DOCUMENT RESUME

ED 065 588 TM 001 857

AUTHOR Proper, Elizabeth C.
TITLE Variable Conceptualization: A New Route to Policy Questions.
PUB DATE Apr 72
NOTE 13p.; 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; *Data Analysis; Information Processing; *Questionnaires; *School Surveys; *Statistical Analysis

ABSTRACT

The variable conceptualization process used in the analysis of data of the 1970 Compensatory Education Programs is discussed. Variables used in the process were designated as being either "univariate" or "constructed." Examples are given of the procedure used in the development of variables. Procedural issues handled in the process of conceptualizing and developing the constructed variables were: the handling of non-respondents, multiple-response items, the multivariate nature of some variables, the population level at which the particular variable should be run, the type of population over which the variable should be broken, the sequence of embedding, and the types of checks necessary to ensure accuracy of the variable development. Each of these issues is discussed. Several problems that arise as a result of attempting to provide a data analysis which will furnish useable information to program personnel are briefly discussed. (DB)

30.4

VARIABLE CONCEPTUALIZATION: A NEW ROUTE TO POLICY QUESTIONS

by

Elizabeth C. Proper
University of Massachusetts

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

Symposium on Data Analysis of the 1970
Elementary Survey of Compensatory Education

Introduction

In the planning stage for the data analysis of the 1970 Compensatory Education Programs, it was decided that a major emphasis should be placed on meeting the needs of program personnel. The contractor who was to do the data analysis and his U.S.O.E. program officer agreed that a major responsibility of the contractor lay in the area of providing data which would both be useful to and actually used by program personnel in addition to meeting the requirements of a report to Congress. It was considered to be important to provide a data analysis which would serve program personnel in two major ways: (1) provide data in a format which would answer the general questions which program personnel specified prior to the analysis, and (2) provide this data in such a way that program people would potentially be able to answer specific questions during the year at the moment in time that those specific questions arose.

Generation of Policy Questions

As a first step, the contractor's program officer established a task force of program personnel to generate a list of their concerns and information needs. After this list was developed, Office of Education personnel under the direction of the program officer generated six policy questions based on the specified

needs of the program personnel: (1) How well are federally-funded programs meeting the needs of the participants in these programs and are there sufficient programs provided to meet the needs? (2) How effectively are the regular and federally-funded sources and programs producing changes in the participating populations and can these changes be attributed to participation in federally-funded programs? (3) Are academic services and programs being concentrated on the neediest populations rather than being used for a large population with varying degrees of need? (4) Are non-academic services and programs being concentrated on the neediest populations rather than being used for a large population with varying degrees of need? (5) What lines of communication have been established between the community, parents, and the school district and particularly are community advisory groups being used during the development and evaluation of federally-funded programs? (6) How effective is inservice training being used to agument the purposes of federally-funded programs?

Through consultation between the program officer and the contractor the conclusion was reached that answers to the policy questions were unobtainable through direct analysis of the available data. Because of the problems created by the type of sample through which the data was gathered, by the absence of bridges between data sources, and by the complexity of programs being studied, a decision was made to hold the statistical analyses of data at the greatest level of simplicity and to construct new variables from the data that maintained a simple one-to-one relationship to the type of information needed. Hence it was decided to construct new variables couched in terms of the questions to be answered and to answer these questions through crosstabulations.

Terminology

Before examining the procedures used in the variable conceptualization process, we need to establish a common frame of reference. In this report the

term "variable" is used quite frequently. These variables are normally of a categorical or discrete nature. When we speak of "levels" we are referring to the various categories over which the variables are allowed to vary. The data in the form provided to the contractor by U.S.O.E. is referred to as "univariate" data or simply as "univariates." Quite often when we used two or more of these univariates to form a new variable we did so by "embedding" one variable into another variable. For example, one might wish to combine grade taught with degree status of teacher. Our sample only contained three grades: two, four and six. If we were to embed the degree status univariate into the grade taught univariate, we would have a twelve-level variable: the first four levels would involve grade two; the second four levels, grade four; and the third four levels, grade six. For each of these three groups of four levels we would have four types of degree status: no degree, bachelor's, master's, or master's plus. Thus, because we would be running through the degree status univariate three times while we were running through the grade taught univariate only once, we would call the degree status univariate our "fastest moving" variable and the grade taught univariate our "slowest moving" variable. We came to talk of our variables as being either "univariate" variables or "constructed" variables, depending upon whether we used them in our tables as they appeared in the data provided by U.S.O.E. or whether we reworked them in various ways before using them.

An Example

In order to illustrate the procedure used in the development of variables we shall look at the process which was involved in attempting to define need, which is one of the recurrent aspects of the policy questions.

One variable was developed at the school level which dealt with the percent of pupils in the school whose families' major means of support was a welfare

program and the percent of sixth grade pupils in the school who were reading one or more years below grade level. Each of these areas of concern was handled in questions on the principal questionnaire; it was expected that such information would give some measure of economic and academic disadvantagement within the given school population. This variable was also broken over school type, whether the school was defined as a Title I or a non-Title I school. On the questionnaire, both the welfare and the reading questions allowed for one response to be given out of seven choices; the Title I participation component was a two-level question. The constructed variable would have contained ninety-eight levels. We were able to generate variables up to sixty-four levels. After looking at the previously completed tabulation of the questionnaire items, it was determined that it was possible to reduce the size of the welfare question to six categories and the reading question to five categories, thus generating a sixty-level variable. It was not desirable to generate such large variables if vital information would not be lost, for two reasons: (1) it is exceedingly difficult to interpret variables which have a large number of categories, and (2) a large number of categories often means that some categories contain very small numbers. For large variables we often developed sister variables without the title break for crossing with variables which contained a title break.

A second variable developed to handle need was built at the pupil level. This variable sought to define pupil academic disadvantagement across Title I and non-Title I schools. In addition to the Title I participation component, it used two multiple-response pupil questionnaire items; the first, a question dealing with pupil needs, was a "mark all that apply" seventeen-option question; the second, a question dealing with target group membership, was a "mark all that apply" ten-item question. These questions, P-11 and P-12, may be found in Figure 1. A fourteen-level variable was developed with seven levels each for

the Title I and non-Title I populations. It was desired that degree of disadvantagement should be presented in this variable. The seven categories which were developed are presented in Figure 2.

FIGURE 1

P-11: In your professional judgment, which of the following creates, or is, a persistent problem for this pupil? (Mark all that apply.)

a) Low achievement in mathematics
b) Low achievement in reading
c) Low achievement in English language arts
d) Poor vision
e) Poor hearing
f) Speech defect
g) Other psychomotor deficiency
h) Physical handicap
i) Chronic disease
j) Mental retardation
k) Social immaturity
l) Emotional problems
m) Anti-social behavior
n) Malnutrition
o) Family instability
p) Other
q) This pupil has no persistent problems

P-12: Should this pupil be classified as any of the following? (Mark all that apply.)

a) Academically disadvantaged pupil
b) Socio-economically disadvantaged pupil
c) Academically gifted pupil
d) Pupil from home where the dominant language is not English
e) Potential dropout
f) Emotionally or mentally handicapped pupil
g) Migrant pupil
h) Neglected or delinquent pupil
i) Physically handicapped pupil
j) None of the above apply

FIGURE 2

Category:

1. A-D1 (Highly eligible. Mark in P-11a,b or c, and mark in P-12a,f,g, h or i)
2. A-D2 (Eligible 1. Mark in P-11a,b or c, and mark in P-12b,c,d or e)
3. A-D3 (Eligible 2. All others having mark in P-11a,b or c)
4. A-D4 (Maybe 1. Mark in P-11d-p, and mark in P-12a,f,g,h or i)
5. A-D5 (Maybe 2. Mark in P-11d-p, and mark in P-12b,c,d or e)
6. A-D6 (Maybe 3. Some mark in P-11d-p)
7. A-D7 (Unclassified)

The types of problems which exist in the handling of non-response are inherent in this question. All variables have a non-response or missing column to handle that portion of the population which cannot be categorized. In this particular variable there is also an unclassified column for each of the Title I participation populations. If you look at the various levels involved in this variable you will note that a pupil could have had either one or both of these questions answered and yet fit into none of the first six categories; or a pupil might have only P-11 answered and potentially fit into Level 3 or Level 6. Obviously if both questions were answered and he fit into one of the first six levels, he was so classified. If he had both questions answered and he fit into none of the first six levels, he was placed in unclassified. If he had neither question answered, he fell into non-response--or did he? He might fall into unclassified if we could place him as a Title I or non-Title I school participant, which we could since those populations were completely defined. In terms of programming and computer time, it would have been a much more rapid procedure if we could have classified those for whom we had a record of positive incidence and simply dumped the remainder either into unclassified or non-response without

bothering to check each question completely, since for positive response we were only dealing with a portion of the complete questions. While a first reaction might be that the amount of computer time necessary to make such a check is inconsequential, it must be remembered that we were dealing with a pupil sample that ran over 84,000 and we had upwards to 500 variables, counting both univariates and constructed variables, on our tapes at all times. The procedure developed for handling non-response for this variable was that we checked the seventeen locations of P-11 and the ten locations of P-12 and that we assigned any pupil who did not respond in some manner to both P-11 and P-12 to the non-response column. He was assigned to the unclassified column if he answered both P-11 and P-12 and was not previously assigned to one of the first six levels. Procedures of this type for the handling of non-respondents were developed for each of the other constructed variables.

Need was handled in a number of other variables. One examined degree of pupils' economic disadvantagement and was built from components of six different questions, two of which were once again "mark all that apply" items. It sought to define economic disadvantagement based on target group membership, pupil need, unemployment, welfare and low income. Another need variable was built around the welfare question and the percent of pupils in the school who received free or reduced-price food at lunch. One variable was developed which handled both academic needs and academic participation.

Procedural Issues

A number of procedural issues had to be handled in the process of conceptualizing and developing the constructed variables: the handling of non-respondents, multiple-response items, the multivariate nature of some variables, the population level at which the particular variable should be run, the type of population over which the variable should be broken, the sequence of embedding,

and the types of checks necessary to ensure accuracy of the variable development.

Multiple response items. The procedure used for handling non-respondents has already been discussed in this paper. The handling of multiple-response items has been alluded to in terms of the type of problem which they presented in determining non-respondents. Many of the questions which were used in the constructed variables were multiple-response items. Each multiple-response item actually contains as many discrete questions as there are components within the given question; any respondent may have answered from one to all of these components. When it was feasible, the variable was developed so that each of all possible combinations was separately categorized. An example of this is in the handling of a district questionnaire item which contained five options regarding services provided by regional centers. The fifth option, stating that no pupils were served in this manner, was distinct from the other four. However, if the fifth option was not marked, either one, two, three or all four of the other options might have been. A sixteen-level variable was built which assigned non-respondents to missing, respondents to the fifth option to the first level and then each of the other fifteen possible combinations to fifteen separate categories.

Most of the multiple-response questions, however, had more options than could be handled in this fashion. In those cases, various options were combined, only certain options were used, or a series of variables were constructed each covering a certain grouping of options.

Multivariate nature of variables. The multivariate nature of certain variables is perhaps most complexly manifested in the variable which was designed to present the extent of improvement in achievement. Achievement was immediately restricted to reading. Four separate variables were developed: one for grade two, one for grade four, one for grade six, and one which combined the three

grades. Nine discrete categories were developed based on whether the pupil's learning rate prior to a pre-test was less than a half year, one-half to three-fourths year, or more than three-fourths year and whether between the time of his pre- and post-tests he was learning at a rate which was more than one-half month, within one-half month, or less than one-half month greater than his previous learning rate. Previous learning rate was computed by dividing the grade equivalent score on the pre-test by the grade level at time of pre-test. Current learning rate was computed by dividing the gain score by the time lapse between pre- and post-tests. In order to develop these categories it was necessary to ascertain that he had taken pre- and post-tests within the same battery and level, the time of the pre-test and the time of the post-test, and his grade equivalent score on each.

Population level. There were four distinct populations available to us on our tapes: district, principal, teacher and pupil. Separate questionnaires had been filled out at each of these levels. While it would not be possible to run pupil information at the district level, it was possible to run district information at any of the other levels; the degree of meaningfulness of running a district question at the principal level may be open to question. In general a variable was run at the population level of the questionnaire from which it was built. However, in a few cases variables crossed questionnaires and in still more cases it was desired to crosstabulate variables which were built from different questionnaires. In these cases the tables were run at the level of the largest population actually involved. An example of this is a table which crossed a district variable and a principal variable. The district variable categorized the types of federal programs within the district. The principal variable categorized the types of specialized facilities available within the school. The table, of necessity, was run at the principal level. However,

the interpretability of discussing specialized school facilities in school districts with various types of funding is difficult.

Type of population. Reference was made earlier to breaking a constructed variable over the school's participation in Title I. Actually there were four ways developed to handle Title I participation. The first was the Title I/non-Title I school participation break already mentioned; the second was a three-way pupil participation break--Title I participant, other federal program participant, non-federal program participant; the third was an embedding of the first two resulting in a six-way break; the fourth way the variable might be run was to not break it over title participation at all. Most of the variables concerned with Title I were produced twice: once with the school split and once with the pupil split.

Embedding process. The procedure used for the embedding sequence of variables depended upon a logical analysis of each particular variable. Some of the constructed variables such as the one mentioned above which dealt with all possible combinations of pupil services provided by regional centers did not involve embedding. Others such as the one dealing with achievement did. In those cases, it was necessary to decide which major categories should be the fastest and the slowest moving. Analysis was made of each of the variables in an attempt to determine which of the sub-categories would be most often compared in table reading. Those sub-categories were chosen as the fastest moving components in order that those columns would be closest together in the table.

Checking Procedures. A series of checking procedures was built to ensure variable accuracy. Following variable conceptualization, a staff member developed specifications of the categories; a logical analysis was then made by another staff member before the constructed variable was given to a programmer. The programmer then wrote a program which would be used to categorize each mem-

ber of our sample into his appropriate level of the constructed variable. Actually, these programs were subroutines or mini-programs which were placed into a larger program designed to prepare the data for the tables. The subroutine was then read and checked by a staff member before being run. As soon as the subroutine was run in a table, each level was checked to ascertain that each level contained appropriate numbers, based upon internal logic and univariate and other data where available.

If problems appeared to exist, a logical identification of problem source was made before submitting the subroutine to a different programmer for checking. This logical identification involved attempting to pinpoint which levels contained too many members and which levels contained too few. An attempt would then be made to determine what problem levels had in common. Then the subroutine was read, with special emphasis and analysis placed on those portions of the subroutine which were involved with the identified problem. The next step was to identify the characteristics required for placement in the levels which contained too few members. The checker would then reread the subroutine and attempt to correctly place a fictitious character into his appropriate level. This process was continued with assignment of a fictitious character to each different level until the checker was satisfied that he had identified and corrected the problem of misassignment. The programmer was provided with a complete identification of the problem but not with the staff member's identification of problem source. After rewriting, the programmer then consulted with the staff member who had identified the problem to ascertain that they were both in agreement as to problem source and solution. The above was a reiterative procedure continued until both the contractor and Office of Education personnel were satisfied that the variable was running as expected. The same general procedures were used for variable modification.

Problems

Several problems arise as a result of attempting to provide a data analysis which will provide program personnel with useable information. The first is the general inaccessibility of the personnel who would use the information below the policy-making level.

A second problem arises in terms of providing information while it is still pertinent. There is a considerable time gap between the initial gathering of the data and extensive feedback. For example, it was the last quarter of 1971 before the U.S.O.E. was provided with information beyond the univariate level for data collected in April of 1970. By the time that this information is generally available, many changes may have occurred throughout the country in program operation. Because of this time lapse, the information tends to be more applicable to summative evaluation at the policy-making level while the information itself was designed to serve a closer monitoring function.

A third problem which constantly besets all who come in contact with the process is the use of language in variable conceptualization. As a concept such as need, which was mentioned earlier, is developed, its definition goes through a series of changes, some obvious, some not as obvious, as attempts are made to utilize information provided by pre-existing instruments. As a result of this reiterative process of matching original concept and available information, the final variable may deviate considerably from the original conceptualization and may or may not serve its intended purpose. When over one hundred variables are being developed in similar fashion over a period of a few months, great care must be taken to ensure that all concerned are constantly kept informed of each change as it occurs so that the accurate identification of variable labels may be maintained.

A fourth problem which may occur is that in constructing and crosstabulating

complex variables we may be forcing relationships which do not actually exist. An obvious problem of this type arose when we crossed two specific variables, each of which used P-11, the student need univariate. One of the variables was designed to tap Pupil Academic Disadvantagement; the other, Degree of Pupil Academic Participation. When large chunks of a table are blank, or when one uses the same univariate data in two variables which are crosstabulated together, it is not difficult to recognize that it is a forced relationship. There may be many other cases, however, where it is not so obvious and yet the relationship is forced. One example would be that of using family income level in one variable and crosstabulating it with another variable which contained welfare data.