DOCUMENT RESUME

ED 065 587 TM 001 856

AUTHOR Hagerty, Michael P.; And Others
TITLE Development of Software.
PUB DATE Apr 72
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Compensatory Education Programs; *Computer Programs; *Data Processing; *Information Storage; *Information Systems; Magnetic Tapes; Surveys; *Tables (Data)
IDENTIFIERS *Survey of Compensatory Education

ABSTRACT

This description of software development for the 1970 Survey of Compensatory Education analysis contract is presented in three sections: (1) specifications for data and output, (2) descriptions of the programs written, and (3) problems encountered while processing the data. The section on Specifications contains the description of the data to be supplied by USOE and the product to be produced by the project staff at the University of Massachusetts. The second section describes the operation of the programs which were written to produce the 3,000 crosstabulation tables. In the third section, problems encountered both prior to and after the arrival of the data tapes from USOE are described. (DB)

DEVELOPMENT OF SOFTWARE

by

Michael P. Hagerty
Frederick A. de Friesse
Stephen Edberg

University of Massachusetts

A paper presented to the 1972 Annual Meeting of
the American Educational Research Association
Chicago, Illinois, April 7, 1970

Symposium on DATA ANALYSIS OF THE 1970
ELEMENTARY SURVEY OF COMPENSATORY EDUCATION

The description of software development for the 1970 Survey of Compensatory Education analysis contract is written in three sections: (1) specifications for data and output, (2) descriptions of the programs written, and (3) problems encountered while processing the data.

Specifications

This section contains the description of the data to be supplied by the Office of Education (USOE) and the product to be produced by the project staff at the University of Massachusetts (UMass). In order to facilitate a clearer understanding of the progress of the project, it is necessary to present this portion in historical fashion.

At the outset of the project, a basic decision was made. This analysis was not to be of the same type as the one conducted by Glass, et al., the previous year (i.e., this project was to produce a series of commented, interpreted cross-tabulation tables, rather than a series of different statistical analyses). It was further decided that the tables produced would be in response to a series of basic policy questions developed by USOE. This series of six to ten questions would cover the entire range of what the Office of Education expected to see from the data. It was anticipated that these questions could be answered in less than 5,000 tables.

Therefore, the project was cleared and USOE was to provide UMass with the data to be analyzed in what will hereafter be referred to as "standard form" (seven-track magnetic tapes recorded unblocked in even-parity BCD at 556 B.P.I. with no parity errors) prior to November 15, 1970.

This deadline was not met, and a meeting was called for December 16-17, 1970. Present at this meeting were representatives of both UMass and USOE, as

well as several outside consultants. The purpose of this meeting was to clarify the conditions and responsibilities defined in the contract concerning the specific dates and products which USOE wanted. At this meeting several points were made, and most of the major details concerning specifications were worked out. The decisions reached at this meeting were as follows.

USOE would provide UMass with the data to be analyzed in standard form by the end of January. The data tapes were to be turned over to the contractor with the weights certified, formats accurate and all appropriate edits made. In return, UMass would provide USOE with approximately 500 tables within six weeks for interpretation and commentary. This submission of tables was scheduled to occur prior to the time USOE was required to report to Congress on the program.

Concerning these tables, it was stated that USOE would provide UMass with the complete descriptions of the variables to be used within one week after the data tapes were accepted. There would be fewer than 100 variables used in these initial tables, and these would be the raw variables straight from the data tapes. The tables were to be two-dimensional and weighted, of 30 by 35 maximum cell size. All the tables were to be run so that both vertical and horizontal variables in a given table were from the same data file whether this file be the pupil, teacher, principal, or district file. Each cell would contain frequency, row, column, and cell percents; each table would contain appropriate marginals. Every variable was to be identified with a 24-character label, and each value of a variable was to have an eight-character code identifier.

When completed, the tables would be presented to USOE for commentary and interpretation through consultation with the UMass team and the USOE editors in 8-1/2 by 11 inch reduced format. The maximum size table that would appear on any page was nine by twelve. Any tables larger than this would be continued on

additional pages, successively numbered. No tests of statistical significance would be applied.

Following this initial report to Congress, a series of approximately 2,500 tables were to be produced in the same format. USOE would provide the description of each table desired (using mostly raw variables crosstabulated with several created composites) prior to the time any of these 2,500 were to be run. The anticipated deadline for these tables was the middle of summer.

The final portion of the contract would be settled with the submission of a technical report describing any difficulties encountered, with suggestions for future improvements if any.

Within several weeks after this meeting, a major reorganization of the project monitoring staff in Washington necessitated a total reconceptualization. The initial report was discarded, although 3,000 tables were still to be produced. Now, however, the maximum size of each table was to be 60 by 65. The number of dimensions, the manner in which weights were to be used, and the selection of the variables remained unchanged.

By late June, when the first useable data tapes arrived, several of the original specifications had changed. Instead of the use of raw variables predominantly, the decision was to use created (or constructed) variables almost exclusively. This necessitated that each variable to be constructed would first have to be completely defined. In addition, the tables were to be specified in piecemeal fashion rather than all at once, requiring many revisions of programs and constant reruns. In many cases a variable was defined and created to be used in a table but was later discarded unused, because the process of variable formulation often preceded the designation of tables and, as information was developed, information needs changed in priority.

The new policy required that the four separate data files be merged into one so that one could read across a response at a weighting level different from the level at which the response was made. A change of this type, apparently quite simple, required massive efforts in program reorganization to meet specifications. The exact details of the programming will be described in the next section.

Programming Considerations and Logic

This section of the report will describe the operation of the programs which were written to produce the 3,000 crosstabulation tables requested by USOE. The two components of this section are: (1) a rationale for the choice of programming logic, and (2) a description of the programs used. In order to facilitate clarity, the rationale will be presented first.

The University of Massachusetts has a CDC 3600 and a CDC 3800 available for on-campus research activities. Both of these machines are late second-generation equipment which feature large quantities of core (128,000 48-bit words). The fact that these machines are fixed-word machines was the most important single consideration, as there were large quantities of data to be used and data packing became a necessity. These machines are also tape-drum machines; transfer rate is good but drum storage is small (approximately 500,000 words). Since the equipment is late second-generation, it is somewhat slower and more prone to system or hardware failures, which made recovery procedures quite important. Finally, the operating procedure at the University of Massachusetts requires users to submit programs for execution as opposed to "hands on" operation, which meant that the programs had to be self-contained.

Also influencing the programming approach were the characteristics of the data. The data from USOE, when considered at the pupil level, was over 1,000

pieces (not bits or bytes) of information per pupil. Since there were over 84,000 pupils, the total data set was over 84 million pieces of data. This large quantity of data was difficult to store and to access rapidly.

The data also came in various forms. There were district, principal, teacher, and pupil questionnaires, each type on a separate tape. There were also update card decks and update tape files. This large group of data had somehow to be handled as a logical group. Since there was a need to cross the four types of questionnaires by identification codes, some sort of programming logic had to be developed which would allow access to all four files simultaneously.

Output format also affected the type of programming logic used. Since the crosstabulation tables to be generated were at times very large (60 by 65 levels with four entries in each cell) and often with certain controls such as adding or deleting specific information, the program logic had to allow for flexible sample size.

When these characteristics were considered, it was decided that it would be best to create a single file which attached the associated teacher, principal and district information to each pupil record. This approach would eliminate the problem of matching files later and would speed access time when creating the actual crosstabulations.

Since the data from USOE was character-packed and the CDC equipment was word equipment, the data had to be unpacked, or decoded. Decoding data is a time-consuming operation on a computer, and once decoded the data would require six to seven times the storage space required for the original data. Because of these two considerations, it was decided to decode and save only specific variables.

The decoded specific variables were then stored on another set of tapes with newly constructed variables which were created by making various types of combinations of original variables.

In order to use the computer time most effectively, as many crosstabulations as possible were run simultaneously. This decision called for an independent program which would take the original and constructed variable file and run specific crosstabulations on this file, dumping the crosstabulated output on another tape for later printing. The combination of parallel crosstabulations with serial data presented core problems which were only controlled by removing all but the essentials for crosstabulations. Thus another program was required to convert the raw crosstabulations into final tables with headings, labels, marginals and cell percentages.

In the process of transforming the data from the tapes provided by USOE into the crosstabulation tables desired, a sequence of five steps was followed: (1) the unblocking of the data, (2) subsequent decoding of the data, (3) the creation of run tapes, (4) the crosstabulation itself, and (5) the final printing sequence. These steps will now be described in detail.

Initially, the data from USOE arrived in separate files. Because of the decision mentioned earlier that all files would be merged together, the first process was to unblock the identification codes for each file (Program Code) and merge the files together into a single run file (Program Merge). The program to perform the file merging had several problems. First, the data was packed; second, it was blocked. It had to be unblocked and unpacked so that it could be matched with the other files. This problem was overcome by reading a physical record into core, unblocking strings with the I.D. codes decoded, and writing these strings out to a temporary drum unit. When enough of these strings were

on the drum unit to fill it, the strings were read serially back into core and matched against their corresponding file. This process represented the use of large amounts of computer time but simplified file matching problems. Two files were merged at once: first the district and principal files were merged, giving each principal a district record; next the combined file was merged with the teacher file, thus giving each teacher a principal and a district; finally the resulting file was merged with the pupil file. This process alone required 120 minutes of computer time and produced a merged file three tapes in length.

The merged file then had to be unpacked, or decoded (Program Decode). On a word machine this is another time-consuming task, so that only those pieces of data which had to be were unpacked. These specific pieces of data were then repacked in an integer word form, which on a word machine is much more rapidly unpacked. The time required to decode the original tapes was eight and one-half hours and produced nine tapes of output.

From the specific original variables saved, another program (Program Kludge) was written to create a run tape containing the original variables and a variable number of created new variables. Each new variable was created by a subroutine call to keep the logic more readable. These new variables and the desired original variables were also integer word packed to conserve space and to increase the speed of processing. This process required eight hours of computer time and produced four reels of tape as a single run tape.

At this point a crosstabulation routine (Program Cross) which allowed for a variable number of crosstabulations up to a total number of 32,000 cells (two 16,000 cell arrays) was employed. The crosstabulation program was able to handle multiple crosstabulations by using index pointers. For example, a five by six crosstabulation required 42 cells (zero responses were included); there-

fore it required a vector of 42 cells. Crosstabulations were placed back-to-back until the 16,000 cells available were used. To address a given crosstabulation, its starting point was taken from an address array and the particular cell was addressed by calculating the number of positions it was down the array.

This program would read the integer word packed variable tape and handle up to 500 individual crosstabulations per logical record. This method cut down the number of times the file had to be accessed and thus cut down the computer time. Program Cross initially read in the labels required for the headings of the tables and the headings of the levels, and stored these on the output tape, the remainder of the tape being filled with the crosstabulation tables themselves. This process, although efficient, required 70 to 90 minutes per pass, each pass producing between ten and 400 crosstabulation tables simultaneously.

Finally, a program (Readem) was written to take the raw crosstabulation output tape and break the two 16,000 cell arrays into the individual crosstabulations, compute the marginals and cell percentages, label the tables and print them out. This program as well as Program Cross kept track of the individual crosstabulations by a series of core-stored pointers. This routine was quite rapid, requiring only about ten minutes to produce a series of ten to 400 crosstabulation tables.

In passing it may be noted that several times, as errors were found in the data or as update decks were supplied by USOE, the entire process had to be followed through from the beginning, including another program (Fixit). This required several weeks of set-up time and reorganization.

If the described process sounds overly complex, perhaps it was. The use of a machine with a large amount of disk storage and byte addressing would have eliminated many of the merging and decoding problems. However, the large core

on the CDC machines did allow for very efficient crosstabulations and therefore offset the time element somewhat.

Problems

This section deals with various problems encountered both prior to and after the arrival of the data tapes from USOE. Problems occurred both in general and specific to the programming and running of data.

Problems which severely affected deadlines and prevented the necessary tape editing and checking were the late arrival of the data tapes, the fact that the tapes never met specifications agreed upon by UMass and USOE (although the final set was readable on our system), and the fact that some of the data was bad or missing and had to be corrected.

The computer systems available for University of Massachusetts research activities are, as previously mentioned, CDC 3600 and CDC 3800. To review, these late second-generation systems feature great processing speed and large amounts of core, but tend to be very slow on input/output functions. The 3600 has only one input/output channel available at any one time, and the 3800 has only two. This severely limited processing times as the nature of the processing required the transfer of large quantities of data. In addition, the equipment was prone to hardware failure, a problem which necessitated restart procedures: the records transferred in the analysis process were long and the processing time was long, both factors sometimes resulting in equipment failure.

Since the equipment sometimes failed after several hours of running, only two hours of processing were done at any one time on both the decoding and run tape procedures; these programs were then restarted. In addition, the filing system for tapes at the University of Massachusetts Computer Center is such that tapes being used for the analysis contract were sometimes destroyed through acci-

dental reuse by personnel outside the project; a second filing system problem involved parity errors due to poor handling or dust. As a result, any tapes representing large amounts of processing time were copied for backup tapes.

In addition to hardware limitations, the programs had to be submitted on a "closed shop" basis for processing, along with other University research and teaching jobs. This meant that perhaps only two to three hours of computer time could be granted to the 1970 Survey of Compensatory Education project on any one night, and sometimes none at all. The Computer Center did, however, occasionally place the computer systems at our disposal when the Center was normally closed. This cooperation helped us meet already pressing deadlines.

Problems were also encountered with the actual programming and running of the programs. The first processes in preparing the data for analysis were to unblock and decode the identification codes and merge the four files (district, principal, teacher, and pupil) on these identification codes. At times there was, to use an example, information for a principal but no corresponding district information, forcing us to fill the district portion of the file with zeros. In other cases, a teacher record might be missing for some district-principal records, resulting in a teacher-principal-district record with zeros filling the teacher portion of the identification code. In addition, some records were out of order and, since the merging process considered only a limited number of records from the two files to be merged (see previous section), the non-sequential record was transferred to the end of the merged file. Elsewhere in the merged file the counterpart of this non-sequential record would contain a zero portion since the misplaced record would be considered missing data. When the next set of files was to be merged, there would be no records to match this partial record. The number of partial records, however, was less than one-

tenth of one percent of the entire number of records in the final file.

The merged file then had to be unblocked and decoded so that our machine could read the information, after which packing was required because of the large amount of information contained in each record. In unpacked form, each record would have required six to seven times as much space for handling and storage. The packing procedure also increased input/output speed, since the data was compact and more could be transferred at one time in this compacted form.

When the analysis contract was awarded in September, 1970, three algorithms for packing data were tried before a workable algorithm was developed. Since the CDC machines are word machines, it was necessary to pack many pieces of information into one word rather than allowing one word for each separate piece of information. Due to software problems in the function for packing data into bits, the first method attempted was unsuccessful. Next an attempt was made to put one piece of information into each byte of our eight-byte words; this process failed when we were unable to recover information from the last byte. Finally, we used a process we called "integer packing," in which each variable was multiplied by some factor of ten and added to the current word. This process reduced the space required to store the information to nearly one-seventh.

Once the original data was readable on our system and the data was arranged such that it could be transferred and stored more easily, certain variables had to be stored on a separate tape, called a run tape, along with variables created from data on the original file. Due to the limited amount of available core, only 500 variables (raw and created) could be stored for each pupil. This fact necessitated the creation of several run tapes. Since the transfer of data on CDC machines is relatively slow, run tape creation represented a large amount of computer time. The information needed to make the run tapes had to be packed to

decrease space and improve data transfer; the run tape file, when packed, was approximately half as long as the original data file.

For the sake of order and readability each constructed variable was coded as a separate subroutine. Some run tapes called as many as 300 subroutines, which necessitated the nesting of the cells in the run tape program as our system will allow only a set number of call statements. This number of subroutines, when punched onto cards, represented 6,000 cards. Card readers sometimes made errors in reading this large quantity; as a result, binary tape programs were made from all programs to eliminate the possibility of card reader mistakes.

Once the run tape was created, the required crosstabulations of specified variables could be done. The data had to be unpacked before being crossed, and the crossed data was stored in a separate array for later processing. This program was changed many times prior to the arrival of the tapes, due to the indecision of USOE and UMass as to the exact limiting specifications of the crosstabulations. Finally, however, the program was able to create a variable number of tables, all totalling 32,000 cells, with a maximum table size of 60 by 65.

The final process in the crosstabulations was to create the tables. This program was modified many times along with Program Cross, as specifications were changed again and again. The table printing process was very efficient, requiring only ten minutes to process 32,000 table cells. Since the printers at the Computer Center were also prone to failures (such as a stuck key or paper jams), the program could easily be run again from the file created from Cross.

In summary, the three greatest problems encountered were (1) delay in the shipment of data from USOE, (2) necessity to edit and check the data even though this process was to have been completed by USOE prior to the time the tapes were

shipped, and (3) lack of rapid access equipment at the University of Massachusetts Computer Center. Since the data was assumed, erroneously, to be complete and correct, problems arose later as errors were discovered which required substantial delays in processing. The slow transfer rate on our system necessitated both previously working programs and useable data. Although the programs were working, the data was sometimes unuseable. Backtracking and checking data in the midst of the processing procedure often resulted in several weeks of delay--delay which could and should have been avoided.