DEVELOPING METHODS FOR COLLECTING AND CODING THE OCCUPATION OF PERSONS WITH COLLEGE DEGREES

Linda P. Hardy, Donna L. Eisenhower

KEYWORDS*:* Occupation coding, occupation collection

In 1993, the surveys of college educated persons, sponsored by the National Science Foundation, were redesigned. This paper describes that part of the redesign which focused on the development of improved methods for collecting data on occupation. The project examined open-ended and structured types of question formats and analyzed the types of errors associated with each of these. In the open-ended format, the individual is asked to describe their job and their response is clerically coded later. The structured method asks the respondent to choose the correct occupation from a list. Later work included the development of a combined open and structured format, cognitive testing, and development and analysis of a clerical review process used for the 1993 surveys. The resulting system for occupational coding is based on a concept of "best code" which utilizes both open-ended and structured responses, along with other relevant information from the questionnaire, to derive the "best" occupational code. The development of a specialized occupational manual and index, and associated procedures is presented. Initial reports on the functioning of the system and plans for further analysis are described.

DEVELOPING METHODS OF COLLECTING THE OCCUPATION OF PERSONS WITH COLLEGE DEGREES

Donna L. Eisenhower, Mathematica Policy Research, and Linda P. Hardy, National Science Foundation

Key words: Occupation Coding, Occupation Data Collection

The SESTAT system:

The National Science Foundation (NSF) is creating a new system of data about scientists and engineers. This system, called SESTAT, provides a variety of data to support analysis of science policy and general research. While the system will contain data from many sources, much of the data comes from three NSF sponsored surveys of college graduates. The surveys are done biennially and produce longitudinal as well as cross-sectional data. 1/ The SESTAT surveys covers most college educated scientists and engineers (S&E) living in the U.S. (There are some coverage problems with foreign educated S&E, however.) The SESTAT data system results from a redesign of NSF data collection efforts of the 1970s and 1980s. In the mid-1980s, NSF sponsored a panel study by the National Academy to review the existing data collection and estimation work and make recommendations for the future of the system. The Committee on National Statistics report, issued in 1989, provided a comprehensive framework for the redesign work that was to follow.2/

An important goal of the redesign was to better serve user needs by more clearly defining the target S&E population. The new definitions allow researchers to study persons either working as or educated as scientists and engineers, or both. (The relevant variables are occupation or degree field.) A related redesign goal was to improve internal and external compatibility of the S&E data. For internal compatibility, the three surveys were made considerably more consistent---re: questions and formats used, processes, editing, etc. For external compatibility, emphasis was placed on using concepts and definitions used by other government agencies. 3/

NSF Occupation Variables:

Persons working as scientists and engineers form one of the two basic groups defined in the SESTAT system. The occupational variables collected in SESTAT are: current occupation; previous occupation--if unemployed or not in the labor force; and occupation *5* years ago. Because of the importance of the occupational variable and its use in defining the scope of the S&E population, the redesign sought to examine and, if possible, improve on methods of occupational data collection. Any improvement, however, would need to work well with all 3 modes used in the surveys: mail, CATI and personal visits.

Nominal Definitions and Methods of Collecting Occupation:

The SESTAT nominal definition of occupation refers to the kind of work done on the job held during a given week. As with the Standard Occupational Classification (SOC) system, the guidelines for classifying the respondent's occupation are (1) on the basis of work performed and (2) on the basis of primary work activity and/or the work activity in which the person spends the majority of their time.

There are two principal question formats used for collecting occupational data: open-ended and structured list. In the open-ended format, the respondent is asked to write/describe their job in one or a series of questions. The responses are then clerically coded, typically using a reference manual system. In the structured list method, the respondent is asked to choose the correct occupation from a list--a self-assigned code. The list is typically structured into occupational groups to help the respondent find the correct occupation

more quickly and to give the occupation some definition through the context of the group.

Previous NSF surveys collected occupational data using a structured list of approximately 90 occupational titles, grouped into 11 categories. In addition to choosing an appropriate code, the respondent also was instructed to write-in the accompanying occupational title from the list. The later was clerically checked to reduce respondent transcription errors. There were no prior studies of the accuracy of the method used.

Some other surveys have used a structured list system for collecting occupational data. The Occupational Employment Statistics (OES) Survey, conducted by the Bureau of Labor Statistics, uses the structured list method for collecting data from employers.4/ OES lists include occupational titles, alternate titles, and short definitions. The length of the OES occupational lists are from about 4 pages, used for small establishments, to 16 pages for some larger ones. Some Department of Education surveys also use a short structured lists for collecting occupational data.

Beginning in 1940, occupational data was collected in the decennial census using an open-ended question format. The information collected was used by coders to assign occupational codes.5/ The decennial form also included questions on type of employer. The employer information was used for industry and ownership code assignment, and in turn, those codes were used to supplement the occupational coding process. Then (as now) the occupation question asked about the “kind of work” done. In 1970, an additional occupational question on the respondent’s usual duties was added. This information helped in coding some occupations where titles were unclear.6/ The basic method of occupation coding using two open-ended responses and supplemental employer information continues to be used in most Census administered survey--including the decennial census, Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP). Coding procedures have evolved over time, as well as an extensive reference system used support the coding process. 7/

Relative Advantages and Disadvantages of Occupational Data Collection Methods:

Open-ended Method: The open-ended question method has a number of advantages. Practical advantages are savings of space on a mail questionnaire and a more straight-forward CATI implementation. A detailed structured list (without definitions) typically takes a minimum of 1-2 pages, while an open-ended format. such as used by Census, takes much less. For mail/CATI surveys, mixed mode effects are expected to be less with the open-ended method, but CATI data capture can introduce problems which lessen this advantage.

A significant advantage with an open-ended format has to do with the ability to control the scope/content of the occupations. This is done through the coding process and the associated reference materials. Also, the reference materials (such as the index of detailed occupations used by the Census Bureau) can provide a direct link between the operational definitions used in data production and the data users. Another advantage is that open-ended responses may be retroactively coded and emerging occupations more readily identified.8/

The major disadvantage of thc open-ended format could be termed the “not-specified" problem. This is a serious problem which results from the respondent’s misunderstanding of the level of specificity needed (to assign an occupational code). For example, a respondent may say they are an “engineer”, yet the *type* of engineering is needed to assign an appropriate code. Even under the relatively controlled conditions of CATI, it is difficult for the respondent/interviewer to understand what information will be needed for the coding process that occurs later. In addition, CATI interviewers have some difficulty recording complete and accurate answers during the interview. (Cantor, 1992). In mail surveys, questionnaire designers hope to lessen the not-specified responses by showing detailed examples of occupational responses. However, the reading of examples increases respondent burden. In CATI versions, the interviewers must be well trained in structured probes--which must be anticipated in advance of the survey start-up.

An example of a serious not-specified problem is that of the post-secondary teacher group of occupations. In the 1990 decennial census (primarily a mail survey), 78% of the teachers were assigned the not specified post-secondary teaching code because respondents did not report their teaching specialty field. In the CPS, (primarily a CATI and personal visit survey) the not-specified rate for post-secondary teachers steadily increased from about 20% of total post-secondary teachers in 1983 to about 40% in 1994 (first quarter 1994 data). This, despite efforts in recent years to make interviewers aware of the data needs of the coding staff.

Another aspect of the not-specified problem is the need to code open-ended responses using general assumptions about the nature of the work performed. For example, in the census system, within the engineering occupations, a not-specified response of "engineer" is coded according to the respondent's industry code. Thus a response of "engineer", when the respondent's industry has been coded organic chemicals, is assigned the occupational code for chemical engineer. However, the OES Survey (1989) shows that more than half the engineers in that industry are *other than* chemical engineers. Information on the proportion of the estimates resulting from not-specified responses is not available. Both of the examples cited above, teachers and engineers, represent serious problems for accurately estimating the NSF S&E population.

Other considerations for the open-ended format include the resources expended in clerical coding and the need to maintain references with current and accurate entries. Inter-code variation is also a potential concern. For the OPS, a 10 percent quality control sample is coded each month after the regular CPS coding is completed. Typical coder difference rates are around 6 percent.

Structured List Method: A structured list method is not commonly used in demographic surveys. Accordingly, there is less understanding of the advantages and disadvantages of this method. The principal advantages are simplicity of data capture and the ability to convey to the respondent the level of detail needed. That is, it greatly reduces the not-specified problem.

A special advantage of the structured list method is primarily cognitive--a list of choices may help respondents make more precise decisions between alternative occupations. While the coder is better informed on the intricacy of the coding system, the respondent knowledge of the details of their occupation is often the key factor in deciding fine distinctions. In S&E occupations, these distinctions taken on added importance. The structured list format implicitly conveys to the respondent some of the basic classification principles being used. This is useful since it is not uncommon for a respondent to feel that several occupational categories apply--or to be confused as to the difference between occupation and industry or education.

The cognitive effect of a structured occupational list can best be described with the example of engineers. There is a very close linkage of the classification of engineering occupational specialties with programs conferring engineering degrees. Likewise, many engineers (as with other professionals) conceive of their occupations relative to their field of degree. 9/ However, some engineering occupations are classified according to application area--e.g. computer or environmental engineering, which generally does not relate well with the title of the degree conferred. It is thought that a structured format improves reporting of the application based engineering occupations by showing the respondent both the traditional and the application engineering occupations as alternative choices. (This issue will be examined further when coding results from the 1993 NSF sponsored surveys are available.)

Redefinition of NSF Occupations:

The SESTAT redesign involved redefining the NSF occupations to be compatible with the Standard Occupational Classification (SOC) System, including use of SOC principles of occupational classification as well as definitions of the S&E sub-groups. 10/ An SOC-based approach is similar to that used by other federal occupational data collection surveys, which are approximately 90% (or more) compatible with the SOC. (Dempsey,

1993). Compatibility in this context means that the crosswalk of the classifications are either one-to-one or can be made so through minor summations of the estimates from one of the system. For example, the NSF classification separates oceanographers from geologist, while the SOC system does not. Compatibility to the SOC can be achieved by summing the two NSF occupations.

Compatibility with the scope of the SOC occupations involves both occupational titles and definitions. In prior NSF surveys, a structured list method was used but definitions of the occupations were not included. Thus, the definition of an NSF occupation would probably differ between respondents and cause the same occupation to be reported under different occupational codes. This prior method did not support compatibility of definitions with the SOC, nor practically allow the data users to understand the definitions associated with the NSF occupational estimates.

Developing a new method for the NSF surveys:

The goals for the new method were accuracy and classification compatibility with other federal occupational estimates, through compatibility with the SOC. In the overall SESTAT redesign, a series of focus groups and think-alouds were used to examine potential problems with the questionnaire. Different focus groups were used for the various sub-populations of college graduates covered by the 3 NSF surveys--groups for recent graduates, older graduates, S&E and non-S&E graduates and PhD graduates. For occupation, both techniques were used to examine advantages/disadvantages of the open-ended and structured list method.

To get a more direct comparison, both open-ended and structured list occupational data were collected from 34 focus group members. Initially the members were asked to provide some background information in writing prior to coming to the focus groups session. This included the standard open-ended occupation and industry questions used in Census surveys. This information was sent to the Census occupational coding group at Jeffersonville to have Census occupation codes assigned. Census codes were used to allow coders use of a system more familiar to them and to take advantage of the existing occupational references. There is good compatibility of the NSF and Census occupations because both are SOC-based systems and subsequent analysis adjusted for differences between Census and NSF code lists. Comments were also collected from the coders as to the difficulties of coding the occupations. (As a side issue coders were also asked to code the open-ended responses without using the industry data. This was to explore the relative utility of the industry data in coding the NSF population.) Focus group participants were also asked to complete an NSF questionnaire which included choosing from a structured occupational list. These two steps provided both open-ended and structured list codes for the same individual.

Since the potential biases of the two methods were to be examined, additional in depth occupational information was obtained later from these individuals so a “true” occupational code could be determined, In addition, all relevant information on the questionnaire was used to evaluate the validity of the two coding systems. Results of this analysis are shown below.

Comparison of Open-ended Clerically Coded Response To Respondent Self-code Chosen from structured NSF List

Total:	34	100%
Both codes correct:	20	59%
Both codes incorrect:	5	15%
Open-ended code correct:	3	9%
Self-code correct:	6	17%

The data show that both the open-ended and structured method produce valid codes in most cases. However, sizable problems exist, with each method having an error rate of 1 in 4 or more. With managerial respondents, some type of problem was found in 10 of 12 cases, and for teachers in 2 of 4 cases. (We realize the small sample and non-random focus group membership selection method puts limitations on these estimates.)

The focus groups and other work verified that the NSF population was prone to similar reporting

errors found in other populations. The not-specified problem was found in the open-ended responses. Misinterpretations of the structured list were also identified. For example, clinical psychologists sometimes reported themselves under medical specialties despite the fact that there was a specific entry in the social science group for this occupation. Special problems with teachers and managers were found in both methods.

Revised NSF Questions on Occupation

The initial analysis confirmed that each method had shortcomings relative to occupational data for the college educated population. The "best code" procedure that was developed and is discussed later attempts to mitigate these problems by using all the relevant data captured on the questionnaire to correct respondent errors.

The occupational portion of the SESTAT forms first ask for an open-ended response. The question is modeled after the first CPS occupational question which was revised in the recent redesign. The occupational question is:

What kind of work were you doing on your principal job during the week of April 15--that is, what was your occupation?

The next question (which is intended to look visually separate) asks the respondent to choose a code from a structured list of 127 occupations arranged in 22 groups on 2 pages. The job list in intentionally placed on 2 facing pages to encourage the respondent to scan all the list before making a choice. The groups are alphabetized rather than hierarchically arranged. It was hoped that this would lessen respondent tendency to pick more socially acceptable jobs. The "upgrading" theory says that some respondents will misreport their occupation as one of higher education and/or social status for self-esteem or other reasons. In the future, we hope to examine this issue relative to the mail and CATI modes using the open and structured list information.

Development of the "Best Code" Procedures, Reference Materials and Coder Training

Occupational coding form and procedures: The coding procedure that was developed is termed a "best code" procedure because it attempts to determine the best occupational code using all the relevant information for the respondent--including the respondent's self-code which they selected from the list. The procedure is intended to be conservative. That is, the respondent's self-code would be changed only when two conditions arc met. The coder has "clear" evidence that the respondent has made an incorrect choice and that another, better, code choice is evident, Of course the meaning of "clear" evidence adds subjectivity to the process and is difficult to define in a practical sense. A few occupations with substantial sources of error, managers and teachers, had more specific decision rules for coders to follow. Special reference materials and training that were developed to lessen coder subjectivity are described later.

The development phase included determining which respondent information should be used by the coders and the order of priority for that information. To facilitate coder review and better control the coding process, it was decided that the information from the form would be organized onto a single page, with one page per respondent. Determining the relevant occupational information to be included on the coding page was an iterative process involving NSF, Mathematica Policy Research (technical consultant for SESTAT), the program staff from the survey organizations, and especially the occupational coding supervisors and staff. The final result was a coding page that included the following data reported by the respondent:

- Open-ended occupational response
- Employer name and address
- Type of employer: type of educational institution or ownership/class of worker information
- Number of persons supervised: directly and through subordinate supervisors
- Relationship of highest degree field and
- Categories of work activities and most important activities

Salary
Respondent's occupation self-code

Coders were encouraged to make notes on the coder form as needed and the procedure included a referral system so that more difficult cases could be resolved more accurately. Since more than one survey contractor was involved, a system was developed for resolving and sharing coding problems.

Typical quality control procedures were used in the coding. In the beginning. each coder's work was reviewed at a 100% rate and at a reduced rate afterwards. Emphasis was placed on limiting coding to those with prior experience in occupational coding. In addition, a 5% quality analysis sample was chosen and then coded a second time using the same procedures. When these results become available, some measures of inter-coder variance may be possible.

Coder Reference Materials: The coder reference materials consisted principally of the procedures manual itself and an alphabetical index of occupations. The index played an important role since it was used to associate an NSF code with the open-ended response. The index was developed using Census' "Alphabetical Index of Industries and Occupations" from the 1990 Census. This index lists over 30,000 detailed occupational titles and assigns them to one of 501 Census occupational categories. The Census index was changed in several ways to suit SESTAT purposes. Based on a crosswalk of Census to NSF occupations, NSF codes were substituted for Census codes in the index and adjustments were made for the lack of an industry code for the respondent. In many cases in the sciences and engineering, NSF occupations were more desegregated than in the Census system. This is understandable since Census' system is broader and does not concentrate on any one area such as professional occupations. For desegregated NSF occupations, each detailed index entry was assigned with a unique NSF code. To enhance the accuracy of the science and engineering detailed entries, a series of interviews was held with various NSF scientists and engineers to solicit additional/updated entries to the manual. As the actual coding work progressed some additions and corrections were made to the index--although this was limited.

Coder Training: The reference materials and coder training were the two principal methods for controlling accuracy and inter-coder variability. Coder training included review of the general coding principals, proper use of the coding form, and emphasis on case examples. The case examples were developed from the survey pretests that were done the year before. Two sets of case examples were developed--an initial set with only straight-forward examples that emphasized basic principles, and a more difficult set that would generate more discussion during the training session. Training included an explanation of the conservative approach that coders should take--changes only when the evidence is clear.

For the Census Bureau coding staff, which would code more than 2/3 of the over 200,000 SESTAT respondents, some adjustments were made to differentiate the NSF project from the coder's usual CPS work. These included emphasizing in training that sufficient time he taken to review each coding sheet thoroughly. (CPS production schedules are necessarily very tight and speed of coding is important.) Also Census codes were excluded from the index and manual despite the fact that many CPS codes have a one-to-one correspondence with NSF codes. This was done to encourage the coders to use the references provided and lessen reliance on memorized codes. Finally, a special code was created to be used when the coder felt the respondent's self-code was incorrect, but no better code was clearly evident. It was hoped that the availability of this special code would reinforce the desired conservative approach.

The final part of the coder training package was a suggested agenda for the coder training session and a coder certification exercise that could be given by the instructor. MPR and NSF staff trained survey trainers who, in turn, trained the coders. Some coders during debriefings mentioned that they felt more sponsor direct involvement in the training would be helpful.

Summary:

For the NSF SESTAT surveys, we examined both the open-ended and structured formats for collecting occupational data. Our analysis showed that both formats bad strengths and weaknesses. Since there was no clear winner, we developed the concept of using both formats and other relevant information from the respondent to arrive at the "best code".

Although estimates are not yet available, preliminary review of partial micro data suggests that the Best Code process makes a difference in the data. The following table shows the percent of the total number of records that were recoded *into* the occupational groupings shown--as a result of the Best Code process: The basis of the percent is the total associated with the revised occupational codes.

Percent of Group Total Resulting from a Code Change Into the Occupational Grouping

Bio/Life Scientists	7%
Computer Science Related	4%
Engineers	3%
Mathematical Scientists	7%
Physical Scientists	9%
Social Scientists	28%

While for the present we assume that the Best Code is the more valid code, we are working on evaluating the survey results and the quality control checks. These will be reported in a forthcoming paper.

Footnotes:

1/ The NSF sponsored SESTAT surveys include: (1) the Survey of Doctorate Recipients, SDR, conducted by the National Research Council has a target population of all PhD recipients from US institutions--sample size of about 50,000 individuals. NSCG is a mail survey with CATI follow-up to non-respondents. (2) the National Survey of Recent College Graduates, NSRCG. conducted by Westat has a target population of recent bachelor's and master's degree recipients--sample size of about 28.000. NSRCG is a CATI survey. (3) the National Survey of College Graduates, NSCG, conducted by the US Census Bureau has a target population of bachelor's and master's degree recipients prior to April 1990--sample size of about 216,000. The NSCG sample also covers PhD graduates from foreign institutions. The NSCG is a mail survey with CATI and PV follow-up to non-respondents. Mathetmatica Policy Research (MPR) provides a variety of technical assistance for the SESTAT surveys and associated redesign issues.

2/ National Academy Press (1989). *Surveying the Nation 's Scientists and Engineers: a Data System for the 1990s.* Constance F. Citro and Graham Kalton, Editors.

3/ More information on NSF work to revise the data system can be found in the American Statistical Association 1992 Proceedings of the Government Statistics Section---Session IV. pp. 78-100.

4/ The Occupational Employment Statistics (OES) Program in the Bureau of Labor Statistics (BLS) surveys employers (primarily establishments) by detailed Standard Industrial Classification (SIC) industry code. The survey, with a sample size of about 700,000 employers, is conducted over a three year cycle by State Employment Security Agencies, in cooperation with BLS. NSF contributes funding to the OES survey and publishes the data for S&E occupations in the surveyed industries annually. OES estimates will be incorporated into the SESTAT data system to provide industry and geography detail for S&E occupations.

5/ Prior to 1940, the "occupational" data collected in the decennial census reflected mostly industry/type of employer associations and occupations were not defined according to the work performed.

6/ A third occupational question on the respondent's occupational title was also used in 1970. This was later dropped in the 1980 census for space allocation and other reasons.

7/ The principal references used in coding occupation in Census surveys arc the Alphabetical (and Classified) Index of Industries and Occupations—Census publication number 1990 CPH-R-3 (4). For the CPS, electronic versions of these references are used.

8/ There are some estimation limitations to retroactively coding occupations not shown on the structured list--primarily due to non-sampling/reporting errors. This is to be expected since one of the functions of a structured list format is to give the occupations meaning through context and to have the respondent select among alternatives. These features are not present for occupations not shown on the questionnaire.

9/ We suspect this effect is stronger with higher degrees, such as PhD recipients, and with those closely associated with degree conferring institutions. Time since degree also seems to be a factor. It is difficult to test for these influences and, at this point, we have not done so.

10/ The Standard Occupation Classification (SOC) system was developed in the late 1970s by the Federal Statistical Policy Office (now at OMB) and the major federal statistical agencies collecting occupational data. The Census occupational systems and the Occupational Employment Statistics (OES) Survey conducted by the Bureau of Labor Statistics underwent conversions of their occupational classifications to become more compatible with the SOC in the late 1970s and early 1980s. The current SOC, 1980, is somewhat out of date, however, and plans are underway to revise the system prior to the 2000 Census.

References for the NSF SESTAT Surveys: 8/9/95

Hardy, L.P., (1992) "Balancing Multiple Demands: Sample Redesign for the NSF System." *Proceedings of the Government Statistics Section,* American Statistical Association.

Hardy. L.P. and Eisenhower, D. (1994) "Collecting Occupational Information from the College Educated Population". *Proceedings of the Government Statistics Section,* American Statistical Association.

Hardy, L.P., Mooney, G. & Eisenhower, D. (1995) "An Overview of Research for the NSF Surveys College Educated Population". To be published in *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Kasprzyk, D. (1992), Discussion for Session IV, *Proceedings of she Government Statistics Section,* American Statistical Association.

McGuiness, R., Owens, D. and Brick, M., Cahalan, M., Lapham, S.J. (1995). "Interviewer Variance in Two Telephone Surveys". To be published in *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Mitchell, S.. Mooney, G. and Shettle, C. (1994) "The Impact of Using Priority Mail in a Mixed Mode Survey". *Proceedings of the Government Statistics Section,* American Statistical Association.

Moonesinghe, R., Mitchell, S. and Pasquini. D. (1995), "An Identification Study of Nonrespondents to the SDR. To be published in *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Rak, B. and Cahalan, M. (1995) "Occupational Autocoding and New Branching Test Results", Westat Special Report.

Shettle, C. F. (1992). "What happens When Your Surveys Need a Radical Overhaul?". *Proceedings of the Government Statistics Section,* American Statistical Association.

Shettle, C. F., Mooney, G. and Giesbrecht, L. (1993), "Evaluation of Using Incentives to Increase Response Rates in a Government Survey". *Proceedings of the Government Statistics Section,* American Statistical Association.

Tremblay, A. (1995), "Nonresponse Issues of the National Survey of College Graduates". To be published in *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Tsapogas, J. and Gannon, J. *P.*, (1992) "Instrument Redesign: Trying to Make a Multitude of Users Happy While Developing a Quality Questionnaire. *Proceedings of she Government Statistics Section,* American Statistical Association.

Wilkinson, R. and Hines, C. J., (1992) "Data Collection Methods in the STPDS Surveys: Improving Response Rates in Mixed Mode Surveys." *Proceedings of the Government Statistics Section,* American Statistical Association.