COMING SOON: TAXPAYER DATA CLASSIFIED BY OCCUPATION

Peter Sailer, Internal Revenue Service
Harriet Orcutt, Social Security Administration
Phil Clark, Internal Revenue Service

In 1916, the fourth year of the U.S. Federal income tax, the tax return (Form 1040) was revised for the first time. Among the revisions, a line was added which invited taxpayers to disclose their occupation to the Internal Revenue Service (IRS). This same year (1916) also marked the beginning of a series of detailed statistical reports on the Federal income tax, entitled Statistics of Income, and the 1916 edition contained several tables which classified taxpayers by occupation. Given the $3,000 minimum net income requirement for filing a return, the occupational groupings chosen concentrated on the high-income area. A total of 36 categories were chosen, most of which were in the professional and business areas (all "labor, skilled and unskilled" was combined into a single category). A copy of one of these tables is reproduced as Table 4 at the end of this paper.

For most of the following 63 years, the occupation question remained on the tax return. In 1933 and 1934, a checklist of eight occupational categories was provided for the taxpayer. In 1936, the single line entry reappeared. However, to the authors' knowledge, IRS never again attempted to use the occupation entry for statistical purposes until the late 1960's, when a number of small-scale pilot studies were undertaken to see whether it was possible to put the taxpayers' entries in some form which could be used for research purposes. (See references [1] through [4] for reports on these pilot studies.) These studies indicated that anywhere from 50 to 88 percent of the tax returns could be coded for occupation, depending on the amount of effort expended on each return. A linked IRS/Current Population Survey (CPS) file created for 1963 indicated a reasonably good correspondence between the codes assigned in the CPS and those derived from the roughly 60 percent of the tax returns which could be readily coded for occupation [5].

Examining the possibilities for creating a Linked Administrative Statistical Sample (LASS) provided a renewed impetus to studying ways of using the tax return to obtain occupational data; the tax return would be the major source of occupation information in that sample. In addition, the establishment of a new "official" occupational coding system, the Standard Occupational Classification (SOC), made it imperative that we test whether this particular system was suitable for coding income tax returns. Therefore, using a previously selected random sample of returns for tax year 1976, we set out to answer the following questions:

1) To what extent can detailed occupation codes be determined from the taxpayers' entries? Obviously, the tax return would not be a good source for occupation data if detailed occupation codes could only be determined for a small percentage of the sample.

2) What is the amount of resources required to create occupation codes?

3) To what extent does the information left by the taxpayer and the codes determined from this and supplemental information accurately reflect individuals' occupations?

If the tax return proved to be a useful source of occupation data, and our methodology proved acceptable, a full-scale project would be undertaken for Forms 1040 filed in 1980.

Codability of Taxpayers' Entries by Themselves

For the pilot study, we obtained a random sample of 6,700 tax returns filed in 1977. The taxpayers' occupation entries were keypunched and sorted in alphabetical order, with identical entries combined. A total of 9,680 entries were transcribed; coding took place in four stages: (1) direct look-up, (2) modified direct look-up, (3) manual search with professional judgment, and (4) use of supplemental information.

The first stage involved clerks simply trying to match a taxpayer's entry exactly with an entry in the SOC index.[1] Only 10% of the entries could be coded in this manner, so we asked the clerks to use some judgment--to substitute synonyms, correct misspellings, change word orders, and drop adjectives. After this second coding stage, roughly 21% of the file was coded. Unfortunately, on closer examination we found that 3% of the file was obviously coded incorrectly. Thus about 18% of the file was now "correctly" coded.

The major problem with this approach turned out not to be the quality of work done by clerks, but the cryptic and often confusing nature of the SOC index. For instance, the SOC index does not list the actual SOC group titles (which appear in the manual) in many cases. Thus, such entries as "author," "brickmason," or "legislator," which are SOC unit group titles, do not appear in the SOC index. Another example of the index's confusing nature is the title "engineer." It appears only once by itself in the SOC index, followed by six variations with qualifying words (engineer-custodial, engineer--soil, engineer--studio, etc.). A clerk coding the entry "engineer" would use the code associated with the single word in the index. However, by looking up that code in the manual, one finds that it refers specifically to marine engineers. While overcoming these difficulties is certainly time-consuming, as coders become more familiar with the index

and manual, progress can be achieved more rapidly.

The third stage of coding involved a degree of subjective decision-making so as to classify all the occupation entries remaining to the highest level possible. Using the SOC books, as well as other occupational manuals, dictionaries, and thesaureses, the authors set out to complete the coding of the file. When we were finished, we had managed to assign some code to 80% of the entries--57% were fully coded (all applicable digits had been assigned), and 23% were partially coded (at least the first digit of the code could be assigned). Unfortunately, over half of the partially coded entries had only the first digit assigned. A full breakdown of codable entries is shown in Table 1.

Table 1
Results of 1976 Pilot Study

Level of coding without industry	Percent of taxpayers
Grand Total	100
Uncodable, total	20
No entry	6
Unusable entry	14
Codable, total	80
Fully codable	57
Direct look-up	9
Modified direct look-up	9
Incorrect direct look-up	3
Professional judgment	36
Partially codable, total	22
Last digit uncodable	3
Last two digits uncodable	6
Last three digits uncodable	13

Use of Supplemental Information to Assign Codes

At this point in the study, we reached two conclusions. First, we wanted to avoid the need for searching through the SOC manual more than once for any given taxpayer entry, especially if this project were to be undertaken on an annual basis. Second, we realized that we needed more information in order to completely code the file. Approaching the first problem, we decided that the best strategy would be to create a computerized dictionary from the pilot study sample. In this dictionary, the taxpayer entries, including all misspellings, abbreviations, etc., would constitute the "words," while the codes which had been found to correspond to the entries would be the "definitions." This way, entries such as college teacher, university professor and all the various misspellings, abbreviations, word combinations, and descriptions of teaching at the university level would be associated with the same code. Whenever any of these entries appear in a future sample, the correct occupation code will be assigned.

To handle the second problem--finding supplemental information in order to complete the partially coded entries--we felt that knowing the industry of the taxpayer's employer would help us differentiate between occupations with similar names but different codes. For example, the entry "senior make-up" might refer to a cosmetologist or a typesetter; the employer's industry would tell us which. We expect that knowing the taxpayer's industry will enable us to code 90% of the file. (See Table 2.)

Table 2
Expected Results of
Coding with Industry

Level of coding with industry	Percent of taxpayers
Grand Total	100
Uncodable	10
Codable	90
Fully codable	89
Partially codable	1

Having anticipated the value of the taxpayer's industry of employment, we had included the employer identification number (EIN) of each taxpayer's employer (whenever available from the Form W-2 attached to the tax return) in the pilot study. By matching the EIN's to the Social Security Administration's (SSA) employer file, we will get an industry code which will be associated with each occupation entry that cannot be fully coded by itself. An attempt is now being made to assign occupation codes to these entries, creating new "words" in our dictionary, each of which would consist of a job title and an industry code. This fourth stage of the pilot study should be completed early in 1981.

Validating the Pilot Study

Plans to evaluate the results of the study involve comparing our occupational distribution with that of other studies. Unfortunately, there are no statistical data currently available based on the SOC coding scheme, and the major groupings of the SOC system are not comparable to those of other structures. However, by reclassifying data from other studies into the SOC scheme at the broadest possible level (using only the first digit), a comparison table can be produced. Table 3 compares the preliminary results of the pilot study with the results obtained in a previous IRS pilot study [1] and with published data from the Current Population Survey [6]. Given the precursory nature of our data (only 80% of the returns coded, with no industry codes used), we are satisfied with the result of this rough test of comparability.

Table 3

Percentage distribution of taxpayers by occupational division: Results of the 1976 pilot study compared to data from the Current Population Survey and from an earlier pilot study for tax year 1973.

SOC Code (First digit)	SOC Description	1976 Pilot Study	1978 CPS	1973 Pilot Study
All	Total	100.0	100.0	100.0
1, 2, 3	Professionals and managers (except farm)	27.9	25.1	27.7
5	Farmers and service workers	13.5	16.6	12.8
6	Craft workers and transportation equipment operatives	13.7	16.9	15.2
7	Operatives (except transportation)	12.8	11.5	12.2
4, 8	Other occupations	32.1	29.9	32.1

Note: This table excludes taxpayers who are not in the labor force, such as students, investors and housewives.

At present, the Bureau of Labor Statistics (BLS) is producing a cross-classification of all major occupational coding systems. We hope that this will not only enable us to make comparisons at a greater level of detail, but, since it will enable clerks to research occupation titles in the Census and Dictionary of Occupational Titles 2/ manuals and convert the results to the SOC system, it will assist clerical personnel in their efforts to code the occupation titles in the 1979 Statistics of Income (SOI) sample.

Future Plans

Assuming that it proves to be possible to produce accurate occupational data from tax returns, we plan to undertake a full-scale effort to code the IRS Statistics of Income file for 1979 (about 190,000 returns). In brief, we plan to proceed as follows:

1) The occupation entry, limited to 20 characters, as entered by the taxpayer will be edited and transcribed from all tax returns in the SOI sample.

2) Using the taxpayer's social security number, a match will be made to a tape containing W-2 information. The EIN's (and the establishment number for those employers with more than one establishments) of each taxpayer's employers will be read into the SOI file.

3) The EIN's and establishment numbers will be sent to SSA, and SSA will supply IRS with the industry code(s) for each taxpayer.

4) After the industry codes are merged into the SOI file, each occupation entry will be matched against the dictionary established during the pilot study and SOC codes entered for matched returns.

5) Return records which do not match against the dictionary will be read out for clerical review. If the clerks, using the SOC manual and the BLS cross-classification, discover obvious entries which are missing from the dictionary, they will create these additions to the dictionary.

6) After the dictionary has been expanded, a second match will be performed. Returns not matched this time will be read out for professional review, and the dictionary will be further expanded.

The occupation - coded SOI file will represent a rich data base which will serve multiple research needs. The following are some of the uses currently under exploration for this file:

1) Roughly 46,000 taxpayers who are in SSA's Continuous Work History Sample (CWHS) have been included in the SOI sample. Thus, once the file is coded, it will become possible to merge occupation data with the other demographic information available from the Social Security Administration. Since the CWHS is a longitudinal sample, repetition of this study in future years will allow comparisons between occupation and mortality and morbidity data.

2) If funds became available, IRS could produce a supplemental report in the SOI series, showing income and tax information classified by occupation and industry. Because of the match to the W-2 file detailed above, the 1979 SOI file will not only contain industry codes, but also separate amounts for husbands' and wives' salaries and

wages, enabling us to further classify much of the data by the sex of the taxpayer.

3) The computerized dictionary created from the SOI sample could prove to be helpful to other researchers trying to code occupational entries, such as those attempting to code death certificates.

Thus, if one assumes that the occupational data provided on tax returns is valid, the pilot effort underway could open many doors for researchers interested in epidemiology and other occupation - related studies.

Acknowledgements

The authors would like to thank June Walters and Karen Hui (both of the Internal Revenue Service) for their work in coding the sample and tallying the results, as well as H. Lock Oh, of the Social Security Administration, for handling all the computer work. Technical assistance by Warren Buckler and Sam D'Avella, of the Social Security Administration; Roger Knaus and John Priebe of the Census Bureau; and Milo Peterson, of the Office of Federal Statistical Policy and Standards, was appreciated. Helpful comments on this paper were received from Wendy Alvey, Beth Kilss, and Fritz Scheuren, of the Internal Revenue Service. Many thanks also to Mary Haigler (IRS), who typed the paper.

Footnotes

1/ By way of explanation, the SOC is a cumulative system, with each digit in the four-digit code representing a greater level of detail. It consists of two volumes: the manual [7], which lists all the codes and titles in numerical order, giving a brief description of each occupational group; and the index [8], which lists occupational titles in alphabetical order and gives the corresponding code.

2/ The DOT, like the SOC, consists of two volumes, one alphabetical and one grouped by occupation [9]. It is primarily intended for use by persons working in the field of employment services. Because the Office of Federal Statistical Policy and Standards of the Commerce Department plans to standardize occupational coding with the SOC, and because the DOT system is too detailed for our purposes, we chose to use the SOC system.

References

[1] Koteen, G., "Occupations Reported on Individual Tax Returns--Tax Year 1973," memorandum dated August 28, 1975, Statistics Division, Internal Revenue Service. See LASS Working Notes No. 2, Office of Research and Statistics, Social Security Administration, January 30, 1979, pp. 1-13.

[2] Koteen, G., and Grayson, P. "Quality of Occupation Information on Tax Returns," paper delivered at the 1979 ASA Meetings in Washington, D.C. See Statistical Uses of Administrative Records With Emphasis on Mortality and Disability Research, Office of Research and Statistics, Social Security Administration, October 1979, pp. 73-81.

[3] Reiser, B.S., "Occupation Data Reported on Individual Income Tax Returns--Tax Year 1968," memorandum dated March 12, 1970, Statistics Division, Internal Revenue Service. See LASS Working Notes No. 2, pp. 21-26.

[4] Sailer, P.J. and Robinson, C. "Feasibility of Occupational Coding from Tax Returns--Tax Year 1970," memorandum dated July 29, 1971, Statistics Division, Internal Revenue Service. See LASS Working Notes No. 2, pp. 14-20.

[5] Aziz, F., Kilss, B. and Scheuren, F., "Occupational Coding from Tax Returns in the Pilot Link Study -- Tax Year 1963." See LASS Working Notes No. 2, pp. 27-58.

[6] U.S. Bureau of Labor Statistics, Department of Labor, Employment and Earnings--1978, Table A30, U.S. Government Printing Office, January 1979.

[7] Office of Federal Statistical Policy and Standards, Department of Commerce, Standard Occupational Classification Manual -- 1977. U.S. Government Printing Office, October 1977.

[8] Office of Federal Statistical Policy and Standards, Department of Commerce, Index, Standard Occupational Classification Manual--1977. U.S. Government Printing Office, June 1978.

[9] Employment and Training Administration, Department of Labor, Dictionary of Occupational Titles--1977. U.S. Government Printing Office, 1977.

Table 4. --Distribution of Individual Income Tax Returns by Occupation for the United States, 1916

[Income returned for the calendar year ended Dec. 31, 1916.]

Occupations.	Returns.		Net income.		Income tax (normal and additional).	
	Number.	Per cent each class is of total number of returns.	Amount.	Per cent each class is of total net income returned.	Amount.	Per cent each class is of total tax.
Accounting professions: Accountants, statisticians, actuaries, etc.	4,229	0.967	$25,932,801	0.412	$183.911	0.106
Architects	1,419	.325	13,701,121	.218	154.108	.089
Artists: Painters, sculptors, etc.	786	.179	7,651.573	.121	118,767	.069
Authors, editors, reporters, etc.	2,529	.579	24,077,768	.382	323,518	.187
Clergymen	1,671	.382	10,868,893	.173	68,560	.040
Engineers: Civil, mining, etc.	6,628	1.517	63.319,332	1.005	960,993	.554
Lawyers and judges	21,273	4.868	245,139,302	3.892	4,289.869	2.474
Medical profession: Physicians, surgeons, oculists, dentists, nurses and other medical specialists	20,348	4.656	143,577,410	2.280	1,256,645	.725
Public service: Civil	2,992	.685	20.427,287	.324	158,726	.092
Public service: Military	5.459	1.249	28,965,909	.460	158,381	.091
Theatrical profession: Actors, singers, musicians, etc.	914	.209	11,128,927	.177	242.854	.140
Teachers: From kindergarten to university; also school and college officials	2,919	.668	19,345.751	.307	117,961	.068
All other professions and occupations	2.913	.667	28,637,580	.455	551.897	.318
Professions or occupations not stated	7,350	1.682	69,038,685	1.096	834.570	.481
Agriculturists: Farmers, stock raisers, orchardists, etc.	14,407	3.297	129,642,432	2.058	1,815,945	1.047
Bankers	6.518	1.492	206,970,133	3.286	12.296,089	7.092
Real estate brokers: Agents and salesmen	6,146	1.406	72,256,877	1.147	1.451,288	.837
Stock and bond brokers	2.839	.649	116,425,299	1.848	5.418,931	3.125
Insurance brokers	1,414	.324	18,314.501	.291	431.157	.249
Brokers: All other	7.479	1.711	152.552.715	2.898	8.063.569	4.650
Capitalists: Investors and speculators	85.465	19.556	1.679.228.016	26.660	55.540,102	32.033
Commercial travelers	12.274	2.809	74.252.624	1.179	466.677	.269
Corporation officials: Secretaries, managers, cashiers, presidents, etc.	53,060	12.141	717,402.707	11.390	15.522.667	8.953
Employees, all other: Superintendents, foremen, office employees, etc.	38.388	8.784	255,234.302	4.052	1,828.982	1.055
Hotel proprietors and restaurateurs	2.752	.629	28.537,581	.453	471.284	.272
Insurance agents and solicitors	7.243	1.657	58.551.346	.930	596.590	.344
Labor, skilled and unskilled	2.304	.527	16.104.057	.256	148.586	.086
Lumbermen	1.319	.302	18.311.335	.291	230.012	.133
Manufacturers	23.631	5.407	589,310.945	9.356	19.354.134	11.162
Merchants and dealers: Storekeepers, jobbers, commission merchants, etc.	54.363	12.439	836.502.071	13.281	21.192.547	12.223
Mine owners and mine operators	2.554	.584	115.288.799	1.830	7.226.758	4.168
Saloon keepers	1.311	.299	9.356.499	.149	92.601	.053
Sportsmen and turfmen	245	.056	1.963.705	.031	22.420	.013
Theatrical business: Owners, managers, etc.	811	.186	12.405.124	.197	268.703	.155
All other business	18.605	4.257	230.550,387	3.660	4.980.662	2.872
Business not stated	12.478	2.855	217,603.826	3.455	6,546.230	3.775
Grand total	437.036	100.000	6,298,577,620	100.000	173,386,694	100.000

Source: United States Internal Revenue (1918), Statistics of Income for 1916, p. 31.