CODING MAJOR FIELD OF STUDY

Larry G. Bobbitt and C. D. Carroll, National Center for Education Statistics
Larry G. Bobbitt, 555 New Jersey Ave.,Washington DC 20208-5652

KEY WORDS: Autocoding, fuzzy search, CATI

Abstract

The National Center for Education Statistics conducts surveys which require the coding of the respondent's major field of study. This paper presents a new system for the coding of major field of study. It operates on-line in a CATI environment and allows conversational checks to verify coding directly from the respondent. The system "learns" by maintaining a database of response/coding pairs which can be incorporated into its algorithm after supervisor review. This paper analyzes the effectiveness of this approach and database in coding major field of study for the **Beginning Postsecondary Students Longitudinal Study Second Followup 1990-1994**.

Introduction

NCES projects have frequently collected data concerning major field of study. Until recently, two procedures have been used to gather this data. First, experts in the *Classification of Instructional Programs* (CIP) examined respondents' text strings and assigned 6-digit codes from about 1,400 possibilities. Second, respondents selected 2-digit codes from a list of about 35 possibilities. The correspondence between the 2-digit and 6-digit codes was artificially high because the expert coders viewed both the text strings and the respondents' 2-digit selections. Inter-rater reliability within the expert pool was typically in the .80-.90 range after 3-4 days of training.

The advent of Computer Assisted Telephone Interviewing (CATI) enhanced the capability for obtaining higher quality text strings describing respondents' major fields of study. However, on-line codings into the 2-digits proved very time consuming. In addition, post-coding of text strings into CIP 6-digit codes delayed file delivery and sometimes resulted in expensive call-back procedures.

Furthermore, researchers were frequently critical of the 6-digit and/or 2-digit codings. Simply put, the 6-digit codes were too complex for analyses and the 2-digit codes did not provide adequate detail for analyses. One of the major critics developed an alternative 3-digit system with 111 possible codes based on patterns of courses described within *A College Course Map: Taxonomy and Transcript Data.*

Finally, as researchers become more sophisticated users of text string data (or as software improves handling of strings), the value of the high quality text strings becomes paramount in the data collection systems, including CATI. Coding becomes primarily a key for sorting or subsetting collections of text strings.

Methodology for Data Collection

All of the factors outlined above contributed to the development of a new approach for coding major field of study in current NCES CATI projects. The new coding approach incorporates on-line coding into our existing CATI system, using a 3-digit classification system. The existing CATI system executes the NCES major field of study coding software. The coding software then takes over all CATI functions for the major field of study question, and returns a response string and a 3-digit code to the existing CATI system. The CATI system then stores this data, and proceeds with the next question for the respondent.

The coding software takes care of prompting the CATI operator throughout the coding session. Initially, the respondent is asked an open question, "What is your major field of study", and the respondent's reply is entered into the coding

software. The coding software breaks up the response into words, and performs a fuzzy search for the words in the response.

The search is "fuzzy" in that

(1) the domain of the search is limited to words which have a similar initial sound.

(2) the object of the search is not only to find the target word (or determine that it is not present in the dictionary), but also to determine a short list of words in the domain of the search which are "closest" to the target word.

Since speed is obviously crucial for an on-line coding system, the search has been refined and "tuned" so that no discernable pause in the CATI operation occurs.

The software maintains a list of "reasonable" codes, initially empty. If a word is found in the dictionary, then all categories which are related to that word are added to the list. If the word is not found in the dictionary, the CATI operator is presented with a short (5 to 7) list of similar words. The words are ordered so that the words which are "closest" to the target word are shown first. The CATI operator can either select a word from the list, or ignore the entered word for purposes of coding. If a word is selected from the list of similar words, then all categories related to that word are added to the list of "reasonable" coding outcomes.

There are several reasons for constructing the list of similar words and having the CATI operator pick from it when exact matches are not found. First, and most importantly, misspellings are extremely common, and this is a quick and efficient way of dealing with them. In most cases, a misspelling will result in the correct spelling being shown as the first word in the list. Secondly, the dictionary for the most part contains root words only. As a general rule, the dictionary does not contain multiple variations of the same word. For example, it would be inefficient to store the words MATH, MATHEMATICS, MATHEMATICAL, etc. when they are all coded to the same major. In some cases, when the distinction helps to indicate the major field of study, multiple variations of the same word are included in the dictionary. For example, the word CHEMICAL is a entry in the dictionary and maps to the major CHEMICAL ENGINEERING, while the word CHEMISTRY contains no engineering associations. While we have experimented with various algorithms to adjust for suffixes, prefixes, etc., the current approach avoids having algorithmic (and possibly wrong) determinations of word variations.

For example, suppose the a CATI operator enters the string "Decison Information Sciences" for a respondent's major field of study. The system converts everything to capital letters, breaks the string into three words, and proceeds to look up the word DECISON, for which no exact match is found. The system presents a screen with a list of possible matches for the input word. The word DECISION tops the list, and is selected by the CATI operator. The program looks up the other two words, and since they are both exact matches for dictionary words, no action is required by the operator.

Once all the words in the response are processed, the set of "reasonable" codes are presented in increasing order of likelihood. The CATI operator may then select one of the "reasonable" codes, override the "reasonable" code list and select a code outside the list, select "uncodeable", or reenter/edit the initial response. The operator is trained to discuss the coding process with the respondent, possibly resulting in changing the initial response to one which more clearly identifies the respondent's major.

Online coding encourages CATI operators to identify shortcomings in the collected text strings while still discussing the respondent's major with the respondent. When such shortcomings are identified, the CATI operator can immediately elicit additional detail in the responses, without the expense of later callbacks. Coding of responses is secondary, the main object of our approach is to

improve the collected text strings.

Returning to our example, the software would find that the word DECISION is associated with only one major field of study, Business/Management Systems. The word INFORMATION is associated with three major fields of study: (1) Business/Management Systems, (2) Computer Programming, and (3) Computer and Information Science. The word SCIENCE is associated with 26 different major fields of study, including Business/Management Systems and Computer and Information Science. The software constructs a screen listing all possible major field of study codes in order from most to least likely. In this case the most likely code is clearly Business/Management Systems so the CATI operator would select that code and continue with the interview.

In our example, the word DECISION tips the scale favor of the code for Business/Management Systems. If the reply had instead been "Information Sciences", then two codes (Business Management Systems, and Computer and Information Science) would have been tied for most likely code. In this case the CATI operator would have probed the respondent for more information. If, for example, the major is in the Department of Computer Sciences this discussion with the respondent will probably result in the correct classification under Computer and Information Science, not under Business/Management Systems. Or perhaps, under probing, the respondent clarifies that his/her real major is really "Decision Information Sciences" in the School of Business. Using a more traditional offline coding approach, we receive only the string "Information Sciences" and must make a judgement which may or may not be correct.

Word Association DataBase

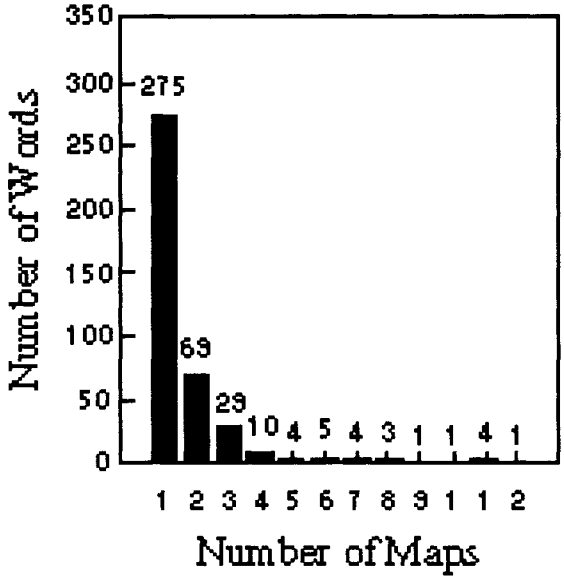
The approach outlined above presumes that we have specific information about how responses map to codes for major field of study. Fortunately, some information on this is available from previous surveys. We know what strings were collected on the previous surveys, and what major field of study was finally coded for each of those strings. Starting with this information we constructed a database of word to code mappings, using our judgement about whether to delete or preserve links between particular words and codes. The final database structure consists of an index of words related to their associated codes, so that for any word the associated codes can easily be identified.

The coding system has been designed so that it has the ability to "learn" new words or new codes for existing words in a controlled way. Because of the importance of the database to the coding system, updating of the database in this way is normally performed by very experienced personnel. The response string, assigned root words, and final major field of study code are gathered and reviewed. In addition, the system indicates what words in the response string were or were not associated with root words, and whether the CATI operator overrode the set of "reasonable" codes presented by the coding system. A computer program identifies potential new database words, and potential new codes to be associated with existing database words. As new database entries are identified, they are presented on a screen and can be added or not added to the database. The update system only presents words or word/code combinations that are (1) not already present in the database and that (2) have not previously been refused entry into the database by the operator.

There are 739 word to code maps in the current word association database. The database has 406 unique words and 113 unique codes. The distribution of the number of word/code maps for the unique words in the dictionary is shown in Figure 1. Figure 1 clearly shows that most words are associated with only a very few major field of study codes. Also, the number of words in the database quickly decreases as the number of maps increases so that very few words map to more than three or four codes. However, as shown in Figure 2, the distribution of the number of word/code

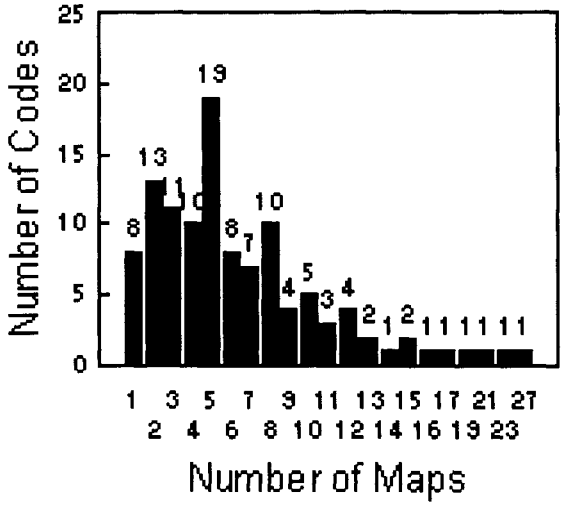
Figure 1.
Distribution of Word Mappings

shows that there are some words which map to a unique code, for example the word "philosophy" maps only the major field of study code for "philosophy".



maps across codes is quite different. Although generally speaking the number of codes in the database decreases as the number of maps increases, there are some exceptions to this general trend. Most of the codes have more than three or four words which map to them. Figure 2 also

Figure 2
Distribution of Code Mappings



Coding Responses

To evaluate how our approach was working, we used our software to review 1427 actual CATI responses. These were responses which our CATI operators collected during the field test for the ***Beginning Postsecondary Students Longitudinal Study Second Followup 1990-94*** (BPS:90/94). None of these test responses were ever used to construct or update the database.

Table 1
Number of "Most Likely" Codes Identified

Number of Most Likely Codes	Number of Responses	Percentage of All Responses	Cumulative Percentage of All Responses
1	545	38.2	38.2
2	0	0.0	38.2
3	258	18.1	56.3
4	73	5.1	61.4
5	2	0.1	61.5

The test responses included 2237 words, of which 2057, or 92%, were found immediately by the computer system in our database. The remaining 180 words were reviewed by an operator who identified 95 of them as misspellings of database words. Only 85 of the 2237 original words were not found in the dictionary, a success rate of about 96%. Most of the words which one typically encounters when asking about a respondent's major field of study appear to be included in the database.

Of the 1427 raw responses, 272, or 19%, mapped to a unique major field of study code. An additional 273 responses had one code which was mapped to by a majority of the words in the response. Although in practice we have the CATI operator review the final code, these 545 responses, or 38% of the total, have a single most likely code and are essentially autocoded by our

system. Of the remaining responses, our approach ranks all codes by the number of word/code pairs found in the dictionary for the response. Using a definition of "most likely code" to mean that the code is tied for the maximum number of word/code pairs found, Table 1 shows the distribution of these responses by the number of most likely codes.

It is important to recognize that for even the remaining responses, the effect of using our coding system is that better information is obtained than would have been obtained otherwise. For example, the initial input word "anthology" was confirmed to be a misspelling by a CATI operator and was identified as the word "anthropology". This immediately clarifies what the major of the respondent is. Under a more traditional offline coding approach it is problematic whether a major of "anthology" would have been correctly coded.

Table 2
Results by the Number of Words in the Raw Response

Number of Words in Raw Response	Number of Raw Responses	Number of Responses With All Words Found	Number of Responses With a Single Most Likely Code
1	656	614	210
2	675	611	275
3	77	68	44
4	16	9	13
5	1	0	1
6	2	0	2
Total	1427	1302	545

Table 2 shows the distribution of our 1427 raw responses by the number of words in the response. The third column of Table 2 shows the number of responses which had all of their words found in the dictionary (possibly with some help from the CATI operator). Over 90% of all the responses contained only one or two words. All the words in the response were found in the dictionary for 1302 of the 1427 responses, a success rate of 91%.

As mentioned above, only 545, or some 38% of the responses were linked to a single most likely code. Even of the 656 responses which consisted of a single word, only 210 (32%) were linked to a single most likely code. The reason for this difference is that many words simp descriptive enough to allow one to narrow down the possibilities to a single code. An example is the single word response "EDUCATION", which could be one of the following major field of study codes:

Early Childhood Ed
Education: Not Phys. Ed.
Elementary Ed
Health/Phys Ed/Recreation (HPER): non-school
Interdisciplinary: all other
Physical Education
Secondary Ed
Special Education

Preliminary results indicate that our new approach is on average taking only a few seconds longer than a simple approach of asking for the major and recording the response. There is also slightly higher cost for training relative to previous approaches. CATI operators must understand how the coding system works. They must recognize what to do when words are misspelled or are not in the database.

Conclusions

Our test results on actual CATI responses can be divided into three groups. Approximately one-third of the responses can be coded with extremely little intervention by the CATI operator. For almost another third, the approach does extremely well and reduces the number of possible codes from 110 to less than five. For the final third, the approach could be characterized as somewhat helpful. It appears that the benefits of correcting spelling errors and of attempting to code the response while the respondent is available for clarification merits the small amount of additional time required for the new coding system.

References

Adelman, C., (1990), *A College Course Map, Taxonomy and Transcript Data*, Washington, DC: U.S. Department of Education.

Knuth, D.E. (1973), *Art of Computer Programming, Volume 3*, Reading, MA: Addison-Wesley.

Morgan, R.L., Hunt, E.S., and Carpenter, J.M.(1991), *Classification of Instructional Programs*, Washington, DC: National Center for Education Statistics.

Wirth, N. (1986), *Algorithms and Data Structures*, Englewood Cliffs, NJ: Prentice-Hall.