of fund (stock funds, tax-exempt bond funds, Government-backed bond funds, other bond funds, combination or mixed funds), value	Type of fund (stock funds, tax-exempt bond funds, Government-backed bond funds, other bond funds, combination or mixed funds), type of institution, value, gain or loss since purchase
Noncorporate Businesses	All businesses, active, nonactive. Value at death, appraisals or balance sheets.	Actively managed: number of businesses, for 3 largest: year formed, type, cost, method of financing, value, income received. For others: total value, cost, income. Nonactively managed: value, cost, type, income received
Trusts	Revocable trusts, marital trusts: detailed listing of assets, value. Split Interest trusts: value, assets invested, charitable beneficiary. Other income trusts may not be reported.	Type (income only, equity), amount of annual income, value, indication of how assets are invested
Bank accounts	Type of account (money market, traditional savings, certificate of deposit), current balance, ownership	Type of institution, type of account (money market, traditional savings, certificate of deposit), current balance, ownership
Life insurance	Face value, accrued interest, policy loan amount	Term and whole life: face value, cash value, policy loans (purpose and payment information), premiums

Mortgages and notes	Amount owed to decedent	Amount owed to respondent
Retirement assets:		
Annuities	Equity: value, detailed listing of assets. Income not usually reported unless there is a death benefit or lump sum value.	Type (income only, equity), amount of annual income, value, indication of how assets are invested
401K, Keogh, etc.	Number of accounts, value. Detailed listings of investments are usually provided	Type (education, Roth, Keogh, rollover), number of accounts, type of institution, value
Pensions	Only pensions where surviving spouse is also a recipient so that a portion is included in the taxable estate	Detailed information on pensions from multiple jobs for primary economic unit including type, contribution amount, benefit amount, timing of payments, death benefits, etc.
Social Security Payments	Not reported	Amount received, reason for payment
Other:		
Art/antiques/collectibles; Depletable/ intangible, livestock, proceeds from lawsuits, lottery winnings, futures	Type, amount	Type, amount (up to three different categories)
Vehicles/boats/etc.	Type; value for all vehicles; model and year usually supplied for automobiles; loan amount	Automobiles: first 4--model, year, financing, value, purchased new or used, Others: financing, value. Other vehicles: first 2--type, financing, value, purchased new or used, Others: financing, value.
Debts:		
Consumer debt	Amount owed	Amount of original loan, type, payment information, balance owed, purpose, collateral, type of institution, payment history
Mortgages	Amount owed	Amount of original loan, type, payment information, balance owed, type of institution, payment history

incorporates both the probability of being selected into the SOI sample of estate tax returns and the age and sex-specific probability of being a decedent in a particular year (see Atkinson and Harrison, 1978, for a description of this methodology). Mortality rates, by age and sex, are used to approximate the probability of being a decedent. Because there is no way to control for the weighted population total, the selection of an appropriate mortality rate is important. Research has shown that the wealthy live longer than the general population due to factors such as access to better health care, safer work environments, and better nutrition. While estimates of patterns of wealth holding appear quite robust over a variety of reasonable alternate assumptions about the longevity of the very wealthy, overall aggregate estimates are relatively sensitive to the selection of the mortality rates. Mortality rates calculated for holders of large dollar value annuity policies are used for these estimates.

Valuation Issues

There are significant differences in the determination of asset values in the ETD and SCF. Estate tax returns are generally accompanied by a great deal of documentation to support reported valuations, including tax returns, brokerage account statements, appraisals, business accounting reports, and legal documents. In contrast, only about 32 percent of SCF respondents use such documents when providing valuation data, although extremely wealthy survey respondents often refer to financial documents or seek assistance from their accountants in order to provide accurate data.

While the more systematic presence of valuation documentation may make ETD a potentially more accurate source of wealth data than survey estimates, the administrative nature of ETD imposes important considerations. Unlike questions on the SCF that have been

carefully constructed to capture data needed for specific research purposes, data reported on estate tax returns are influenced by provisions in the tax law, estate planning mechanisms, and the point in the life cycle at which data are collected. For example, the tax code allows certain adjustments in asset values, such as the special valuation of real estate used for farming or certain business purposes, and includes some items, particularly the face value of life insurance and trust property over which a person had a limited power of appointment, that might not ordinarily be considered part of lifetime wealth [6]. In addition, the tax code generally exempts from tax other wealth to which a person has an income interest, but not necessarily actual title, such as defined-benefit pension plans, simple trusts, and Social Security benefits.

A number of other factors can contribute to differences in the values of assets captured in the ETD and those collected on the SCF. While estate tax returns are generally prepared by professionals and are, therefore, likely to be more precise in detail than survey responses, the values are used to compute tax liability; so, there is a natural tendency for the values to be as conservative as legally permissible. This is especially true for hard-to-value assets, such as businesses and certain types of real estate. It should also be noted that the ETD collected by SOI are pre-audit figures. While we believe that the relatively high audit rate for estate-tax returns ensures that complete evasion is relatively rare, the values reported may be subject to underreported and missing values, the later due to informal transfers of small items such as jewelry [7]. In addition, it is common to claim substantial discounts when valuing ownership interests of less than 50 percent in small companies, partnerships, and other nonliquid assets. The creation of family lim-

ited partnerships and other estate planning techniques can significantly reduce the asset values included in a decedent's estate by taking advantage of these discounts [8]. Finally, the wealth of some estate tax decedents may differ significantly from that of the general population in the same age cohort, due to expenses related to final illnesses. In addition, when death is anticipated, decedents may have altered the composition of their assets in order to simplify their finances, to provide liquidity to pay for health-related expenses, and to ensure that family-owned business operations are not disrupted by their deaths.

Direct Comparisons Between SCF and ETD Data

The study of wealth includes many goals, only one of which is the determination of point estimates for various populations and subpopulations. The previous section pointed out important structural differences between the SCF and ETD. A key research question then is do these two datasets provide similar analytical results, despite these differences? Focusing on total assets as the measure of wealth, the SCF data show that there were more than 13.4 million households with total assets of $675,000 or more, while the ETD data show that there were more than 6.1 million individuals at or above that wealth threshold. The mean age for heads of household in the SCF was 56, and the median age was 54. For ETD, the mean and median ages were both 60 [9]. Estimates for widowed, single, separated, or divorced persons provide the best opportunity for direct comparisons between the two datasets since the units of observation should be closely aligned. Figure 2 provides a direct comparison of wealth components for the SCF and

Figure 2
Comparisons of SCF and Estate Tax Data Estimates of Wealth, by
Marital Status, for Households or Estates with >= $675,000 in Assets
(Money amounts are in thousands of dollars)

	Survey of Consumer Finances				Estate Tax Estimates			
	% reporting	Mean	Median	Total	% reporting	Mean	Median	Total
Single/widowed/div/sep								
Total assets	100.0	2,102	1,099	4,564,262,000	100.0	1,833	1,068	4,822,014,000
Financial assets	100.0	1,122	653	2,435,399,000	100.0	1,189	745	3,108.671,000
Nonfinancial assets	98.5	980	488	2,128,862,000	96.0	678	343	1,713,343,000
Personal residence	85.0	286	230	620,366,000	67.1	320	240	564,534,000
Other real estate	50.7	270	17	586,918,000	36.1	386	215	367,051,000

Note: SCF and ETD estimates are based on samples.

ETD, for unmarried or unpartnered units with at least $675,000 in total assets. The SCF data show that there were 2.17 million single/widowed/divorced/separated households in 2001 with total asset holdings worth nearly $4.6 trillion, while ETD estimates show 2.6 million such individuals with more than $4.8 trillion in total assets. Financial assets compose 53 percent of total assets in the SCF, but account for nearly 65 percent of the total in the ETD estimates. Nevertheless, the mean and median values for financial assets are similar between the two groups, with SCF values somewhat lower than ETD values. Total nonfinancial assets have somewhat higher mean and median values in the SCF estimates. The mean and median values for personal residences in both datasets are remarkably similar, despite the higher incidence of this asset reported in the SCF and the fact that personal residences account for a smaller portion of total assets in the ETD estimates.
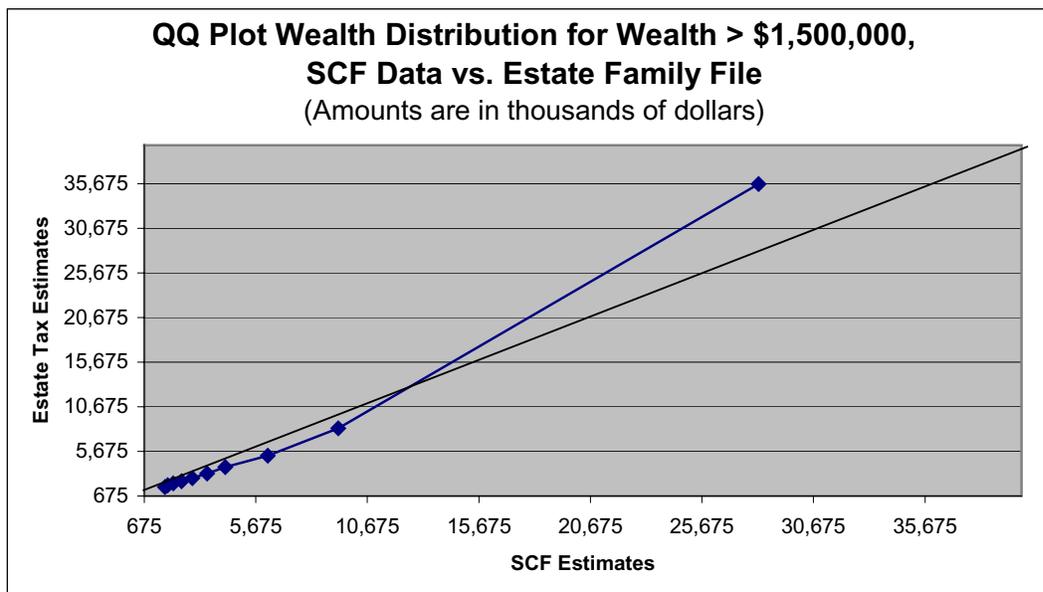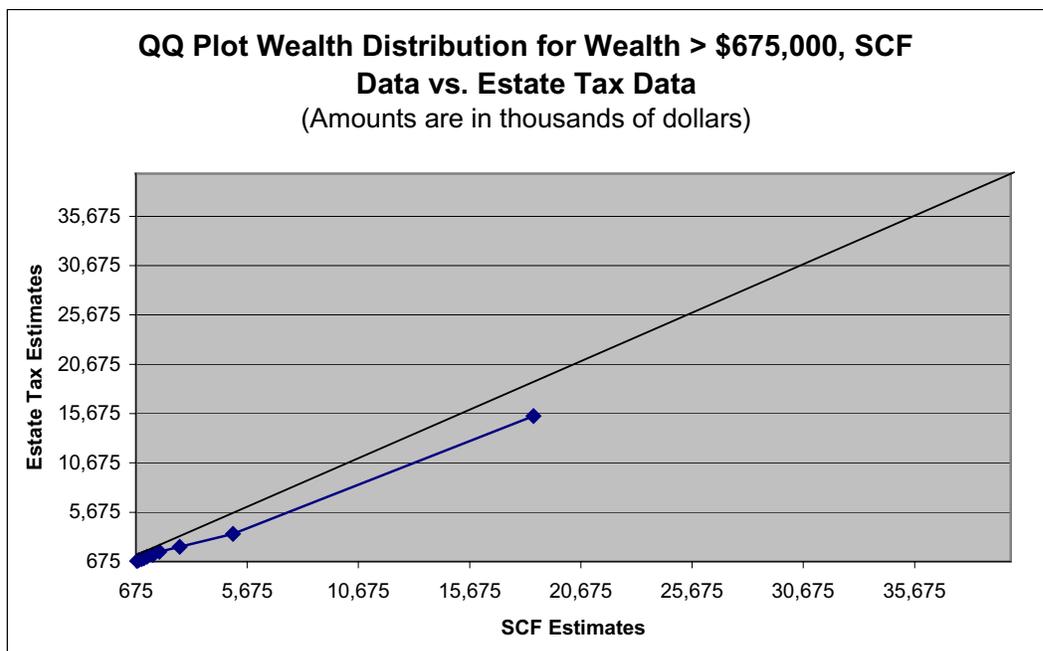
Because point estimates for married households in the SCF include assets of both partners while estimates from the ETD are for only one of a pair, direct comparisons are not meaningful. However, it would be useful to know whether differences in the estimates are primarily attributable to differences in the unit of measurement and population coverage, or if these differences are masking more fundamental structural differences between the two data sets. In order to examine these issues, it is necessary either to divide households in the SCF to create individuals, or to impute households from individuals in the ETD. There have been a couple of attempts to simulate the estate tax filing population using SCF data (see for example Poterba and Weisbrenner, 2001; Eller et al., 2001). However, these efforts have been limited by the sample size of the SCF and the sensitivity of the resulting estimates to assumptions about the relative share of household assets attributable to each separate spouse. We choose instead to impute households for married individuals in the ETD. A sketch of the procedures follows (see Johnson and Woodburn, 1994 for a full description of this process).

While estate tax returns provide detailed information on property held jointly with a surviving spouse, they provide virtually no other information on the wealth owned separately by the survivor, making model-based imputation of households infeasible. Instead, hotdeck imputation is used to approximate the wealth of a survivor spouse (see Hinkins and Scheuren, 1986, for a detailed discussion of hotdeck imputation). Married decedents are separated into two groups, based on sex, under the simplifying assumption that decedents on the file, as a group, had characteristics similar to those of the surviving spouses [10]. Adjustment cells are constructed based on the value of jointly held property, within broad age strata, and male decedents were paired randomly with a female decedent, within adjustment cells, to form families. Additional weight adjustments are needed to account for households where the female decedent's wealth is above the estate tax filing threshold, but where the separate wealth of her spouse is below the threshold. Still missing from this simulated household file are households where each partner's independent wealth is below the estate tax filing threshold, but where their combined gross assets exceed $675,000. By choosing a high enough threshold, for example $1.5 million, the effects of these missing households on final estimates should be minimized.

The resulting imputed family data set, while only crudely approximating household wealth for married individuals and ignoring nontraditional households that would be included in the SCF, can nevertheless be used to test whether the two data sources are measuring the same underlying wealth distribution. Figure 3 graphically compares the distributions of total assets using quantile-quantile (QQ) plots. If the distributions implied by the data sets being compared are similar, the plots will form a straight line. Deviation from the 45-degree line indicates variance between the two sets of estimates. The first graph compares the ETD with the SCF. Note that the QQ plot is nonlinear, meaning that the distributions are functionally different. The second graph compares the imputed family data set to estimates from the SCF and truncates the distributions at $1.5 million. In this graph, the plots for the 10th through 90th percentiles are approximately linear and much closer to the 45-degree line than was the case for the untransformed ETD estimates. The values in the SCF are still somewhat larger than ETD, as would be expected. Differences at the 99th percentile, where the ETD estimates are much higher, reflect the sample variance of both datasets, particularly the SCF, which has very few observations at this level of wealth. Overall, these results suggest that the two

Figure 3

QQ Plot Wealth Distribution for Wealth > $675,000, SCF Data vs. Estate Tax Data
(Amounts are in thousands of dollars)

QQ Plot Wealth Distribution for Wealth > $1,500,000, SCF Data vs. Estate Family File
(Amounts are in thousands of dollars)

data sets produce roughly equivalent measures despite having different units of observation.

If both the ETD and SCF are observing essentially the same population characteristics, they should provide similar estimates of economic trends. One trend that is often

considered an important measure of the overall economic well-being in the U.S. is the concentration of wealth, defined here as the share of total wealth owned by a fixed portion of the population. As shown in Figure 4, the SCF estimates reveal that the wealthiest 1 percent of households owned between 30 percent and 35 percent of total household

wealth between 1989 and 2001, with an increase between 1992 and 1995 and a slight decrease after that. Estimates for individuals in the top 1 and top ½ percent of the population constructed from ETD show a similar trend, with a slight increase in the middle of the period, but with concentration in 2001 about the same as in 1989.

▶ Income Data

Both the SCF and the ITF file are important sources of data on the different types of income received by households and tax filers. The main differences between the two sources are the unit of observation, sample size, and the motivations people face in providing data. While much has been said about the differences in the unit of observation in the two data sources, it is also worth noting the difference in the sample size. The ITF file is a sample of approximately 175,000 tax records, but the sample size for the 2001 SCF is a much smaller 4,449 households. Although the SCF has a smaller sample, the detail and scope of the data allow for a broader range of research than is possible with the tax data.

Valuation Issues

The income questions in the SCF are structured to allow the respondents to reference their tax forms when answering the income questions. Figure 5 shows the correspondence between the income questions in the SCF and the line number on IRS Form 1040. The SCF variable numbers that correspond to each line of the IRS Form 1040 are listed on Figure 5. As shown in Figure 5, the SCF income questions were designed to cover most forms of income that a household reports on its tax form. Since the SCF is interested in all sources of household income and not just income subject to taxation, the questions on pensions, IRA/401(k) distributions, annuities, and Social Security payments refer to the total amounts. The SCF also asks about any income from nontaxable investments, such as municipal bonds, and any income received from Government transfer programs (such as TANF, SSI, and food stamps). Households are not questioned about any adjustments to total income (lines 23-31a on Form 1040), but households are questioned about their Adjusted Gross Income (AGI, line 33). All
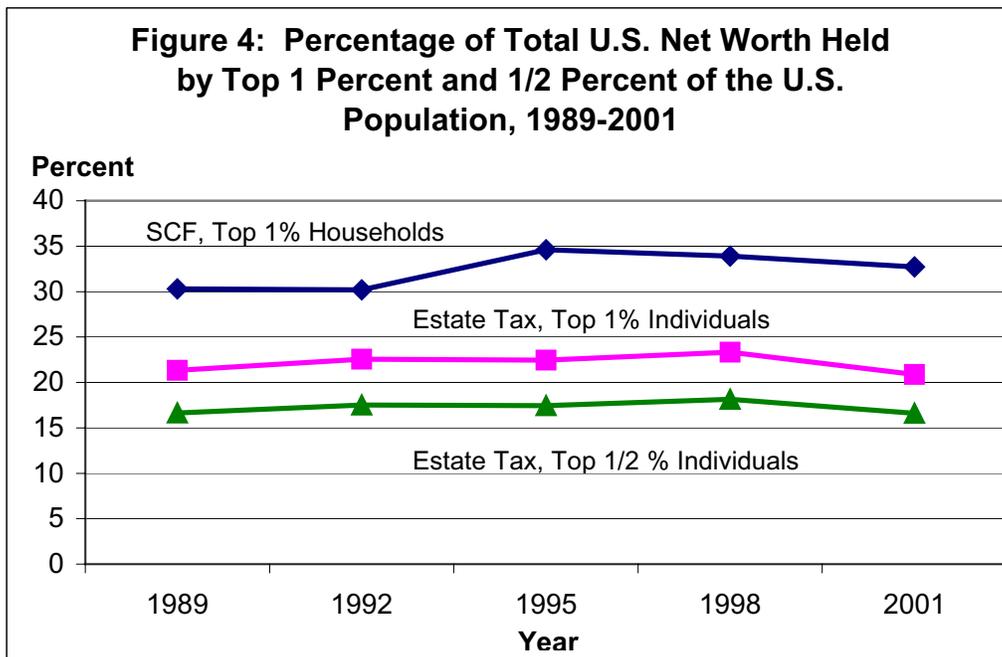
Figure 4: Percentage of Total U.S. Net Worth Held by Top 1 Percent and 1/2 Percent of the U.S. Population, 1989-2001

Percent

SCF, Top 1% Households

Estate Tax, Top 1% Individuals

Estate Tax, Top 1/2 % Individuals

Year

Figure 5

Form **1040** Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return **2000** (99) IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2000, or other tax year beginning , 2000, ending , 20 | OMB No. 1545-0074

Label
(See instructions on page 19.)
Use the IRS label. Otherwise, please print or type.

Your first name and initial | Last name | Your social security number

If a joint return, spouse's first name and initial | Last name | Spouse's social security number

Home address (number and street). If you have a P.O. box, see page 19. | Apt. no.

City, town or post office, state, and ZIP code. If you have a foreign address, see page 19.

▲ **Important!** ▲
You **must** enter your SSN(s) above.

Presidential Election Campaign (See page 19.)
Note. Checking "Yes" will not change your tax or reduce your refund.
Do you, or your spouse if filing a joint return, want $3 to go to this fund? ▶
You ☐Yes ☐No Spouse ☐Yes ☐No

Filing Status
Check only one box.
1 ☐ Single
2 ☐ Married filing joint return (even if only one had income)
3 ☐ Married filing separate return. Enter spouse's social security no. above and full name here. ▶
4 ☐ Head of household (with qualifying person). (See page 19.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 ☐ Qualifying widow(er) with dependent child (year spouse died ▶). (See page 19.)

Exemptions
6a ☐ **Yourself.** If your parent (or someone else) can claim you as a dependent on his or her tax return, **do not** check box 6a
b ☐ **Spouse**
c Dependents:
(1) First name Last name | (2) Dependent's social security number | (3) Dependent's relationship to you | (4) ✓ if qualifying child for child tax credit (see page 20)

If more than six dependents, see page 20.

d Total number of exemptions claimed

No. of boxes checked on 6a and 6b
No. of your children on 6c who:
● lived with you
● did not live with you due to divorce or separation (see page 20)
Dependents on 6c not entered above
Add numbers entered on lines above ▶

Income
Attach Forms W-2 and W-2G here. Also attach Form(s) 1099-R if tax was withheld.

If you did not get a W-2, see page 21.

Enclose, but do not attach, any payment. Also, please use Form 1040-V.

7 Wages, salaries, tips, etc. Attach Form(s) W-2 | 7 | X5702
8a **Taxable** interest. Attach Schedule B if required | 8a | X5708
b Tax-exempt interest. **Do not** include on line 8a | 8b | X5706
9 Ordinary dividends. Attach Schedule B if required | 9 | X5710
10 Taxable refunds, credits, or offsets of state and local income taxes (see page 22) | 10 |
11 Alimony received | 11 | X5718
12 Business income or (loss). Attach Schedule C or C-EZ | 12 | X5704
13 Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ ☐ | 13 | X5712
14 Other gains or (losses). Attach Form 4797 | 14 | X5712
15a Total IRA distributions | 15a | X5724 | b Taxable amount (see page 23) | 15b |
16a Total pensions and annuities | 16a | X5722 | b Taxable amount (see page 23) | 16b |
17 Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E | 17 | X5714
18 Farm income or (loss). Attach Schedule F | 18 | X5704
19 Unemployment compensation | 19 | X5716
20a Social security benefits | 20a | X5722 | b Taxable amount (see page 25) | 20b |
21 Other income. List type and amount (see page 25) | 21 | X5724
22 Add the amounts in the far right column for lines 7 through 21. This is your **total income** ▶ | 22 |

Adjusted Gross Income
23 IRA deduction (see page 27) | 23 |
24 Student loan interest deduction (see page 27) | 24 |
25 Medical savings account deduction. Attach Form 8853 | 25 |
26 Moving expenses. Attach Form 3903 | 26 |
27 One-half of self-employment tax. Attach Schedule SE | 27 |
28 Self-employed health insurance deduction (see page 29) | 28 |
29 Self-employed SEP, SIMPLE, and qualified plans | 29 |
30 Penalty on early withdrawal of savings | 30 |
31a Alimony paid b Recipient's SSN ▶ | 31a |
32 Add lines 23 through 31a | 32 |
33 Subtract line 32 from line 22. This is your **adjusted gross income** ▶ | 33 | X5751,X7651,X7652

For Disclosure, Privacy Act, and Paperwork Reduction Act Notice, see page 56. Cat. No. 11320B Form **1040** (2000)

Tax and Credits	**34**	Amount from line 33 (adjusted gross income)	**34**
	35a	Check if: ☐ **You** were 65 or older, ☐ Blind; ☐ **Spouse** was 65 or older, ☐ Blind. Add the number of boxes checked above and enter the total here . . . ▶ **35a**	
	b	If you are married filing separately and your spouse itemizes deductions, or you were a dual-status alien, see page 31 and check here . . . . ▶ **35b** ☐	

Standard Deduction for Most People

Single: $4,400

Head of household: $6,450

Married filing jointly or Qualifying widow(er): $7,350

Married filing separately: $3,675

36	Enter your **itemized deductions** from Schedule A, line 28, **or standard deduction** shown on the left. **But** see page 31 to find your standard deduction if you checked any box on line 35a or 35b **or** if someone can claim you as a dependent	**36**
37	Subtract line 36 from line 34	**37**
38	If line 34 is $96,700 or less, multiply $2,800 by the total number of exemptions claimed on line 6d. If line 34 is over $96,700, see the worksheet on page 32 for the amount to enter .	**38**
39	**Taxable income.** Subtract line 38 from line 37. If line 38 is more than line 37, enter -0- .	**39**
40	**Tax** (see page 32). Check if any tax is from **a** ☐ Form(s) 8814 **b** ☐ Form 4972 . . .	**40**
41	Alternative minimum tax. Attach Form 6251	**41**
42	Add lines 40 and 41 . . . . ▶	**42**
43	Foreign tax credit. Attach Form 1116 if required	**43**
44	Credit for child and dependent care expenses. Attach Form 2441	**44**
45	Credit for the elderly or the disabled. Attach Schedule R . .	**45**
46	Education credits. Attach Form 8863	**46**
47	Child tax credit (see page 36)	**47**
48	Adoption credit. Attach Form 8839	**48**
49	Other. Check if from **a** ☐ Form 3800 **b** ☐ Form 8396 **c** ☐ Form 8801 **d** ☐ Form (specify) _____	**49**
50	Add lines 43 through 49. These are your **total credits**	**50**
51	Subtract line 50 from line 42. If line 50 is more than line 42, enter -0- . . . . ▶	**51**

Other Taxes		
	52	Self-employment tax. Attach Schedule SE	**52**
	53	Social security and Medicare tax on tip income not reported to employer. Attach Form 4137	**53**
	54	Tax on IRAs, other retirement plans, and MSAs. Attach Form 5329 if required	**54**
	55	Advance earned income credit payments from Form(s) W-2	**55**
	56	Household employment taxes. Attach Schedule H	**56**
	57	Add lines 51 through 56. This is your **total tax** . . . . ▶	**57**

Payments		
	58	Federal income tax withheld from Forms W-2 and 1099 . .	**58**
	59	2000 estimated tax payments and amount applied from 1999 return	**59**
If you have a qualifying child, attach Schedule EIC.	**60a**	**Earned income credit (EIC)**	**60a**
	b	Nontaxable earned income: amount . . ▶ _____ and type ▶ -------	
	61	Excess social security and RRTA tax withheld (see page 50)	**61**
	62	Additional child tax credit. Attach Form 8812	**62**
	63	Amount paid with request for extension to file (see page 50)	**63**
	64	Other payments. Check if from **a** ☐ Form 2439 **b** ☐ Form 4136	**64**
	65	Add lines 58, 59, 60a, and 61 through 64. These are your **total payments** . . . . ▶	**65**

Refund		
Have it directly deposited! See page 50 and fill in 67b, 67c, and 67d.	**66**	If line 65 is more than line 57, subtract line 57 from line 65. This is the amount you **overpaid**	**66**
	67a	Amount of line 66 you want **refunded to you** . . . . ▶	**67a**
	▶ **b**	Routing number _____ ▶ **c** Type: ☐ Checking ☐ Savings	
	▶ **d**	Account number _____	
	68	Amount of line 66 you want **applied to your 2001 estimated tax** . ▶ **68**	

Amount You Owe		
	69	If line 57 is more than line 65, subtract line 65 from line 57. This is the **amount you owe**. For details on how to pay, see page 51 . . . . ▶	**69**
	70	Estimated tax penalty. Also include on line 69 . . . . . **70**	

Sign Here

Joint return? See page 19.

Keep a copy for your records.

Under penalties of perjury, I declare that I have examined this return and accompanying schedules and statements, and to the best of my knowledge and belief, they are true, correct, and complete. Declaration of preparer (other than taxpayer) is based on all information of which preparer has any knowledge.

Your signature	Date	Your occupation	Daytime phone number ( )
Spouse's signature. If a joint return, **both** must sign.	Date	Spouse's occupation	May the IRS discuss this return with the preparer shown below (see page 52)? ☐ **Yes** ☐ **No**

Paid Preparer's Use Only

Preparer's signature ▶	Date	Check if self-employed ☐	Preparer's SSN or PTIN
Firm's name (or yours if self-employed), address, and ZIP code ▶		EIN	
		Phone no. ( )	

Form **1040** (2000)

income amounts reported in the SCF are for the year prior to the survey year.

Even with the close correspondence between the income questions in the SCF and IRS Form 1040, accurate classification and reporting of income amounts are still a potential problem in the SCF. While households are encouraged to reference documents during the interview, in the 2001 SCF, only about 32 percent of households referenced any type of documents. However, of those households that used documents, 43 percent referenced their tax forms. The ability of households that did not reference their tax forms to accurately recall and classify income introduces potential bias or inefficiency into the SCF income estimates. Although the legal penalties for misreporting income provide a strong incentive for filers to report accurate amounts to the IRS, evasion and misclassification may still bias the estimates and introduce inefficiencies.

Direct Comparisons Between SCF and SOI Data

Figure 6 provides a comparison of SCF and SOI income for the 2000 tax year. The first row of Figure 6 highlights the difference in the unit of observation between the two data sources. In the SCF, the unit of observation is the household, which can often contain more than one tax unit. The SCF asks the filing status of the core individual or couple in a household, thus allowing married or partnered households filing separately to be counted as two returns. The SCF underestimates the number of returns, no doubt in large part because the SCF does not ask about the filing status of other individuals within the household. These individuals include dependents who may also file a return and other members of the household who are not financially dependent on the household head or the core couple.

Figure 6

Comparing Components of Total Income from the SCF to the IRS Values, All Returns

(Money amounts in thousands of dollars)

Tax Year	2000		
Data Source	SCF	IRS	% Diff
Number of Returns	102,825,058	129,373,500	-25.8
Components of Total Income			
Wages and salary	4,985,506,700	4,456,167,438	10.6
Business income	651,515,251	213,865,353	67.2
Nontaxable interest	54,929,226	54,511,136	0.8
Taxable interest	138,970,069	199,321,670	-43.4
Dividends	107,561,912	146,987,679	-36.7
Capital gain/loss	492,696,443	630,542,431	-28.0
Rent, royalties, s-corp	180,621,157	238,022,618	-31.8
Unemployment	14,625,905	16,913,305	-15.6
Alimony	26,683,086	6,192,307	76.8
Pensions, annuities, SS	459,542,345	738,596,530	-60.7
Other income	49,438,841	25,370,158	48.7
Total	7,162,090,935	6,726,490,625	6.1
Memo item:			
Broad business income	1,324,832,851	1,082,430,402	18.3

Notes: SCF values are for households who filed or intend to file a tax return.

IRS values from Tables 1.3 and 1.4 in *Statistics of Income—2000, Individual Income Tax Returns*.

Broad business income includes business income, capital gain/loss, and rent, royalties, and

S corporation income.

For the components of total income, Figure 6 shows no clear pattern in the comparison of the two data sources; the SCF overestimates five and underestimates six of the income components relative to the SOI estimates. Of the eleven income components, the SCF and SOI estimates are within +/- 30 percent for wage and salary, nontaxable interest, capital gains, and unemployment income. The differences for the seven other income components are quite large; SCF alimony income is 76 percent larger than the SOI estimate, and the amount of SCF pensions, annuities, and Social Security income is 60 percent less than the SOI estimate. The larger differences deserve further investigation.

Some of the differences in the SCF and SOI estimates are due to how each source defines an income component. For example, the SCF question on alimony income instructs the respondent to include child support payments. Since child support payments are nontaxable, such payments should not be included in the SOI estimate. One possible method for removing child support payments from SCF alimony income is to restrict the estimate of alimony income to households who report alimony income but have no children under the age of 25 in the household. This restriction reduces the amount of alimony income to $3.6 billion, which is about 58 percent of the SOI estimate ($6.2 billion).

The SCF underestimates the amount of taxable interest and dividends by 43 percent and 36 percent, respectively. A possible reason for these lower estimates is that households that receive small amounts of taxable interest or dividend income may forget to report these amounts in the SCF questionnaire. Even households with large interest income may find such income less salient if they are not in a phase of life where they would rely on such income for spending. Since the SCF collects extensive information on assets, it is possible to indirectly estimate the amount of income households might receive from their interest and dividend-producing assets. Unfortunately, the estimates of interest and dividend income obtained by applying average rates of return to these types of assets are even lower than the estimates derived from the SCF income questions. Two reasons for this difference are heterogeneity in the rates of return for different households and the sale or consumption of assets during the time prior to the survey interview.

Business income estimated by the SCF is over three times as large as the SOI estimate. However, note that the amount of capital gains and the amount of rent, royalties, and subchapter S corporation income reported in the SCF are about 30 percent lower than SOI estimates. The SCF definition of business income should be analogous to income reported on lines 12 and 18 of SOI Form 1040 (see Figure 5), but it is not unlikely that households may be misclassifying capital gains or rent, royalties, and subchapter S corporation income as business income. This may be partially due to the order of the income questions in the SCF, since the business income question is asked early in the income sequence, while the capital gains and rent, royalties, and subchapter S corporation income questions are asked later in the sequence. A broader definition of business income might include all three of these income measures; summation of the three measures reveals that the SCF estimate is about 18 percent larger than the SOI estimate.

Another large difference between the income estimates is that the SCF understates the total of pension, annuity, and Social Security incomes by 60 percent. By using information reported in other sections of the SCF, it is possible to compute alternative estimates of pension, annuity, and Social Security income. The sum of the three alternative estimates of these components is less than 2 percent larger than the estimate f total pension, annuity, and Social Security income derived from the summary income questions in the SCF. Furthermore, the SCF estimate of Social Security income is about 26 percent larger than the SOI estimate. Thus, the problem appears to be the estimate of pension and annuity income, not the estimate of Social Security income.

The estimate of "other" income, the final income component in Figure 6, is about 50 percent larger using the SCF data than the estimate using the SOI data. One possible reason for the difference is that the SCF definition of other income includes distributions from Individual Retirement Accounts (IRA) or 401(k) plans. If income from these sources is removed, the SCF estimate of other income falls by about $13.3 billion and is now only 30 percent larger than the SOI estimate.

As an attempt to shed further light on the differences between the two data sources, tax units and households are grouped by AGI class. One motivation for this grouping is that households in the SCF with at least $50,000 in AGI are twice as likely to have referenced tax forms during the interview as households with less than $50,000 in AGI (21.5 percent versus 10.3 percent). This suggests that households in the SCF with higher AGI should do a better job of reporting and classifying income. Another motivation for grouping filers or households by AGI is to determine if the differences between the two data source are driven by many small errors throughout the AGI distribution, or one specific segment of that distribution. Figure 7 presents the results of this exercise. For the less $50,000 AGI group, only the estimates of wages and salary and pension, annuity, and Social Security income are within +/- 30 percent. This stands in contrast to the $50,000 plus AGI group, in which all but five income components are within +/- 30 percent.

For the less than $50,000 AGI group, the largest differences are for taxable interest, dividends, and rent, royalties, and subchapter S corporation income. As discussed previously, the differences for taxable interest and dividend income may be due to many households neglecting to report relatively small amounts of these types of income. For example, for households with less than $50,000 in AGI that own interest-bearing assets, about 75 percent of these households do not report any interest income. Furthermore, the median amount of interest-bearing assets for the households that do not report any interest income is only $1,900 [11].

The large difference in the estimates of rent, royalties, and subchapter S corporation income for the less than $50,000 AGI group may be partly due to the treatment of losses in the SCF. Although the SCF allows households to record negative amounts for certain income questions, often households report zero instead of the actual loss. Given the tax treatment of losses, it is not surprising that losses are more likely to be reported to the IRS.

In contrast to the income estimates for all households, the amount of business income reported in the SCF for the less than $50,000 AGI group is lower than the SOI estimate. Again, for business income, it may be more useful to combine business income, capital gains, and rent, royalties, and subchapter S corporation income into one broad measure of business income. For the less than $50,000 AGI group, the SCF estimate of this broad business income measure is less than 1 percent larger than the SOI estimate.

Turning to the bottom panel of Figure 7, for households with $50,000 or more in AGI, the lack of large differences in the estimates for most of the income components is evidence that households referencing tax forms are good for the data. As for the large differences in the estimates of business income and rent, royalty, and subchapter S corporation income, using the broader definition of business income reduces this difference substantially. Under the broad business income definition, the SCF estimate is only 20 percent larger than the SOI estimate. Whether this difference is due to reporting error in the SCF or evasion in the SOI data is unclear.

The most striking result for the $50,000 or more AGI group from Figure 7 is that the SCF estimate of pension, annuity, and Social Security income is less than one-half the SOI estimate. As with the estimates for all households, the summation of the alternative SCF estimates of pension, annuity and Social Security incomes are only about 2 percent less than the SCF estimate derived directly from the income questions. Also, the SCF estimate of Social Security income is only about 17 percent less than the SOI estimate. Thus, the bulk of the difference between the SCF and SOI estimates is due to pension and annuity income. One possible reason for the discrepancy is the treatment of rollovers from one tax-deferred retirement to another tax-deferred retirement account. For example, if a household transfers the balance of one IRA account to another IRA account, the transfer is not taxable, but the transfer amount should appear on line 16a of Form 1040 (see Figure 5). Often households neglect to report these rollovers on their tax forms since there are no tax implications. However, the SOI estimate will include these rollovers, even if the household does not include them on its tax form [12]. Since households in the $50,000 or more AGI group are about twice as likely to have some sort of tax-deferred retirement account, these households may have more rollovers.

Figure 7

Comparing Components of Total Income from the SCF to the IRS Values,

By AGI Class, All Returns

(Money amounts in thousands of dollars)

Tax Year	2000		
Data Source	SCF	IRS	% Diff
AGI < $50,000			
Number of Returns	63,504,207	77,370,713	-21.8
Components of Total Income			
Wages and salary	1,495,908,100	1,514,257,995	-1.2
Business income	71,562,974	94,459,352	-32.0
Nontaxable interest	6,367,893	7,253,787	-13.9
Taxable interest	27,735,062	60,487,940	-118.1
Dividends	17,297,297	41,826,985	-141.8
Capital gain/loss	22,558,717	37,621,491	-66.8
Rent, royalties, s-corp	17,365,370	-21,255,979	222.4
Unemployment	9,033,543	12,204,865	-35.1
Alimony	14,568,265	4,357,077	70.1
Pensions, annuities, SS	272,705,769	294,763,093	-8.1
Other income	17,835,043	7,616,376	57.3
Total	1,972,938,034	2,053,592,982	-4.1
Memo item:			
Broad business income	111,487,061	110,824,864	0.6
AGI >= $50,000			
Number of Returns	39,320,851	32,798,001	16.6
Components of Total Income			
Wages and salary	3,489,598,600	2,941,909,441	15.7
Business income	579,952,277	119,406,001	79.4
Nontaxable interest	48,561,333	47,257,350	2.7
Taxable interest	111,235,007	138,833,728	-24.8
Dividends	90,264,615	105,160,694	-16.5
Capital gain/loss	470,137,727	592,920,941	-26.1
Rent, royalties, s-corp	163,255,787	262,335,219	-60.7
Unemployment	5,592,363	4,708,441	15.8
Alimony	12,114,821	1,821,107	85.0
Pensions, annuities, SS	186,836,576	443,833,436	-137.6
Other income	31,603,798	17,753,782	43.8
Total	5,189,152,905	4,675,940,140	9.9
Memo item:			
Broad business income	1,213,345,791	974,662,161	19.7

Notes: SCF values are for households who filed or intend to file a tax return.

IRS values from Tables 1.3 and 1.4 in *Statistics of Income—2000, Individual Income Tax Returns*.

Broad business income includes business income, capital gain/loss, and rent, royalties, and

S corporation income.

A final item to note from Figure 7 is that the SCF and SOI estimates of total income for each AGI group are remarkably close. This provides evidence that, although households may misclassify the components of income, the aggregate level of income is fairly consistent.

▶ Conclusions

Our research has shown that, while ETD and SCF data seem to be capturing very similar portfolio data for the wealthiest people in the U.S, differences in population coverage and the unit of observation make it very difficult to declare estimates from one source superior to the other. There is a great deal of evidence that the financial characteristics of the very wealthy are sufficiently heterogeneous to require quite large samples to make meaningful estimates for small subpopulations. It is also clear that the increasingly complicated financial and business arrangements practiced by the very wealthy require a great deal of attention to the definition of data variables when attempting any sort of analysis. Here, we are thinking about the proliferation of nontraditional investment instruments, such as derivatives, strips, options, and futures, as well as complex ownership arrangements, such as trusts, family limited partnerships, and holding companies. Lifecycle effects are also an important consideration; the portfolios of working individuals are different from those of the retired, which are also going to be different from individuals who face the end of their lives.

For studying broad trends in the population or for an overview of the top of the wealth distribution, the SCF provides more complete coverage than ETD. By focusing on households, the SCF data are uniquely suited for answering many complex economic questions and provide comparability with other publicly available national datasets. The availability of extensive savings, income, debt, work history, and demographic data also makes the SCF a much richer source of data than ETD for many research purposes. In addition, the sample design ensures that individuals at all phases of the lifecycle are included in the sample, thus providing a broad measure of the economic behavior of all households.

Data from U.S. estate tax returns provide a unique source of data on wealthy individuals. For many pur-

poses, such as the study of intergenerational wealth transfers, they are the only viable data source. The large sample size permits detailed study of individuals at the highest levels of the wealth distribution. ETD can also support detailed study of the wealthy in various demographic groupings, particularly by age, marital status, and sex, while these groups are not sufficiently represented in the SCF to allow reliable estimates. These demographic characteristics seem to be key determinants of behaviors such as portfolio choice, charitable giving, and bequest decisions. In addition, the abundance of valuation documentation provided with ETD provides unique opportunities to study in detail the financial planning and business arrangements employed by the wealthy to both minimize tax liability and to ensure that a legacy of wealth accumulation is preserved beyond their lifetimes.

Estimates for households made up of single, widowed, divorced, or separated individuals in the ETD and SCF were remarkably similar, and our simulations suggest that data for married or partnered households are likewise comparable. Overall, values reported on estate tax returns appear to be conservative relative to those in the SCF, reflecting the difficulty of valuing some assets, especially businesses; practical considerations, such as the difficulty of finding a willing buyer for a fractional interest in a basket of market goods; and the natural desire to minimize tax liability to the great extent possible within the constraints of the tax code. In addition, differences between the mean and median ages reported in the ETD and those in the SCF suggest that the use of mortality rates that reflect the longevity advantages enjoyed by the wealthy in constructing wealth multipliers may not completely compensate for overrepresentation of the elderly in the decedent population, perhaps introducing a slight bias. The ETD may also be biased by effective financial and estate planning, by expenses associated with a long final illness, and by changes in asset holdings made in anticipation of death.

In terms of the comparison between the SCF and SOI income data, our research has shown that, although there are differences in the unit of observation and issues with the definition of certain income types, the two data sources compare quite favorably. One reason for this is the close correspondence between the SCF income

questions and the income categories on IRS Form 1040. While it appears that households often misclassify income, the total amount of income reported by households in the SCF is only 6 percent larger than the SOI estimate. Due to the detail and scope of the SCF data, it is often possible to use data from other sections of the survey to make adjustments to better align the SCF and SOI income definitions. The detail and scope of the other data collected in the SCF also allow for a broader range of research than the SOI tax data. However, the large sample size and administrative nature of SOI tax data make it an appealing source for certain types of research, such a tax policy.

The direct comparison of the SCF and SOI income data reveals that encouraging households to reference their tax forms is critical for the accuracy of the SCF income data. Households with lower AGI may feel it is unnecessary to check their tax forms given the few types of income they receive, but it clearly makes a difference, as Figure 7 demonstrates. Households with higher levels of AGI are more likely to receive more types of income due to the increasing complexity of their financial situations. Thus, it is potentially even more difficult for these households to correctly report and classify their incomes without referencing their tax forms.

Overall, the message for researchers is that the SCF and SOI data are complementary sources of data on both wealth and income. The goal of our research is not to declare one data set superior to the other; that is a difficult judgment to render. What we have attempted to show in this paper is that there are many important issues to understand when comparing administrative and survey data. The key, then, is that each data source has strengths and weaknesses that need to be understood and carefully considered before attempting to use them to answer any set of research questions.

▶ References

Aizcorbe, Ana; Kennickell, Arthur B.; and Moore, Kevin B. (2003), "Recent Changes in U.S. Family Finances: Evidence from the 1998 and 2001 Survey of Consumer Finances," *Federal Reserve Bulletin*, Volume 89, pp. 1-32.

American Statistical Assocation (1977), "Report of the Ad Hoc Committee on Privacy and Confidentiality," *The American Statistician*, Volume 31, pp. 59-78.

Atkinson, A.B. and Harrison, A.J. (1978), *Distribution of Personal Wealth In Britain*, Cambridge University Press, Cambridge, England.

Brackstone, C.J (1987). "Statistical uses of Administrative Data: Issues and Challenges," *Statistical Uses of Administrative Data Proceedings,* pp. 5-26.

Brown, Robert E and Mazur, Mark J. (June 2003), "IRS' Comprehensive Approach to Compliance Measurement," 2003 National Tax Association Spring Symposium, http://www.irs.gov/pub/irs-soi/mazur.pdf.

Cartwright, David W. and Armknecht, Paul A. (1979), "Statistical Uses of Administrative Records," *Proceedings, Section on Survey Research Methods*, American Statistical Association, pp 73-76.

Eller, Martha Britton (2001), "Audit Revaluation of Federal Estate Tax Returns," *Internal Revenue Service Statistics of Income Bulletin,* Winter 2000-2001, Washington, D.C., pp. 100-139.

Eller, Martha Britton; Erard, Brian; and Ho, Chih-Chin Ho (2001), "Noncompliance with the Federal Estate Tax," in *Rethinking Estate and Gift Taxation*, William G. Gale, James R. Hines, and Joel Slemrod, editors, Brookings Institution Press, pp. 375-421.

Frankel, Martin and Kennickell, Arthur B. (1995), "Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods,* American Statistical Association.

Fries, Gehard (2003), "Disclosure Review and the 2001 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Garnick, Daniel H. and Gonzalez, Maria Elana (1979), "Statistical Uses of Administrative Records: Where Do We Go From Here?," *Proceedings, Section on Survey Research Methods*, American Statistical Association .

Hinkins, Susan and Scheuren, Frederick (1986), "Hot Deck Imputation Procedure Applied to a Double Sample Design," *Survey Methodology*, Volume 12, pp. 181-196.

Jensen, Paul (1987), "The Quality of Administrative Data From a Statistical Point of View, Some Danish Experience and Considerations, *Statistical Uses of Administrative Data Proceedings,* pp. 291-300.

Johnson, Barry W. and Mikow, Jacob M. (Spring 2002), "Federal Estate Tax Returns, 1998-2000," *Statistics of Income Bulletin*, Internal Revenue Service, Washington DC, pp. 113-186.

Johnson, Barry W. and Woodburn, Louise (1994), "The Estate Multiplier Technique, Recent Improvements for 1989," in *Compendium of Federal Estate Tax and Personal Wealth Studies*, Barry Johnson, editor, Internal Revenue Service, Publication 1773, pp. 391-400.

Kennickell, Arthur B. (2003), "A Rolling Tide: Changes in the Distribution of Wealth in the U.S., 1989-2001," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (2001), "Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (1999), "Revisions to the SCF Weighting Methodology: Accounting for Race/Ethnicity and Homeownership," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. (1998a), "List Sample Design for the 1998 Survey of Consumer Finances,"

working paper, Board of Governors of the Federal Reserve System.

Kennickell, A. (1998b), "Multiple Imputation in the Survey of Consumer Finances," *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association.

Kennickell, A. (1991), "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kennickell, Arthur B. and McManus, Douglas (1993), "Sampling for Household Financial Characteristics Using Frame Information on Past Income," working paper, Board of Governors of the Federal Reserve System.

Kennickell, Arthur B. and Woodburn, R. Louise (1999), "Consistent Weight Design for the 1989, 1992, and 1995 SCF's, and the Distribution of Wealth," *Review of Income and Wealth*, Series 45, 2, pp. 193-215.

Lessler, Judith T. and Kalsbeek, William D. (1992), *Nonsampling Error in Surveys*, John Wiley and Sons, New York.

Menchik, Paul (1991), "Economic Status as a Determinant of Mortality Among Nonwhite and White Older Males: or, Does Poverty Kill?," Institute for Research on Poverty, Discussion Paper Number 93891.

Petska, Tom and Strudler, Mike (1999), "The Distribution of Individual Income Taxes: A New Look at an Old Issue," *Turning Administrative Systems Into Information Systems, Internal Revenue Service*, pp. 7-22.

Poterba, James and Weisbenner, Scott (2001), "The Distributional Burden of Taxing Estates and Unrealized Capital Gains at Death," in *Rethinking Estate and Gift Taxation*, William G. Gale, James R. Hines, and Joel Slemrod, editors, Brookings Institution Press, pp. 422-456.

Sailer, Peter and Weber, Michael (1996), "Creating Household Data from Individual Income Tax Returns," *Proceedings, Section on Survey Research Methods*, American Statistical Association .

Sailer, Peter and Weber, Michael (1999), "The IRS Population Count: An Update," *Proceedings, Section on Survey Research Methods*, American Statistical Association.

Scheuren, Fritz (1994), "Historical Perspectives on IRS Wealth Estimates With a View to Improvements," *Compendium of Federal Estate Tax Data and Personal Wealth Studies*, Department of the Treasury, IRS Publication 1773, pp. 355-362.

Statistical Policy Working Paper 6--Report on Statistical Uses of Administrative Records (1980), Federal Committee on Statistical Methodology, United States Office of Management and Budget.

Tourangeau, Roger; Johnson, Robert A.; Qian, Jiahe; Shin, Hee-Choon; and Frankel, Martin R. (1993), "Selection of NORC's 1990 National Sample," working paper, National Opinion Research Center at the University of Chicago.

Wilk, M. B. and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, Volume 55, pp. 1-17. (as described in Hoaglin, David C. (1985), "Using Quantiles to Study Shape," *Exploring Data Tables, Trends, and Shapes*, John Wiley and Sons, New York.

Wilson, Robert (Fall 1988), "Statistics of Income: A By-Product of the U.S. Tax System," *Statistics of Income Bulletin*, Department of the Treasury, Internal Revenue Service, Volume 8, Number 2, pp. 103-114.

▶ **Footnotes**

[1]   In nine U.S. States, nearly all property acquired by a married couple is considered owned equally by both parties. Property acquired separately by gift or bequest is generally exempted.

[2]   In States where there are no community property rights, assets titled legally as joint tenants are considered owned equally by both partners in a marriage, usually without regard to how much consideration each party contributed to purchase the asset.

[3]   Details of the SCF list sample design are provided in Kennickell, 2001.

[4]   Due to the difficulty of gaining cooperation from the wealthiest individuals, the SCF uses as its upper sample threshold the minimum amount of wealth required for inclusion in the listing of the wealthiest 400 individuals in the U.S., as estimated by *Forbes* magazine. Kennickell (2001) discusses the methodology used for selecting the SCF list sample.

[5]   Gross estate is a measure similar to total assets, but which includes the full face value of life insurance, certain gifts made prior to death, and certain assets placed in trust.

[6]   Where possible, we modify the data to compensate for these reporting anomalies. For example, the full face value of life insurance is included in the decedent's total gross estate for tax purposes; however we impute a cash value using data from the SCF.

[7]   Examination rates vary by size of estate. In 2003, about 6.4 percent of all returns were examined, while 27.5 percent of those reporting estates of $5 million or more were subject to examination. A recent Statistics of Income (SOI) study, based on the results of IRS audits of estate tax returns filed in 1992, estimated that detected undervaluation of assets was about 1.2 percent of total asset holdings for all audited returns (Eller, et al., 2001).

[8]   A family limited partnership is a business arrangement in which a wide array of business and market assets are transferred to a partnership, with general partner interests held by parents

and limited partner shares distributed to children through annual tax-exempt gifts. This results in fractured ownership interests in the individual assets, qualifying them for large valuation discounts for tax purposes.

[9] The mean and median ages for heads of households with total assets of $1,500,000 from the SCF were both 57, virtually the same as for individuals in the ETD with this level of wealth, for whom the mean and median ages were 58.

[10] This approach will tend to overpredict wealth since some surviving spouses would in reality have less wealth than those available for matching in the ETD.

[11] For households with $50,000 or more in AGI that own interest-bearing assets, about 53 percent do not report any interest income. Median interest-bearing assets for these nonreporting households is $6,200.

[12] A rollover transaction generates a Form 1099-R that SOI matches to Form 1040. If a filer neglects to report the rollover on his or her tax form, the value from Form 1099-R is added to the filer's Form 1040.

4

Disclosure Protection Techniques

Greenia

Developing Adoptable Disclosure Protection Techniques: Lessons Learned From a U.S. Experience

Nicholas H. Greenia, Internal Revenue Service*

The development of new disclosure protection techniques is useful only insofar as those techniques are adopted by statistical agencies. For technical experts in disclosure limitation to be successful, they are likely to need to interact with the appropriate statistical offices. This paper discusses just such a successful interaction in the United States.

Since 2001, interagency efforts have been under way on a synthetic data approach to produce a public-use file (PUF), which would combine selected statistical and administrative data from three U.S. agencies: the Census Bureau's Survey of Income and Program Participation (SIPP), retirement and disability benefits data from the Social Security Administration (SSA), and limited earnings data from tax records filed with the Internal Revenue Service (IRS). Based on progress so far, the outlook for this work is promising. The confidentiality and research benefits of this approach, if successful, could be substantial, but details of that technical discussion are left for other papers.

It is important to note, however, that technological advances in disclosure protection are necessary, but not sufficient, conditions for the adoption of new techniques. This paper focuses primarily on describing the evolution of the legal, institutional, and bureaucratic environment that was the critical precursor of the interagency effort. Out of the story come lessons that may help other national statistical offices cope with similar challenges.

This story is largely a confluence of separate but related events:

- The development of an institutional interagency trust, after a serious test of the fundamental relationship;

- The recognition by the Census Bureau of the deteriorating tradeoff between data quality and data protection in the release of previous SIPP public-use files, which was influential in decid-

ing to pursue the synthetic data PUF approach; and

- The development of a new program (Longitudinal Employer-Household Dynamics) that brought in the technical know-how that permitted the integration of statistical and administrative data within the new program, and the creation of the aforementioned SIPP/SSA/IRS PUF.

This paper focuses primarily on the first of these, but also notes the relevance of the other events.

▶ Background

Statistical agencies have become increasingly aware that two relatively new challenges may seriously affect their ability to release data into the public domain, whether in tabular or public-use file format. Increasing capabilities of computing power and advances in mathematical/statistical techniques have led to the increase in technical reidentification capacity. This challenge is matched by a practical increase in this capacity due to the proliferation of datasets in the public and private/commercial domain. In spite of these challenges, the need for publicly collected confidential data to inform decisions in both government and the private sector is not expected to abate.

The U.S. tax administration agency, the Internal Revenue Service (IRS), faces additional challenges in its role as an important administrative data provider for the Federal statistical system. Tax data have always been particularly susceptible to reidentification, both because of their relatively widespread distribution in public form and because of their sensitive content. In addition, because publicly and privately available datasets are often directly based on entities also in the tax system, there is more potential to match to tax data and reidentify taxpayers. Moreover, IRS views the protection of taxpayer confidentiality as an essential component of

successful voluntary tax compliance, upon which the tax system relies. Because of the several U.S. statistical agencies authorized to receive confidential tax data, IRS must not only preserve tax data confidentiality within its own administrative system, but also oversee the safeguarding of tax data in the systems of the recipient statistical agencies. In a related vein, IRS must ensure that the numerous products produced by each statistical agency cannot be statistically "cross-matched" and thereby enable complementary disclosure of identifiable information.

Because of these additional challenges, IRS must insist that its safeguarding standards be met by a recipient statistical agency, regardless of the agency's standards for data it collects directly. This requirement of compliance with administrative data provider standards also influenced the authorization process for statistical use of tax data by Census, as will be shown later, but this requirement may differ for other countries. For example, the United Kingdom's Office of National Statistics stipulates that "the same confidentiality standards will apply to data derived from administrative sources as apply to those collected…for statistical purposes"[1]. Nevertheless, the unmistakable conclusion is that it is becoming increasingly difficult to release even aggregate tabular data into the public domain, and public-use files (often of most use to researchers without access to the original source data) pose special challenges that are exacerbated over time in the public domain. Although closer coordination of all releases is advisable, new methods of confidentiality protection may afford the most hope for data users, data providers, and ,ultimately, the respondents themselves.

While issues surrounding the disclosure of confidential data are common to all Federal statistical agencies, IRS also has its own idiosyncratic issues [2]. Confidential tax data, also known as Federal Tax Information (FTI), have several uses, including specifically authorized statistical purposes. The homogeneous treatment of FTI results from restrictions in the tax statute, the Internal Revenue Code (IRC), which do not allow IRS to distinguish among FTI data elements--even as to age. That is, there is no statute of limitations as there is for confidential microdata at statistical agencies such as the U.S. Census Bureau. In addition, the tax statute does not distinguish among different types of data or taxpayers, so that the Social Security Number of John Q. Citizen in Anywhere, USA, would receive the same protection as that of Bill Gates which, in turn, would be protected as much as all the financial information on any business tax return which Microsoft Corporation might file. Accordingly, all FTI--whether entity or tax module information--must be treated and protected in perpetuity as equally sensitive and confidential [3]. This task of protecting confidentiality, given the ever-increasing amount of data for which IRS becomes responsible over time, is expensive and technically challenging.

The tax law's anonymity standard is indiscriminate and absolute in requiring that all tax data, whether business or individual, be released in anonymous form. The anonymity requirement for data publicly released by IRS also applies to statistical agencies authorized to receive FTI. However, although the general standard applies, the actual disclosure protection methodology is not specified. The requirement is simply that whatever methodology is used be either identical to that employed by IRS or else an equivalent approved by IRS.

The practical question confronting any methodology attempting to meet the absolute anonymity standard is: From what sort of intrusion must the data be protected? Must it be absolutely impossible to reidentify a taxpayer using any means available, or is there some less rigid methodological standard? Traditionally, the answer has been that tax data must be protected from potential intruders who, using "reasonable means," might attempt to make such a reidentification. Reasonable means include the use of reasonably available computer technology, mathematical/statistical techniques, and a working knowledge of the subject matter to which the data apply. The reasonable means standard is a good effort to keep the entire system from shutting down and being replaced by a policy of no data release at all--probably the only way to guarantee no reidentification. The problem, as can probably be imagined in 2004, is that the concept of reasonable means is a technology-relative concept and may be a moving target too elusive to be relevant for the absolute standard of anonymity. As a result, in a time of increasingly tight budgets, protecting the confidentiality

of tax data is becoming a task virtually impossible to execute successfully.

▶ Developing Interagency Trust

A Breakdown in the Relationship

In 1999, IRS began its mandated triennial safeguards review of a principal U.S. statistical agency, the Census Bureau. Although the U.S. statistical system is more decentralized than that of many European Union countries, Census receives the preponderance of confidential tax data for statistical purposes as a result of the statutory authorization conferred by section 6103(j)(1)(A) of Title 26 of the United States Code (USC). The implementing Income Tax Regulations specify both the actual items authorized for access and their access purpose or Title 13, Chapter 5, USC.

The mandated IRS safeguards review of Census (and other recipient agencies of confidential tax data) is a result of the same section, 6103, which authorizes such access in the first place. As a result of the 1999 IRS safeguards review, deficiencies in the oversight process were uncovered by IRS, some of which reflected poorly on both Census and IRS. For example, Census used tax data for some projects which had not received explicit IRS approvals, but IRS had made explicitly clear neither the need for such approvals nor the process for effecting them in a coordinated fashion.

As it became clear that neither Census nor IRS could resolve the resulting crisis, intervention at high levels of government became necessary. Eventually, the U.S. Office of Management and Budget (OMB), which has broad oversight responsibilities for Federal statistical agencies, helped broker an understanding between the two agencies based upon three essential points:

(1) Census must comply with IRS safeguard standards in order to protect the confidentiality of tax data,

(2) informed decisions by policymakers inside and outside government require the best possible data available, and

(3) tax data are so important to these information decision systems that their exclusion is not a viable option.

Thus, the conclusion of this process was that IRS, as an administrative data provider, and Census, as an administrative data user, would have to find a way to make their relationship work in order to satisfy the several stakeholders involved; that is, an interagency "trainwreck" or shutdown was viewed as unacceptable and would not be tolerated.

As a result, IRS and Census recognized that the increasingly murky and implicit boundaries within which their relationship had been struggling were inadequate as guidance. Further, a relationship was needed which would not only work but which would better accommodate the increasingly complex needs of the many end users. Essentially, the relationship needed to be not only re-evaluated but also recalibrated, especially to accommodate a new form of confidential data access created by Census for outside researchers meeting new Census study needs: the Research Data Center (RDC) consortium operated by its Center for Economic Studies. Like statistical agencies in other countries, Census had realized the need to explore other venues for purposes of improving its statistical knowledge base and data quality, but only as a result of the IRS safeguards review did this realization include the need to integrate its RDC's into the overall process encompassing its other longstanding functions [4].

To meet especially the need for new statistical research uses of FTI, a clear and detailed understanding that met the mandates of both agencies needed to be documented. Accordingly, an IRS-Census policy agreement, *Criteria for the Review and Approval of Census Projects that Use Federal Tax Information*, better known as the Criteria Agreement, was mutually devised and eventually signed into effect by both agencies in September 2000. At the core of this agreement, available at www.ces.census.gov, was the understanding that any data use or access had to be authorized by an explicit approval process involving both the data provider, IRS, and the data user, Census, and that, especially for outside researcher access, the predominant purpose

of such access had to be the benefit of Census under its own statutory mandate, namely, Title 13, Chapter 5, United States Code.

In effect, the Criteria Agreement established and refined not only the protocols, but most importantly, the authorization to fully legitimize Census use of confidential tax data. It was implicit in this agreement that exclusively statistical use was a necessary but insufficient condition for authorized access. Instead, an explicit approval by the data provider and user was required which attested to the access authorization under the statutes of both IRS and Census, the IRS implementing regulations, and the Census-IRS Criteria Agreement's specific requirements in order to satisfy the record for a particular programmatic use. This point is worth emphasizing, as it was not enough that data provider and user agreed to the general imprimatur provided by the statutory and regulatory bases for proposed access by the user. Because the Census-RDC model was seen as at the vanguard, if not the frontier, of data access, it was especially important that the record explicitly demonstrate the data provider was convinced of the proposed statistical use's justification. This type of specific dual approval is also necessary for another unique data access model with similar high visibility disclosure risk, namely, the public-use file.

Implicit to this interagency relationship is the notion that the record of all actions taken must be able to demonstrate not only authorized intent but credibility--for some pending audience of critics. This inevitable, critical eye is known as third party scrutiny, and it is neither hypothetical nor irrelevant, instead consisting of both explicit and implicit oversight bodies such as the U.S. Congress's Government Accountability Office, the U.S. Treasury Inspector General's Office, privacy advocates, the media, and ultimately, the respondents themselves. In preparing for third party scrutiny, the record underlying data access should credibly demonstrate that the process has anticipated as many factual questions as possible and that it has also considered perceptions as well. Thus, the process needs to demonstrate consistently that it operates within not only the letter of the agreement but also its intent--so that accountability, authorization of the access granted, and purpose are never in doubt.

To address both outside perceptions and the reality of third party scrutiny, Census and IRS agreed on the importance of exceeding the literal requirement of the agreement whenever possible. For this reason, both agencies agreed that it would be a rare occasion demanding minimum adherence to predominant purpose as an acceptable criterion; that is, only over 50 percent of the access purpose. Consequently, approval on the margin would not be the rule, but the exception.

Perceptions, in conjunction with concerns about third party scrutiny, played a large role in this need for dual explicit authorization by data provider and user, especially for outside researchers engaged by a national statistical agency such as Census. Again, it was vital that access to the provider's administrative data not be construed as a type of unauthorized usage disassociated from or only loosely associated with the statistical user's mandate and mission, especially when the resulting analytical data had the potential for affecting groups of respondents. Without explicit evidence, that is, the mutual approvals of both the administrative data provider and the statistical user signifying that the specific use was authorized, third party scrutiny might raise troubling questions as to the type of confidentiality protection assured by the administrative data provider, which assumes virtually all risk with its respondent population. This issue goes to the heart of accountability in data stewardship.

One reason for the IRS-Census impasse in 2000 is that there is a fundamental and inexorable tension due to the conflicting nature of their respective mandates. Census is mandated to use administrative data to the maximum extent possible in order to reduce respondent burden and processing costs. IRS is mandated to provide confidential tax information only to the minimum extent necessary. This inherent tension imposes a sort of *de facto* equilibrium in the intersection of the agencies' confidentiality cultures, and only the strongest part of each culture is allowed relevance. It is thereby critical to protecting confidentiality, including perceptions of abuse, as both data provider and user must bargain hard for an acceptable access transaction that satisfies their respective mandates. Critical to such success is a set of clearly defined terms and processes, and the documen-

tation of subsequent actions following such a process. Equally critical is the devotion of sufficient resources to ensure the needed safeguards. Because resources are finite, so must be the amount of access whose safeguarding can be demonstrably credible. Without resource commitment to verifiable standards of protection, the clear implication is that access can approach infinite levels, suggesting both an inability and a lack of commitment to safeguard the data effectively.

Rebuilding the Relationship: Implementation of the Criteria Agreement

It was clear at the inception of the Criteria Agreement that the many new proposals of RDC outside researchers would be tied to the Census Bureau's future viability, especially its ability to keep up with the new statistical needs of decisionmakers. That is, the RDC project proposals were seen as critical to maintaining the statistical heartbeat at Census.

In fact, most of the FTI access proposals came from Census RDC's, and, initially, Census and IRS reviewed these proposals concurrently. This arrangement was soon abandoned for primarily one reason. Although it was inefficient for IRS, the administrative data provider, to spend time reviewing proposals ultimately rejected by Census, it was critical that the fundamental criterion of all tax data access, that is, a proposal's predominant purpose of benefiting Census under Title 13, Chapter 5, be demonstrated in proposals that Census, as data user, first approved. That is, the Census review process was supposed to consider not only scientific merit but also Title 13, Chapter 5, predominant purpose, while IRS review considered only the latter. Once it became clear that Census needed to take responsibility for both aspects of review (although IRS, as data provider, maintained ultimate control as data owner), the human review capital, especially regarding requirements for tax data access, could be transferred upstream from IRS to Census, and then from Census to the researcher community. Thus, the confidentiality culture needed by the data provider to assuage its third party scrutiny concerns was necessarily integrated into the data user's confidentiality culture as well as that of its researcher community. In turn, this culture colonized prospective researchers.

Outside researchers realized they had two critical interests in helping such a system succeed. First, the perpetuation of the Census-IRS arrangement allowed the researcher community access to FTI for authorized purposes, which required undertaking only proposals within scope. Second, by learning the needed culture, researchers could help increase the probability of their own proposals being approved, and even increase the number of proposals which might be possible, by theoretically and *ceteris paribus*, shortening the review process itself.

However, to counter the potential for insincere or even fraudulent researcher behavior, IRS, as administrative data provider, and Census, as data user, also conveyed three fundamental understandings to the researcher community. First, cheating on proposal purpose would inevitably be self-defeating, as it would destroy the process. Thus, implicit, if not explicit, peer-policing among the researcher community was essential to the process succeeding, and was encouraged by both Census and IRS. In fact, both agencies took pains conveying directly to the researcher community that, while it might be possible to deceive both agencies' reviews, it would be at a cost fatal to the process. Second, a post-project certification process would be necessary not only to satisfy the potential dangers of third party scrutiny by completing the authorization process, but also to help increase the knowledge capital of the proposal process itself. Third, the entire process was dynamic and was likely to be re-evaluated whenever necessary, to ensure that practice kept up with the multiple needs of decisionmakers, which included not only adequate data but also confidentiality concerns and related perceptions.

The notion of "Census benefit" may require some amplification, as it might differ from the statistical benefit required by other countries. For example, in the U.K.'s ten principles of protocol, access to confidential data is granted only "where it will [emphasis added] result in a significant statistical benefit" [5]. This type of arrangement appears to require certainty of tangible success, but it may also include a type of benefit implicitly recognized by the flexibility in the IRS-Census arrangement. That is, to reassure researchers that a fall from the "high wire" of Title 13, Chapter 5, predominant purpose attempted by ambitious projects would not necessarily

be "fatal," IRS and Census agreed that a safety net of sorts would exist for all projects, especially those that failed to meet the criteria in their proposals but made a demonstrably good faith effort to do so. However, the good faith effort of failure needed to be documented, as did that of success, so that the future proposal process could use these outcomes as a learning device for both reviewers and prospective researchers.

▶ Recognition of the Deteriorating Tradeoff

In the late 1990's, Census became concerned about potential confidentiality problems in a previously released SIPP public-use file. These had been detected through analytical techniques used by a professional intruder whom Census had engaged contractually for just such a purpose. At the January 2002 conference, in which the book, *Confidentiality Disclosure, and Data Access Theory and Practical Applications for Statistical Agencies* was showcased and released by Census, Sweeney (2001) presented some of her methods and how they might be used to reidentify survey respondents. Part of this methodology relied upon the possibility that variables in the public-use file might also be individually identifiable in other publicly available datasets [6]. In some respects at least, this event served as a type of catalyst for not only the current synthetic data approach for the SIPP/SSA/IRS public-use file, but also for re-examining disclosure risk in the Federal statistical community.

Although the success of the new Census-IRS relationship was largely predicated on a more collegial attitude, it was clear at the outset that this could not be a coequal partnership, as confidential data flowed only from the administrative data provider, IRS, to the data user, Census, and not vice versa. However, benefits did accrue. Partly as a result of the Sweeney (2001) work, IRS's own Statistics of Income Division decided to subject its public-use file, the tax model file based upon a sample of individual tax return filings, to such an examination and contracted with Sweeney's laboratory at Carnegie Mellon University for a professional intruder assessment of its confidentiality protections. In addition, because IRS approval of the synthetic data SIPP/SSA/IRS public-use file would be required (just as

the Census RDC proposals required specific approvals) before its public release, IRS was also brought in by Census early in the process as a collaborator, not just a reviewer. If the synthetic data approach is successful at Census, it will help increase the utility to researchers of nonconfidential tax data at the same time it reduces the need for access to confidential tax data, possibly even at Census RDC's where the beta testing will occur. Such a win-win outcome would benefit not only the confidentiality protection of administrative tax data but also the utility of researcher analysis for decisionmakers in both government and the private sector.

▶ The Creation of a New Program

In late 2000, as both agencies began to resolve their differences with work on the Criteria Agreement, another Census-IRS crisis was brewing. Namely, a Census request to amend the Income Tax Regulations had been submitted in order to enhance Census estimates of poverty and income for the SIPP program. The detailed earnings items sought were also deemed critical for an emerging Census flagship program, the Longitudinal Employer-Household Dynamics study, which sought, among other goals, to track more closely employment flows in the U.S. economy. Both requests initially encountered opposition, but the justification for each emphasized the minimal need for FTI in these mandated uses. Eventually, the regulations were approved in February 2001, and immediately after work began on the SIPP/SSA/IRS PUF. It is ironic, but not coincidental, that the regulations were approved so soon after the Criteria Agreement's implementation in September 2000. That is, the process, which had prepared both agencies for the Criteria Agreement, also galvanized them for purposes of these new proposed uses of FTI by making them focus on the criteria within the agreement as well as the protocols and process which would govern such access. It is also not a coincidence that one of the goals set forth in the Census justifications for the IRS regulations amendment was the production of a SIPP public-use file, which was to include associated administrative data from SSA and IRS. The utility of this product was clearly seen as not only a predominant Title 13, Chapter 5, benefit for Census, but also a confidentiality boon for administrative tax data in general. However,

without the items requested for regulation amendment, both SIPP and the potential robustness of the proposed LEHD program would have been seriously weakened. In fact, had the regulations items not been approved, it is likely that the LEHD program as it is known today would not exist. Had the Criteria Agreement, and even its early implementation not been developed as the SIPP and LEHD requests were prepared and later considered, it is possible, if not probable, that neither would have been approved.

▶ Lessons and Recommendations

One consequence of the modern Census-IRS relationship is that the Criteria Agreement process undergone to protect confidentiality also laid the groundwork for further legitimate access meeting these requirements, for example, the SIPP/SSA/IRS public-use file and the LEHD program described above.

Another lesson is that the record can probably be satisfied for posterity's perceptions of the past by ensuring that clear and sufficient documentation exists to explain those past intentions and actions.

The final lesson learned is that agencies must look outside themselves for the talents and skill sets needed to help them protect confidentiality and meet the needs for which confidential data are collected in the first place. In a time of dwindling budgets and competing priorities, such considerations are no longer options--they are imperatives.

In sum, one of the most important services that government agencies can perform is communicating to decisionmakers the need to learn the above lessons. If avenues are closed to such pursuits in the future, decisionmakers need to understand not only that their decisions will be based upon inadequate information--including its quality--but also that the imprimatur for intruding on the privacy of respondents-citizens will not exist. That is, the mandate for data collection will cease, but so will the ability of decisionmakers to lead and govern.

▶ Footnotes

* The author presented this paper in June 2004 at the Conference on Privacy in Statistical Databases, in Barcelona, Spain.

[1] P. 6, Working Paper No. 11, *Contexts for the Development of a Data Access and Confidentiality Protocol for UK National Statistics*, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003.

[2] Confidential data are any identifiable data whose public release is unauthorized. The removal of identifier information, such as name, address, and identification numbers, is a necessary but insufficient condition to render such data anonymous or unidentifiable.

[3] An abbreviated course in IRS master files might summarize data maintained on these systems (whether individual or business master file) as being one of two types: entity information or tax module information. Entity information refers to information used to identify and locate a taxpayer such as Taxpayer Identification Number (Social Security Number--SSN, Employer Identification Number--EIN), Name, Address, and perhaps Industry Classification Code (NAICS or SIC-based) for a business. Everything else is tax module information.

[4] For example, see Working Paper No. 10, *Research Data Centres of Official Statistics*, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003.

[5] P. 7, Working Paper No. 11, *Contexts for the Development of a Data Access and Confidentiality Protocol for UK National Statistics*, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003.

[6] Latanya Sweeney (2001), "Information Explosion," in *Confidentiality, Disclosure, and Data Access Theory and Practical Applications for Statistical Agencies*, North Holland.

5

Multivariate Analysis

Chen

Some New Tables for Upper Probability Points of the Largest Root of a Determinantal Equation with Seven and Eight Roots

William W. Chen, Internal Revenue Service

We revisit the Fisher-Girshick-Hsu-Roy distribution (1939), which has interested statisticians for more than six decades. Instead of using K.C.S. Pillai's method of neglecting higher order terms of the cumulative distribution function (C.D.F.) of the largest root to approximate the percentage points, we simply keep the whole C.D.F. and apply its natural non-decreasing property to calculate the exact probabilities. At the duplicated percentage points, we found our computed percentage points to be consistent with existing tables. However, our tabulations have greatly extended the existing tables.

In Chen (2002), we were concerned with the distribution of the largest characteristic roots in multivariate analysis when there are two to six roots. Now, we will extend the size to seven and eight roots.Fisher-Girshick-Hsu-Roy(1939) discuss this in detail and present the joint probability density function in general. This well-known distribution depends on the number of characteristic roots and two parameters m and n, which are defined differently for various situations, as described by Pillai (1955). The upper percentage points of the distribution are commonly used in three different multivariate hypothesis tests: tests of equality of the variance-covariance matrices of two p-variate normal populations, tests of equality of the p-dimensional mean vectors for k p-variate normal populations, and tests of independence between a p-set and a q-set of variates in a (p+q)-variate normal population. When the null hypotheses are true, these three proposed tests depend only on the characteristic roots of matrices using observed samples. The problem can be stated as follows: using a random sample from the multivariate normal population, we will compute the characteristic roots from a sum of product matrices of this sample. We will then compare the largest characteristic root of the matrices with the percentage points tabulated in this paper to determine whether or not the null hypothesis is rejected at a certain probability confidence.

There are already many published tables that focus on upper percentage point tabulations or chart the various sizes of roots. The best-known contributor in this area is Pillai, who gave general rules for finding the C.D.F. of the largest root and tabulated upper percentage points of 95% and 99% for various sizes of roots. Other contributors, including Nanda (1948, 1951), Foster and Rees (1957, 1958), and Heck (1960) will be discussed in more detail in section 2. Section 3 contains the joint distribution of s non-null characteristic roots of a matrix in general form and the C.D.F. of the seven and eight largest characteristic roots. The algorithm used to create the tables in this paper is the same as in Chen (2003), and we will not repeat it. Also, we will ignore the discussion of precision of the results.

▶ Cumulative Function and Historical Work

The joint distribution of s non-null characteristic roots of a matrix in multivariate distribution was first given by Fisher-Girshick-Hsu-Roy (1939) and can be expressed in the form of (3.1). We further extended the distribution of the largest characteristic root to seven and eight roots. Even though the form of the joint density function is known, it is not easy to write out the C.D.F. of the largest characteristic root to seven roots. To solve this problem, two methods can be used to find the C.D.F. more easily. Pillai (1965) suggests that the C.D.F. of the largest characteristic root could be presented in determinant form of incomplete beta functions. Since the numerical integration of each of the s factorial multiple integrals when the determinant is expanded is difficult, he suggests an alternative reduction formula that gives exact expressions for the C.D.F. of the largest root in terms of incomplete beta functions or functions of incomplete beta functions for various values of s. An alternative method suggested by Nanda (1948) yields the same results. He started with the Vandermonde determinant and expanded it in minors of a row, then repeated applied

integration by part to find the C.D.F. of the largest characteristic root. In this paper, we use the Pillai notation and present the case with seven roots in equation (3.2). Following this C.D.F. and the algorithm previously used, we tabulate the upper percentage points.

Here, it is useful to review some of the published tables and reasons to extend the tables. Pillai (1956a, 1959) published tables that focus only on two percentage points: 95% and 99% for s =2,6, m = 0(1)4, and n varying from 5 to 1000. Foster and Rees (1957) tabulated the upper percentage points 80%, 85%, 90%, 95%, and 99% of the largest root for s=2, m=-0.5, 0(1)9, n=1(1)19 (5)49,59,79. Foster(1957, 1958) further extended these tables for values of s=3 and 4. Heck (1960) has given some charts of upper 95%, 97.5%, and 99% points for s=2(1)5, m=-0.5, 0(1)10, and n greater than 4. These table values can be applied to our statistical analysis with some standard textbooks as references. For example, recently, Rencher included the percentage point 0.950 in two textbooks Rencher (1998 and 2002).

Without a modern computer, it is difficult and tedious to compute the whole C.D.F. (3.2) at each percentage point. Therefore, deleting higher order terms and retaining a few lower order terms to approximate the roots is a reasonable solution. However, this approach involves intolerable error at lower percentage points, such as 80%, 82.5%, 85%, 87.5%, 90%, or 92.5%. These percentage points are usually ignored due to the difficulty of their computation, and not due to their lack of use. Traditional methods treat intermediate percentage points by interpolation, but without, for example, 85% or 90% percentage points, it is difficult to interpolate 87.5%. In recent years, computers have gradually matured in memory, speed, and flexibility in usage, which has greatly changed the methods by which we study statistics. In this study, we use one of the most basic properties of C.D.F. and revisit this most important distribution. As many percentage points as are needed in one computer run are included: these are 0.80, 0.825, 0.850, 0.875, 0.890, 0.900, 0.910(0.005), 0.995. Different authors have selected different m and n parameter values, but we selected these parameters such that all existing table values are included. For the parameters m=0(1)10 and

n=3(1)20(2)30(5)80(10), 150, 200(100)1000, our table provides the percentage points and probabilities while avoiding the interpolation problem.

▶ The Distribution Function of Seven Characteristic Roots

Suppose $x = \{x_{ij}\}$ and $x^* = \{x^*_{ij}\}$ are two p-variate random matrices with n_1 and n_2 the degree of freedom, respectively. Assume the two multivariate populations have the same covariance matrix: for example, $S_1 = xx^T / n_1$ and $S_2 = x^* x^{*T} / n_2$. When the null hypothesis is true, both S_1 and S_2 are independent estimators of the unknown but equal covariance matrices. The joint distribution of the roots of the determinantal equation $|A - \theta(A + B)| = 0$ where $A = n_1 S_1$ and $B = n_2 S_2$ has been given by Fisher-Girshick-Hsu-Roy(1939) and can be written as :

$$f(\theta_1, ... \theta_s) = C(s, m, n) \prod_{i=1}^{s} \theta_i^m (1 - \theta_i)^n \prod_{i>j} (\theta_i - \theta_j)$$

$$(0 < \theta_1 \leq .... \leq \theta_s \leq 1),$$

where

$$C(s, m, n) = \frac{\pi^{s/2} \prod_{i=1}^{s} \Gamma(\frac{2m + 2n + s + i + 2}{2})}{\prod_{i=1}^{s} \Gamma(\frac{2m + i + 1}{2}) \Gamma(\frac{2n + i + 1}{2}) \Gamma(\frac{i}{2})}. \quad (3.1)$$

and the parameters m and n are defined differently for various situations as described by Pillai (1955, pp. 118). Following Pillai's method, the cumulative distribution function of the largest characteristic root for seven and eight is given below:

When s = 7, the C.D.F. of the largest characteristic root is:

$$Pr(\theta_7 \leq x) = \frac{C(7, m, n)}{m + n + 7}[-IO(x, m + 6, n + 1) * v\_0654321x(x, m, n)$$

$$- 2I(x, 2m + 6, 2n + 1) * v\_054321x(x, m + 1, n) + 2I(x, 2m + 7, 2n + 1)$$

$$* v\_05432x(x, m, n) - 2I(x, 2m + 8, 2n + 1) * v\_05431x(x, m, n)$$

$$+ 2I(x, 2m + 9, 2n + 1) * v\_05421x(x, m, n) - 2I(x, 2m + 10, 2n + 1)$$

$$* v\_05321x(x, m, n) + 2I(x, 2m + 11, 2n + 1) * v\_054321x(x, m, n)]$$

$$\Pr(\theta_8 \le x) = \frac{C(8, m, n)}{m + n + 8}[-IO(x, m + 7, n + 1) * v\_07654321x(x, m, n)$$
$$+ 2I(x, 2m + 7, 2n + 1) * v\_0654321x(x, m + 1, n) - 2I(x, 2m + 8, 2n + 1)$$
$$* v\_065432x(x, m, n) + 2I(x, 2m + 9, 2n + 1) * v\_065431x(x, m, n)$$
$$- 2I(x, 2m + 10, 2n + 1) * v\_065421x(x, m, n) + 2I(x, 2m + 11, 2n + 1)$$
$$* v\_065321x(x, m, n) - 2I(x, 2m + 12, 2n + 1) * v\_064321x(x, m, n)$$
$$+ 2I(x, 2m + 13, 2n + 1) * v\_0654321x(x, m, n)]$$

$$C(7, m, n) = \frac{\Gamma(2m+2n+14)*\Gamma(2m+2n+12)*\Gamma(2m+2n+10)*\Gamma(2m+2n+8)}{46080*\Gamma(2m+6)*\Gamma(2n+6)*\Gamma(2m+4)*\Gamma(2n+4)*\Gamma(2m+2)*\Gamma(2n+2)*\Gamma(m+4)*\Gamma(n+4)}$$

$$C(8, m, n) = \frac{\Gamma(2m+2n+17)*\Gamma(2m+2n+15)*\Gamma(2m+2n+13)*\Gamma(2m+2n+11)}{8847360*\Gamma(2m+8)*\Gamma(2n+8)*\Gamma(2m+6)*\Gamma(2n+6)*\Gamma(2m+4)*\Gamma(2n+4)*\Gamma(2m+2)*\Gamma(2n+2)}$$

▶ References

Chen, W.W.S.(2002), Some new tables of the largest root of a matrix in multivariate analysis : a computer approach from 2 to 6, Proceeding of Statistical Computing Section, pp. 503-508.

Chen, W.W.S.(2003), Table for upper percentage points of the largest root of a determinantal equation with five roots, February, http://Inter-Stat.stat.vt.edu/InterStat/

Fisher, R.A.(1939), The sampling distribution of some statistics obtained from non-linear equations, Ann. Eugenics, Vol. 9, pp. 238-249.

Foster, F.G. and Rees, D.H.(1957), Upper percentage points of the generalized beta distribution, I, Biometrika, Vol. 44, pp. 237-247.

Foster, F.G.(1957), Upper percentage points of the generalized beta distribution, II, Biometrika, Vol. 44, pp. 441-453.

Foster, F.G.(1958), Upper percentage points of the generalized beta distribution, III, Biometrika, Vol. 45, pp. 492-503.

Girshick, M.A.(1939), On the sampling theory of the roots of determinantal equations, Ann. Math. Stat., Vol. 10, pp. 203-224.

Heck, D.L.(1960), Charts of some upper percentage points of the distribution of the largest characteristic root, Ann. Math. Statist., 31, pp. 625-642.

Hsu, P.L. (1939), On the distribution of roots of certain determinantal equations, Ann. Eugenics, Vol. 3, pp. 250-258.

Nanda, D.N.(1948), Distribution of a root of a determinantal equation, Ann. Math. Stat., Vol. 19, pp. 47-57.

Nanda, D.N.(1951), Probability distribution tables of the largest root of a determinantal equation with two roots, J. Indian Soc. Of Agricultural Stat., Vol. 3, pp. 175-177.

Pillai, K.C.S.(1955), Some new test criteria in multivariate analysis. Ann. Math. Statist. 26, pp. 117-121.

Pillai, K.C.S.(1956a), On the distribution of the largest or smallest root of a matrix in multivariate analysis, Biometrika, Vol. 43, pp. 122-127.

Pillai, K.C.S.(1956b), Some results useful in multivariate analysis, Ann. Math. Statist. Vol. 27, pp. 1106-1114.

Pillai, K.C.S. and Bantegui, C.G.(1959), On the distribution of the largest of Six roots of a matrix in multivariate analysis, Biometrika, Vol. 46, pp. 237-240.

Pillai, K.C.S. (1965), On the distribution of the largest characteristic root of a matrix in multivariate analysis, Biometrika 52, pp. 405-414.

Roy, S.N.(1939), p-statistics or some generalizations in analysis of variance appropriate to multivariate problems, Sankhya, Vol. 4, pp. 381-396.

Rencher, A.C.(1998), Multivariate statistical inference and application, John Wiley & Sons, Inc.

Rencher, A.C.(2002), Methods of multivariate analysis, John Wiley & Sons, Inc.

Upper percentage points of .900 of theta(p,m,n),
the largest eigenvalue of |B-theta(W+B)|=0,when s=7

m

n	0	1	2	3	4	5	6	7	8	9	10
3	.9040	.9188	.9295	.9378	.9442	.9495	.9538	.9576	.9608	.9630	.9644
4	.8650	.8842	.8986	.9097	.9186	.9259	.9320	.9371	.9415	.9453	.9490
5	.8266	.8497	.8671	.8809	.8920	.9012	.9090	.9156	.9212	.9261	.9307
6	.7899	.8160	.8362	.8522	.8653	.8763	.8855	.8935	.9004	.9064	.9116
7	.7552	.7838	.8062	.8242	.8391	.8515	.8622	.8714	.8794	.8865	.8927
8	.7226	.7533	.7774	.7971	.8135	.8273	.8392	.8495	.8586	.8665	.8737
9	.6923	.7244	.7501	.7711	.7888	.8038	.8168	.8281	.8380	.8469	.8548
10	.6639	.6973	.7241	.7463	.7650	.7811	.7950	.8072	.8180	.8276	.8363
11	.6376	.6717	.6995	.7226	.7423	.7592	.7740	.7869	.7985	.8088	.8181
12	.6130	.6478	.6763	.7002	.7206	.7382	.7537	.7674	.7796	.7905	.8004
13	.5901	.6253	.6543	.6788	.6999	.7182	.7342	.7485	.7613	.7727	.7832
14	.5687	.6042	.6336	.6586	.6801	.6989	.7155	.7303	.7436	.7556	.7664
15	.5487	.5843	.6140	.6394	.6613	.6806	.6976	.7128	.7266	.7390	.7503
16	.5300	.5656	.5955	.6211	.6434	.6630	.6804	.6960	.7101	.7229	.7346
17	.5124	.5480	.5780	.6038	.6263	.6462	.6640	.6799	.6943	.7074	.7194
18	.4959	.5314	.5614	.5873	.6100	.6302	.6482	.6644	.6791	.6925	.7048
19	.4805	.5157	.5457	.5717	.5945	.6148	.6330	.6495	.6644	.6781	.6906
20	.4659	.5009	.5308	.5568	.5797	.6001	.6185	.6351	.6503	.6642	.6769
22	.4391	.4736	.5032	.5291	.5520	.5726	.5912	.6081	.6236	.6378	.6509
24	.4152	.4490	.4782	.5039	.5268	.5474	.5661	.5832	.5988	.6133	.6267
26	.3937	.4267	.4554	.4809	.5036	.5242	.5429	.5600	.5758	.5904	.6040
28	.3743	.4065	.4347	.4598	.4823	.5027	.5214	.5386	.5544	.5691	.5828
30	.3567	.3881	.4158	.4404	.4627	.4829	.5015	.5186	.5344	.5492	.5629
35	.3190	.3486	.3748	.3984	.4198	.4394	.4576	.4744	.4901	.5047	.5184
40	.2885	.3162	.3410	.3635	.3840	.4030	.4205	.4369	.4523	.4667	.4802
45	.2632	.2894	.3128	.3342	.3538	.3720	.3889	.4048	.4197	.4338	.4471
50	.2420	.2666	.2888	.3092	.3279	.3454	.3617	.3770	.3915	.4052	.4181
55	.2240	.2472	.2683	.2876	.3055	.3223	.3380	.3528	.3668	.3800	.3926
60	.2084	.2304	.2504	.2688	.2860	.3020	.3171	.3314	.3449	.3578	.3700
65	.1949	.2157	.2348	.2524	.2688	.2842	.2987	.3124	.3255	.3379	.3498
70	.1830	.2028	.2210	.2378	.2535	.2683	.2822	.2955	.3081	.3202	.3317
75	.1725	.1914	.2087	.2248	.2399	.2541	.2675	.2803	.2925	.3042	.3154
80	.1631	.1811	.1977	.2131	.2276	.2413	.2542	.2666	.2784	.2897	.3005
90	.1470	.1636	.1788	.1931	.2065	.2192	.2313	.2428	.2538	.2645	.2747
100	.1339	.1492	.1633	.1765	.1889	.2008	.2121	.2229	.2333	.2432	.2529
110	.1229	.1371	.1502	.1625	.1741	.1852	.1958	.2060	.2158	.2252	.2343
120	.1136	.1268	.1390	.1506	.1615	.1719	.1819	.1915	.2007	.2096	.2182
130	.1056	.1179	.1294	.1403	.1506	.1604	.1698	.1788	.1876	.1960	.2042
140	.0986	.1102	.1211	.1313	.1410	.1503	.1592	.1678	.1761	.1841	.1919
150	.0925	.1035	.1137	.1234	.1326	.1414	.1498	.1580	.1659	.1735	.1810
200	.0706	.0792	.0872	.0948	.1021	.1091	.1158	.1223	.1287	.1348	.1408
300	.0480	.0539	.0595	.0648	.0699	.0749	.0796	.0843	.0888	.0932	.0975
400	.0363	.0409	.0451	.0492	.0532	.0570	.0607	.0643	.0678	.0712	.0746
500	.0292	.0329	.0364	.0397	.0429	.0460	.0490	.0519	.0548	.0576	.0604
600	.0244	.0275	.0305	.0332	.0359	.0386	.0411	.0436	.0460	.0484	.0507
700	.0210	.0237	.0262	.0286	.0309	.0332	.0354	.0375	.0396	.0417	.0437
800	.0184	.0208	.0230	.0251	.0271	.0291	.0311	.0330	.0348	.0367	.0384
900	.0164	.0185	.0205	.0224	.0242	.0260	.0277	.0294	.0311	.0327	.0343
1000	.0148	.0167	.0184	.0202	.0218	.0234	.0250	.0265	.0280	.0295	.0309

**Upper percentage points of 0.900 of theta(p,m,n),
the largest eigenvalue of |B-theta(W+B)|=0,when s=8**

m

n	0	1	2	3	4	5	6	7	8	9
5	0.8517	0.8702	0.8845	0.8959	0.9052	0.9124	0.9225	0.9563	0.9570	0.9882
6	0.8183	0.8396	0.8563	0.8698	0.8808	0.8897	0.8978	0.9014	0.9096	0.9727
7	0.7862	0.8099	0.8286	0.8439	0.8566	0.8667	0.8781	0.8788	0.9019	0.9532
8	0.7558	0.7814	0.8019	0.8187	0.8328	0.8446	0.8559	0.8763	0.8994	0.9279
9	0.7270	0.7542	0.7761	0.7943	0.8096	0.8226	0.8347	0.8428	0.8664	0.8882
10	0.7000	0.7283	0.7515	0.7708	0.7872	0.8012	0.8139	0.8237	0.8385	0.8695
11	0.6744	0.7038	0.7280	0.7482	0.7656	0.7806	0.7937	0.8070	0.8163	0.8390
12	0.6505	0.6807	0.7056	0.7267	0.7448	0.7606	0.7746	0.7874	0.8013	0.8057
13	0.6280	0.6587	0.6843	0.7061	0.7248	0.7412	0.7555	0.7688	0.7826	0.7865
14	0.6069	0.6381	0.6641	0.6864	0.7057	0.7227	0.7376	0.7513	0.7665	0.7798
15	0.5870	0.6185	0.6449	0.6676	0.6874	0.7048	0.7202	0.7351	0.7468	0.7731
16	0.5683	0.5999	0.6267	0.6498	0.6699	0.6877	0.7036	0.7182	0.7341	0.7421
17	0.5507	0.5824	0.6094	0.6327	0.6532	0.6713	0.6874	0.7028	0.7153	0.7347
18	0.5340	0.5658	0.5929	0.6164	0.6371	0.6555	0.6720	0.6866	0.6981	0.7066
19	0.5183	0.5500	0.5772	0.6009	0.6218	0.6404	0.6572	0.6719	0.6841	0.6935
20	0.5035	0.5351	0.5623	0.5861	0.6071	0.6259	0.6428	0.6579	0.6713	0.6799
22	0.4761	0.5074	0.5345	0.5583	0.5796	0.5986	0.6159	0.6313	0.6450	0.6560
24	0.4514	0.4823	0.5092	0.5330	0.5542	0.5734	0.5908	0.6067	0.6205	0.6319
26	0.4291	0.4595	0.4861	0.5097	0.5309	0.5501	0.5676	0.5837	0.5982	0.6101
28	0.4089	0.4387	0.4649	0.4883	0.5094	0.5286	0.5461	0.5622	0.5768	0.5897
30	0.3904	0.4196	0.4454	0.4686	0.4895	0.5085	0.5260	0.5422	0.5569	0.5699
35	0.3507	0.3784	0.4031	0.4254	0.4457	0.4643	0.4816	0.4975	0.5123	0.5258
40	0.3182	0.3444	0.3679	0.3893	0.4089	0.4270	0.4438	0.4595	0.4741	0.4874
45	0.2912	0.3160	0.3383	0.3588	0.3776	0.3951	0.4114	0.4267	0.4410	0.4542
50	0.2684	0.2918	0.3131	0.3327	0.3507	0.3676	0.3833	0.3982	0.4121	0.4250
55	0.2488	0.2711	0.2913	0.3100	0.3274	0.3436	0.3588	0.3731	0.3866	0.3994
60	0.2319	0.2531	0.2724	0.2903	0.3069	0.3225	0.3371	0.3510	0.3641	0.3766
65	0.2172	0.2373	0.2557	0.2728	0.2888	0.3038	0.3179	0.3313	0.3441	0.3562
70	0.2042	0.2234	0.2410	0.2574	0.2727	0.2871	0.3008	0.3137	0.3261	0.3378
75	0.1926	0.2110	0.2278	0.2435	0.2583	0.2722	0.2854	0.2979	0.3098	0.3212
80	0.1824	0.1999	0.2160	0.2311	0.2453	0.2587	0.2715	0.2836	0.2952	0.3062
90	0.1647	0.1808	0.1958	0.2097	0.2229	0.2354	0.2473	0.2587	0.2696	0.2800
100	0.1502	0.1651	0.1790	0.1920	0.2043	0.2159	0.2271	0.2378	0.2480	0.2578
110	0.1380	0.1519	0.1648	0.1770	0.1885	0.1994	0.2099	0.2200	0.2296	0.2389
120	0.1277	0.1406	0.1527	0.1641	0.1749	0.1852	0.1951	0.2046	0.2138	0.2226
130	0.1188	0.1309	0.1423	0.1530	0.1632	0.1730	0.1823	0.1913	0.2000	0.2084
140	0.1110	0.1225	0.1332	0.1433	0.1530	0.1622	0.1711	0.1796	0.1879	0.1958
150	0.1042	0.1150	0.1252	0.1348	0.1439	0.1527	0.1611	0.1692	0.1771	0.1847
200	0.0798	0.0882	0.0962	0.1038	0.1111	0.1181	0.1248	0.1313	0.1377	0.1438
300	0.0543	0.0602	0.0658	0.0711	0.0762	0.0812	0.0860	0.0907	0.0952	0.0997
400	0.0412	0.0457	0.0500	0.0541	0.0580	0.0619	0.0656	0.0692	0.0728	0.0762
500	0.0331	0.0368	0.0403	0.0436	0.0469	0.0500	0.0530	0.0560	0.0589	0.0618
600	0.0277	0.0308	0.0338	0.0366	0.0393	0.0419	0.0445	0.0470	0.0495	0.0519
700	0.0238	0.0265	0.0290	0.0315	0.0338	0.0361	0.0383	0.0405	0.0426	0.0447
800	0.0209	0.0232	0.0255	0.0276	0.0297	0.0317	0.0337	0.0356	0.0375	0.0393
900	0.0186	0.0207	0.0227	0.0246	0.0265	0.0283	0.0300	0.0317	0.0334	0.0351
1000	0.0168	0.0187	0.0205	0.0222	0.0239	0.0255	0.0271	0.0286	0.0301	0.0316

Index
of IRS Methodology Reports
on Statistical Uses of Administrative Records

Special Studies in Federal Tax Statisics--2003
Selected papers given primarily at the 2003 Annual Meetings of the American Statistcal Association in San Francisco, CA. The volume is divided into four major sections. It begins with four papers presented in the same session under the topic, "Are the Rich Getting Richer and the Poor Getting Poorer?" Section 2 includes a paper on survey methods. Section 3 presents five papers on new developments in tax statistics and administrative records. Finally, Section 4 contains a paper on survey nonresponse and imputation.

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