OCCUPATION DATA FROM TAX RETURNS: A PROGRESS REPORT

Patricia Crabbe, Peter Sailer, and Beth Kilss
Internal Revenue Service

For some time the Internal Revenue Service (IRS) has been involved in interagency efforts to link data from its administrative record systems to produce statistical samples for use in many diverse research fields, ranging from the labor force and environmental health to life-cycle earnings and the distribution of wealth. The present paper focusses on IRS' research in using the tax return to obtain occupational data, a key variable in all of these areas.

Until recently, coding occupation data was a manual operation. Computerized dictionary systems, however, are fast replacing the slower, tedious manual methods [1]. Coding efforts at the Internal Revenue Service have followed the same path of development as elsewhere [2]. Presently, we are engaged in our first large-scale computerized effort in this area. (All the smaller pilot efforts at the Internal Revenue Service over the past two decades were done manually.) The outcome that we envision could prove to be helpful to other researchers trying to code occupational entries (on death certificates, for example [3]) and could open many doors for researchers interested in epidemiology and other occupation-related studies.

The purpose of this paper is to take a look at our progress to date in coding occupation for the entire 1979 Statistics of Income (SOI) sample of individual income tax returns. Organizationally, the material is divided into four parts. The first of these is a brief look at the pilot study that preceded our present project. In particular the importance of the pilot study for the overall project is discussed.

The focus of the second part of the paper is on the Internal Revenue Service methodology used to develop the computerized occupation dictionary. An examination of some of the problems involved in coding occupation entries made by taxpayers who have no instructions to follow is also discussed. The preliminary activities, during which the 1979 Statistics of Income file was prepared for coding, are described, as well.

The next section discusses a joint project with the Census Bureau to validate the occupation codes on the 1979 Statistics of Income file. This is followed by some early products from the coding.

The paper concludes with a discussion of potential uses of occupation-coded files and areas for future study.

1. PILOT STUDY

In 1979, two developments spurred renewed interest in the codability of income tax returns. One was the establishment of a new coding system as the Government standard for producing and presenting occupational data. This was the Standard Occupational Classification system (SOC). The other was the research conducted at the Social Security Administration (SSA) and elsewhere [4] into the possibility of creating a Linked Administrative Statistical Sample (LASS).

In connection with the LASS project, the Social Security Administration, which had been looking into this area for sometime, agreed to fund one more occupation pilot study, on a somewhat larger scale than the previous ones [5 through 8], to see whether tax returns were codable to the new Standard Occupational Classification system. In addition, the pilot study was to address itself to the problem of cost--if we were ever to do a large-scale coding project, we were going to have to devise a system that did not require professional judgment to code each return. The details of this pilot, for which we used a simple random sample of 6,700 returns filed for Tax Year 1976, were presented in a paper given at the 1980 American Statistical Association meetings [9]. Two major conclusions came out of that study:

1. Without any knowledge of the industry in which an individual worked, it was difficult to code taxpayers' entries, except into the broadest Standard Occupational Classification groupings--basically, those represented by the first digit of the four-digit Standard Occupational Classification code. However, once the industry code was added, as many as 89 percent of the returns appeared codable, especially if certain assumptions were allowed (for instance, a taxpayer calling himself a "worker" in the steel industry could be coded as a steel worker).
2. The manual available for the Standard Occupational Classification system, by itself, was inadequate for use by clerks in coding tax returns. Either a new manual would have to be developed, or we could try to break new ground by developing a computerized dictionary that would incorporate all the decisions painfully arrived at during the pilot study. Adventurous as we were, we chose the latter course.

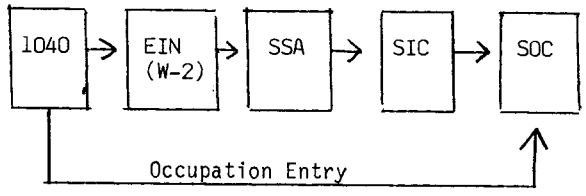
The problem of figuring out which industry the taxpayer worked was solved by using the W-2's attached to the tax returns in our sample. The W-2's contained employer identification numbers (EIN's) which, when matched to the Social Security Administration's Employer Identification File, gave us an industry code. The combination of taxpayer occupation entries on the 6,700 returns in our sample and the industry codes we chose to go with each entry (based on our match to SSA's files) became the basis of the first phase of our computerized dictionary for occupation coding returns.

Based on the favorable results of the Pilot Study, the Social Security Administration and the National Cancer Institute (NCI) agreed to sponsor a large-scale coding effort. This was performed on a stratified probability sample of about 203,000 tax returns--the entire Statistics of Income (SOI) sample for Tax Year 1979.

2. CODING THE 1979 INTERNAL REVENUE SERVICE STATISTICS OF INCOME FILE

When we began the preliminary activities to prepare the 1979 Statistics of Income file for coding, the keypunchers were instructed to key-enter the word or words the taxpayers had written in the occupation box on their individual income tax returns (Form 1040 or 1040A).

The Social Security Number on each return led us to a Form W-2, Wage and Tax Statement. The Employer Identification Number (EIN) from the W-2 was sent to the Social Security Administration, whose files contain industry codes for most employers [10]. We searched our own files for industry codes for self-employed persons. We now had not only an occupation entry, but also a Standard Industrial Classification (SIC) code for most taxpayers.



It should not be assumed from the brief description above that the matching of all these files was a simple matter. Above all, the matching of Employer Identification Number's from our W-2 file to Social Security Administration's Employer Identification File posed legal and technical problems that set the whole project back by about one year: (1) The lawyers had to work out mutually acceptable agreements that allowed the Social Security Administration and the Internal Revenue Service to share information without disclosing any protected data about individuals; and (2) the systems analysts had to figure out how to make seemingly incompatible computer systems compatible to each other.

Even after the matching, we still had numerous problems trying to interpret our merged file. Among the things we had to cope with were:

1. Not all 1040 salaries were backed up by W-2's.
2. Not all W-2 salaries were reported on 1040's.
3. Some 1040 salaries were backed up by a W-2P, not a W-2 (these were largely disability pensions).
4. Even if there was a W-2, we did not necessarily have a salary amount, due to the illegibility (to man and machine) of the entries. Some of these W-2's with no money amounts did, however, contain valid EIN's.
5. Even if we had a W-2 with a valid EIN, we did not necessarily get an industry code from the Social Security Administration [11].

All in all, we ended up with industry codes for about 71.5% of salaried employees in our sample. However, we only needed an industry code in about 40% of all cases in order to generate an occupation code. Furthermore, even some of the 40% may be occupation-codable at the one- or two-digit level without SIC information.

As the various match problems began to drag on, we decided to give ourselves a head-start on expanding the occupation dictionary by reading out any occupation titles that did not match to the dictionary created in the pilot study. Obviously, this would only give us a portion of the potential nonmatches, since some of the returns which matched on occupation titles might still not match on industry code. Nonetheless, we would at least get a partial answer to the one question that remained from the pilot study: Would enough entries on the 1979 tax returns match those on a small sample of 1976 tax returns to validate the whole concept of a computerized coding system? Below are the preliminary results from matching the 203,000 tax return sample (359,048 taxpayers) of the pilot study dictionary, which consisted of 2,944 occupation titles [1, 12].

Status	Number of taxpayers	Percent of total
Total	359,048	100.0
Matched	229,939	64.0
Unmatched	129,109	36.0

Even these figures tend to overstate the nonmatching problem, since they include both taxpayers who did not give any occupation and those taxpayers with an unmatched title. Each time a title did not match it counted as a nonmatch, even though several of these may have been for the same title. The total number of unmatched taxpayer entries was 39,347; many of these were simply misspellings or abbreviations of titles that had already been coded; some were cases that could not be coded at all. With each subsequent use and updating of the dictionary to incorporate these variations, the percentage of returns automatically codable should grow higher.

Coding Structure

The purpose of the computerized occupation dictionary is to convert taxpayers' entries in the occupation box to standard occupation classification codes. The structure of the standard occupation classification codes is generated by a four-level system. As seen below, the first level usually represents the division. The division level for production working occupations is 7000. The second level represents the major groups. 7500 is the major group for machine operators and tenders. The third level represents the minor groups. The minor group 7510 includes occupations involving operating and tending metal working and plastic working machines. The fourth level represents the unit groups--as shown below, 7513 is the unit group which includes occupations involving operating and tending lathe and turning machines.

Standard Occupation Classification System

1.	Division.....Production Workers	7000
2.	Major Group..Machine Operators and Tenders	7500
3.	Minor Group..Metal and Plastic Machine Operators and Tenders	7510
4.	Unit Group...Lathe and Turning Machine Operators and Tenders	7513

In some cases, there was a direct one-for-one conversion from occupation title to the standard occupation code. A few examples of these are lawyer, secretary, physician, and waiter. In other cases, it was necessary to consider both the taxpayer's occupation title and his or her industry code before assigning a standard occupation classification code. An example is the word "stripper". In industry 2752, which is establishments primarily engaged in manufacturing printing by the lithographic process, the standard occupation classification code is 6868, which is stripper photolithographic. Stripper appears in our dictionary two other times with different industry codes (3861 and 7399). The corresponding standard occupation classification code for these industries is 6842. However, one industry (5813) with the word stripper could not be determined directly from the published SOC Manual. This industry consists of establishments primarily engaged in the retail sale of drinks, such as beer, ale, wine, liquor, and other alcoholic beverages for consumption on the premises. After considering the industry, a judgmental decision was made to include these strippers in standard occupation classification code 3270, which is a type of dancer.

Dictionary Examples

Title	SIC	SOC
Lawyer	*	2110
Secretary	*	4622
Physician	*	2610
Waiter	*	5213
Stripper	2752	6868
Stripper	3861	6842
Stripper	5813	3270
Stripper	7399	6842

* No SIC needed.

Dictionary Structure

Here's how the dictionary works. The left-hand column contains words taken from actual taxpayers' returns. The middle column contains industry codes necessary to interpret the taxpayer's entry. The right column contains the standard occupation classification code chosen for each taxpayer entry.

At this point, it should be noted that certain occupation entries we found on the 1979 Statistics of Income file simply were not appropriate for entry in our computerized dictionary. For example, entries such as human being, slave, bum, bore, beatnik, and dunce/fool appeared to be more of psychological than of demographic significance. Fortunately, there were less than .04 percent of these occupations listed in the occupation box on the tax return. In fact, close to 89 percent of the tax returns did have comprehensible occupations listed. In total, our dictionary now contains close to 39,000 lines [13], which represents the number of different combinations of occupation entries and industry codes found in the 1979 Statistics of Income file [14]. Since the 1979 file contains about 359,048 taxpayers, this means that, on the average, each dictionary entry coded about 11 taxpayers.

3. VALIDATING OCCUPATION CODES ON THE 1979 STATISTICS OF INCOME FILE

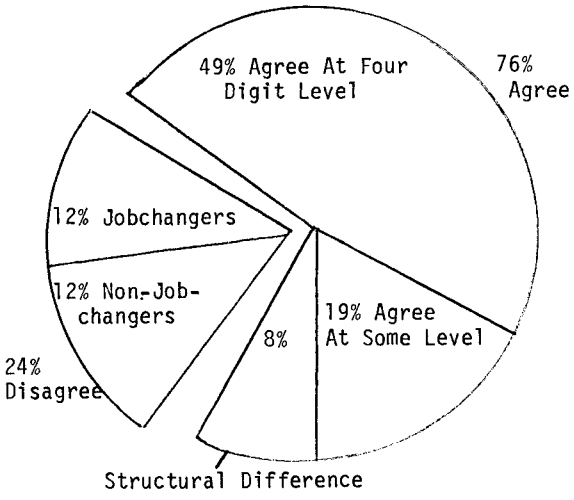
We have a fairly high level of confidence in the reliability of the coding process. The creation of the computerized occupation dictionary was quite an expensive undertaking; hence, it was obviously essential that a certain amount of work be conducted to carefully validate the effort.

The Census Bureau has considerably more experience in coding occupation than we do at IRS. In addition, their surveys contain additional questions (such as major activities or duties of respondent) which help them interpret their respondents' occupation entries. Moreover, most of their surveys contain educational level. The sample of one of the Census Bureau surveys was assigned to coincide with the SOI sample for roughly 800 individuals in order to evaluate the results of our coding system. Briefly stated, the project involved a comparison of the occupation provided in the survey with occupation codes produced for the SOI with our occupation coding system. It should be noted that the sample selection, clerical matching and review were conducted at the Census Bureau to protect the confidentiality of the survey respondents [15]. It should be noted, further, that even at the Census Bureau, occupation coding is not an exact science. For example, in a study where the Census Bureau compared occupation information obtained by interviewers to self-reported data by respondents in selected engineering, scientific, and technical occupations, 14% of the sample had to be eliminated because of insufficient response, and only 44% of the remainder agreed exactly at the most detailed level of occupation coding [16]. This result should be kept in mind, since the IRS information is self-reported, while the Census information being used to validate it was obtained in personal interviews.

Census/IRS Validation Study

As a result of the Validation Study, just about half of the returns agreed at the four-digit level of coding with the code derived from the Census survey (See below). An additional 8 percent would have agreed at the four-digit level if the Census Bureau had not made structural changes to the standard occupation classification system. An additional 19 percent of the cases agreed at the division level, although not at the four-digit level. In analyzing the returns where the codes disagreed, we found that roughly one-half of these taxpayers appeared to have changed jobs at some time between the beginning of the taxable year and the interview by the Census Bureau. This determination was made by comparing the name of the employer on the largest W-2 to the employer name shown in the Census survey. So there is at least a possibility that both the Census and the Internal Revenue Codes are correct for these taxpayers, only for different points in time.

COMPARISON OF OCCUPATION-CODED RESPONDENTS/TAXPAYERS IN THE CENSUS/IRS VALIDATION STUDY



Obviously, the results are based on a very small sample and, therefore, are far from conclusive. Hence, we are exploring the possibility of using a larger IRS statistical sample of taxpayers who have been interviewed by IRS agents to do a large-scale test of our computerized dictionary. Nevertheless, the Census Bureau validation study did point out some problem areas which need future study.

Shown below is the degree of correspondence between IRS and Census codes for ten major occupation divisions. For some occupations the correspondence is remarkably good. However, the chart also points out that a major problem exist in the area of unskilled workers. Typically this group gives the most cryptic entries in the occupation box (often showing industry rather than occupation); interpreting these entries is a major challenge. (See the Appendix for a sample list of the actual occupation titles as they appeared on the tax returns.)

Level Of Correspondence Between IRS And Census Coding, By Major Occupational Divisions

Census Occupation	Percent
Overall Level Of Correspondence *	76
Executives and Administrators	94
Professional Specialty Occupations	84
Technicians	83
Sales Occupations	68
Administrative Support Occupations	83
Service Occupations	82
Precision Production and Production Working Occupations	80
Transportation Workers	83
Handlers, Equipment Cleaners, Helpers, and Laborers	25

* Farmers were excluded because the Census results are based on only two returns.

4. FUTURE PLANS

Our immediate plans for the occupation data are to publish a supplemental report in the Statistics of Income series, showing income earned and taxes paid by taxpayers classified by occupation and sex. Together with the National Cancer Institute, we plan to look into ways of using occupation-coded taxpayer data to study various epidemiologic problems. In fact, it was with these long range goals in mind, that we joined with the National Cancer Institute in designing the 1979 Statistics of Income sample to coincide with cases which fall into Social Security's 0.1 percent Continuous Work History Sample (CWHS). In that way, we hope to eventually introduce occupation data from tax returns to the CWHS, thus enhancing the industry data already on that file. If serval legal hurdles can be overcome and we find a way to merge age data from the Social Security Administration and cause-of-death data from death certificates with this file, we will be well on our way to creating a large-scale, low-cost longitudinal sample for monitoring occupational health issues. We are also exploring with the National Center for Health Statistics the possibility of using their National Death Index to lead us to the death certificate [17].

There are many other (not epidemiologic) potential uses for our new-found ability to occupation-code tax returns and we can only touch on a few of them here. For instance, a sample of tax returns is currently being used to produce statistical tabulations which monitor the progress of Indochinese refugees [18]. If these data were enhanced by occupation-coding, we could tell not only whether progress was being made, but in what occupation, with obvious implications for public policy in the area of job-training.

Another sample of tax returns is used to produce statistical tabulations of persons leaving the Armed Forces [19]. Being able to tell in which occupations exservicemen and servicewomen are doing best would give Congress and the Department of Defense a better understanding of the pay structure needed to encourage experienced technicians to remain in the Armed Forces.

Finally, it should be noted that the methodology developed for coding income tax returns may also prove usable for coding other documents which contain occupation information, including death certificates. Furthermore, it has applicability for the development of coding schemes for similar open-ended questions of administrative records and survey questionnaires.

ACKNOWLEDGEMENTS

The authors would like to take this opportunity to extend special thanks to the National Cancer Institute for the funding of this project. Additional thanks are due to Victor M. Valdisera and Doug Sater, of the Census Bureau for the very timely analysis they performed on the Census/IRS validation study. At IRS, a special thanks also goes to Victoria Pazulski and William Bradley, for their help in the conversion of occupation titles to computerized

coding. Thanks to Al Schreier, Cynthia Brown, and Marianne Hanoien of the IRS Data Center for their help merging the data files. Much appreciation is also due to Wanda Thomas and Roselind Vinson for their extensive efforts in typing this report.

However, since all comments from the reviewers may not have been incorporated, due to the deadlines on this paper, any shortcomings would have to be blamed on the authors.

NOTES AND REFERENCES

[1] Anderson, Ronald and Lyberg, Lars "Automated Coding at Statistics Sweden" 1983 American Statistical Association Proceedings, Section on Survey Research Methods.

[2] Appel, Martin and Hellerman, Eli. "The Census Bureau Experience with Automated Industry and Occupation Coding," 1983 American Statistical Association Proceedings, Section on Survey Research Methods.

[3] Rosenberg, Harry M., Burnham, Drusilla, Spirtas, Robert, and Valdisera, Victor. "Occupation and Industry Information from the Death Certificate: Assessment of the Completeness of Reporting," 1979 American Statistical Association Proceedings, Section on Survey Research Methods, pp. 286-292.

[4] Kilss, Beth, Scheuren, Fritz and Buckler, Warren. "Goals and Plans for a Linked Administrative Statistical Sample," 1980 American Statistical Association Proceedings, Section on Survey Research Methods, pp. 450-455.

[5] Koteen, Gloria and Grayson, Paul. "Quality of Occupation Information on Tax Returns," 1979 American Statistical Association Proceedings, Section on Survey Reserch Methods, pp. 280-285.

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

[7] 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, Office of Research and Statistics, Social Security Administration, January 1979, pp. 21-26.

[8] Sailer, P.J. and Robinson, C. "Feasability 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, Office of Research and Statistics, Social Security Administration, Janaury 1979, pp. 14-20.

[9] Sailer, Peter, Orcutt, Harriet and Clark, Phil. "Coming Soon: Taxpayer Data Classified by Occupation," 1980 American Statistical Association Proceedings, Section on Survey Research Methods, pp. 467-471.

[10] This is the most economical means for the Internal Revenue Service to acquire a Standard Industrial Classification Code.

[11] This occurred because the Social Security Administration has never been given the funds to follow up on employers who fail to supply them with industry information.

[12] Actually this is an unexpectedly high match rate. By comparison a similar experiment was conducted at Statistics Sweden. For Statistics Sweden to get an overall coding degree of 65% they ended up with 4,230 dictionary descriptions.

[13] However, mentioned earlier there were 39,397 occupations to be coded. The difference reflects uncodable occupation titles.

[14] The pilot dictionary contained only about 3,000 entries.

[15] In this August 1980 Income Survey Development Program (ISDP) Special Frames Test, the Census Bureau included a sample of 800 individuals who were to be included in the 1979 Statistics of Income file. The sample was selected from the 1978 IMF extract which contained mailing addresses at the time of filing (between January 1979 and September 1979). Therefore, there were sample persons not found due to business vs. residence addresses and due to moving between 1979 and August 1980. For the purpose of the present study, Census nonrespondents and refusals were eliminated. This left 436 primary taxpayers in the comparison. Of these, 43 gave no response to the occupation question on the tax return, and 50 were non-labor force taxpayers. Therefore the comparison is based on 343 matched individuals.

[16] U.S. Department of Commerce Bureau of the Census An Evaluation of 1970 Census Occupational Classification.

[17] Patterson, John E. "Evaluation of the Matching Effectiveness of the National Death Index," 1983 American Statistical Association Proceedings, Social Statistics Section.

[18] Reimbursable Contract with Internal Revenue Service and the Department of Health and Human Services, Social Security Administration, Office of Refugee Resettlement.

[19] Reimbursable Contract with Internal Revenue Service and Department of Defense Survey Market Analysis Division, Defense Manpower Data Center.

Appendix: Occupation Entries on Tax Returns

A random sample of 100 occupation labels were extracted from the 1979 Statistics of Income file. Listed below are the occupation labels, standard industrial classification code(s) (SIC) and occupational classification code (SOC) where applicable.

Table 1.--Random Sample by Industry and Occupation Code

Occupation Title Listed by Taxpayer (COD)	Standard Industrial Classification Code (SIC)	Standard Occupation Classification Code (SOC)	Occupation Title Listed by Taxpayer (COD)	Standard Industrial Classification Code (SIC)	Standard Occupation Classification Code (SOC)
1. Insurance	6411	4122	51. Hospattndt		5236
2. Lawyer		2110	52. /	8221	
3. Attorney	8111	2110	53. Executive	2052	1210
4. Asst. Bldr./Expedite	1500	1330	54. Executive	2899	1210
5. Dentist	8021	2620	55. Waiter/Student		5213
6. Wholesaler		4020	56. Counselor	8059	1419
7. /			57. Salesman	2319	4243
8. Laborer		8700	58. Executive	2099	1210
9. Self/Employed	4230	1342	59. Stock Broker	6211	4124
10. /			60. V/P Sales	3674	1210
11. Investments		9200	61. Physician	8011	2610
12. Dentist	8021	2620	62. Auto Mechanic	5511	6111
13. Attorney	8111	2110	63. Cook	5812	5214
14. Physician	8011	2610	64. Clerk/Typist		4624
15. Self/Employed	1300	1360	65. /	9011/3613	
16. Produce Manager			66. Physician	8071	2610
17. Auto Sales	5521	4030	67. Asst. Director	2051	1210
18. Teacher		2000	68. Driver	5100	8210
19. Racehorse Trainer	0754	3400	69. Retired		9300
20. Secretary		4622	70. Executive	2034	1210
21. Physician		2610	71. Student	7393	9500
22. Physician	8011	2610	72. Sprinkler filler	1711/3599	6450
23. Banker	6020	1220	73. Executive	1542	1210
24. Bank Officer	6031	1419	74. Investments		9200
25. Printer	2761	7643	75. Work Processing		4793
26. Molder	3321	8769	76. Work Processing		4793
27. Clerk	8399	4600	77. Student	5411	9500
28. Financial Analyst	3691	1419	78. Mechanic	5541	6111
29. Machine Operator	7218	7673	79. Parts Salesman	7622	4367
30. /	4213		80. Mechanic	5541	6111
31. /	3079		81. Self/Employed	4119/7538	1342
32. Retired		9300	82. Clerk	5812	4364
33. Manager of Store	5921	4030	83. Physician	8062	2610
34. Administrator		1390	84. Medical Doctor	8062/8011	2610
35. Self/Employed	8111	2110	85. Construction Worker	1623	6479
36. Teacher	2065	2390	86. Insurance Salesman	6311	4122
37. Retired		9300	87. /	5087	
38. Counselor			88. Auto Repair		6111
39. Postal Clerk	9011	4742	89. /		
40. Corp President	3341	1210	90. Mechanic	5711/3711	6130
41. A/Painter	5170	6442	91. Driver	5141	8210
42. Dentist	8021	2620	92. Student	8641/5082	9500
43. Service Manager	5511	6000	93. Control Clerk	5812	4363
44. Set Operator	2258	7650	94. Outside Sales	2751	4366
45. Military	9011	9100	95. Farm Hand	6211	5612
46. /		1521	96. Executive	3743	1210
47. Executive	5511	1210	97. Lawyer	8111	2110
48. Stock Clerk	8062	4754	98. College Prof./Ind. C	8221/7392	2200
49. Assist. Supervisor	6143	4529	99. Retired		9300
50. Physician	8011	2610	100. /	8011	

Note: The symbol / denotes that no taxpayer entry for occupation was provided.