DOCUMENT RESUME

ED 174 653 TM 009 512

AUTHOR Greene, Jennifer C.; Kellogg, Theodore
TITLE Use of Existing Data Bases in a Large Scale Correlational/Regression Study. for Period January 1977-January 1978.
PUB DATE [77]
NOTE 15p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Data Analysis; *Data Bases; Educational Assessment; Elementary Education; Grade 4; Grade 8; *Institutional Characteristics; Research Design; *Research Methodology; *Research Problems; Research Utilization; *State Surveys; Student Characteristics
IDENTIFIERS Rhode Island

ABSTRACT

Statewide assessment data available from two school years, two grade levels, and five sources (achievement tests; student, principal, and teacher questionnaires; and principal interviews), were aggregated to more closely investigate the relationship between student/school characteristics and student achievement. To organize this large number of distinct data sets (old and new), all questionnaire and interview data were assigned to conceptual clusters. Separate analyses were conducted for each year (1975-6 and 1976-7) and for each grade (4 and 8). A single indicator of school achievement was used--mean standardized scores on the subtests of the Iowa Tests of Basic Skills. In response to the problem of different collection times for different data sets, the primary analyses were conducted on the 1976-77 set. A series of strategies were employed to overcome the statistical problems associated with a large number of variables and limited degrees of freedom. Despite these conceptual, methodological, and statistical problems, research with existing data bases was worth the effort; however, the appropriateness of an existing data base for answering new questions depends on the similarity between the original and the new questions. (CP)

# Use of Existing Data Bases in a Large Scale Correlational/Regression Study

In recent years the Rhode Island Statewide Assessment Program (SAP) has been designed to provide policy-makers with state level student achievement information and with information regarding the student and school factors significantly related to achievement. Data collected during the 1975-76 SAP and the 1976-77 SAP included achievement test data from fourth and eighth grade students, data on student characteristics from a questionnaire administered along with the test, and data on school characteristics from a questionnaire completed by the school principal. Analysis of these data left state policy-makers with the belief that they needed additional information about important student and school factors to interpret and use their test results. The State of Rhode Island therefore sponsored a Study of Schools (SOS) research project for the purpose of:

(1) examining more closely the relationships between educational factors and student achievement uncovered by the Statewide Assessment Program during 1975-76 and 1976-77, and

(2) investigating more systematically relationships between a wider range of educational factors and student achievement.

This paper examines a variety of methodological issues and problems which arose during the SOS study, describes alternative solutions to these issues, and assesses the overall value of conducting similar research with previously collected data. It should be noted that other

investigators might arrive at different solutions to these same issues or different conclusions regarding the value of studies like the present one. However, the purpose of this paper is not to debate the merits of alternative solutions but rather to share general approaches to problems arising out of the use of existing data.

Table 1 outlines the existing and newly collected data sources for the SOS project. As indicated, the SOS data sets varied in date of collection, level of collection, and sampling design.

## Methodology--Issues and Solutions

It is first important to discuss the methodological parameters under which the SOS project was conducted:

(1) The study was performed under a contract in which the request for proposals and the proposal assumed many design considerations as given.

(2) The study was conducted in the spring of 1977 with the requirement that all new data be collected prior to the end of the academic year.

(3) There was no mandate that all schools which had participated in the SAP had to participate in this study.

These methodological parameters highlight a common and important problem in research studies using existing data. Such studies often charge researchers with a specific question or problem but leave them

TABLE 1

Study of Schools Data

| Data Source | Collection Date | Level of Data | Sample Design | Approximate n |
|---|---|---|---|---|
| Existing Data: | | | | |
| 1. ITBS - 1975-76 | October, 1975 | Individual 4th and 8th grade students | Statewide Matrix Sample | 6770 |
| 2. ITBS - 1976-77 | October, 1976 | Individual 4th and 8th grade students | Statewide Matrix Sample | 7700 |
| 3. Student Questionnaire | October, 1975 | Individual 4th and 8th grade students | Statewide Matrix Sample | 6770 |
| 4. Student Questionnaire | October, 1976 | Individual 4th and 8th grade students | Statewide Matrix Sample | 7700 |
| 5. Principal Questionnaire | October, 1975 | Each School | All Schools in SAP | 200 |
| 6. Principal Questionnaire | October, 1976 | Each School | All Schools in SAP | 240 |
| Newly Collected Data: | | | | |
| 7. Teacher Questionnaire | May, 1977 | Individual Teachers | Sampled Within 1976-77 SAP Schools | 1300 |
| 8. Principal Interview | May, 1977 | Each School | All Schools in 1976-77 SAP | 150 |

little flexibility to design the type of study needed to answer the question. This problem can be stated in more general terms as:

* Can the question being asked in the evaluation study be answered within the framework of the existing data bases?

In the SOS study the questions asked were similar to the original intent of the SAP. Therefore, the framework of the existing SAP data was generally appropriate for SOS analyses. However, the SOS investigators used an additional strategy to compensate for constraints imposed by given design factors in the existing data bases. This strategy was to collect the new data needed for the study within the same general framework as the existing data. With this strategy, discrepancies between the old and new data wre minimized and matches between the two sets of data were facilitated.

## Conceptualization -- Issues and Solutions

As presented in Table 1, this study had an enormous amount and variety of data available for two different years at two different grade levels from five different sources (achievement tests; student, principal, and teacher questionnaires; and principal interviews). The initial challenge of organizing these varied data sets raised two major conceptual problems early in the study.

* How can a large number of distinct data sets (new and old) be organized to form a logical, integrated structure that has theoretical and/or practical relevance?

In response to this problem, the following solutions were applied in the SOS study:

(1) At the beginning of the project all data were organized into conceptual clusters. Newly collected information from the teacher questionnaire and principal interview were organized into a conceptual framework using the following categories: demographic information, school setting, curriculum and philosophy, instructional setting, and school climate. Existing data from the student and principal questionnaires were organized into two clusters: student factors and school factors.

(2) The analyses of data for each year were conducted separately for two main reasons. First, a substantially different population of schools participated in each year of the SAP. Second, the collection of new data was restricted to schools which had participated in the 1976-77 SAP. (Schools which had participated in the 1975-76 SAP only were excluded from this study because of a lack of reliable information as to the teachers in those schools during that year and a lack of a means of gaining access to these teachers.)

(3) Separate analyses were conducted for fourth and eighth grade schools.

(4) A single indicator of school achievement was used as the dependent variable.

The second major conceptual issue was:

* How to define the dependent variable?

Use of existing data created a major problem in defining the dependent variable. In the SAP sampling procedure, from one to four groups of students were selected from within schools to take one of the four subtests into which the ITBS battery had been divided. Each student, therefore, took only one subtest; and each school, therefore, had from one to four mean student achievement scores available, each representing a different subtest.

The design of the study, however, required that the dependent variable be expressed as a single indicator of achievement for each school. In order to derive this single indicator, standardized z-scores were first calculated for each subtest. Then, school means for each subtest were derived from individual students' z-scores. Finally, a single achievement indicator for each school was computed by averaging the one to four mean standardized scores available for that school (representing the one to four subtests taken by students in that school).

Comment. While challenges to the conceptual framework established for the SOS data can be made, particularly in reference to the derivation of a single achievement score for each school, it is not the purpose of this paper to defend the methodology but rather to share the type of approach used in working with existing data. We found the establishment of the conceptual framework highly valuable. It helped us to recognize the limitations of the data early on (e.g., in the achievement measure), alerted us to further conceptual and statistical problems, and facilitated a systematic and thorough exploration of the data.

Data Analysis--Issues and Solutions

The following describe the primary statistical questions we faced during the SOS study, as well as the solutions to these problems we utilized.

* What consideration was given to the fact that data were collected at different times?

In response to the problem of different collection times for different data sets, it was decided that the primary analyses in the study would be conducted on data collected during the 1976-77 academic year. A complete data set of both old and new data existed within this year, while the 1975-76 analyses required the use of data collected during 1976-77. It was further decided to use the 1975-76 data set to test, explore, or verify significant findings from the 1976-77 analyses. (Examples of such findings include correlations, factor patterns, and regression equations.) With this strategy, consistency within the data set was strengthened without sacrificing a substantial portion of the available data.

* Was the existing data in a metric suitable for direct use?

Often it was not and when not, substantial work was involved in altering the metric. One example, given above, required conversion of student scores on each test section to a standardized metric. As described above, scores were calculated for each student.

A second example involved the determination of an appropriate indicator for socioeconomic status (SES). Several variables related to SES (e.g., mother's and father's education, mother's and father's occupation) were included on the student questionnaire. We first attempted to obtain a single SES indicator by recoding these variables into approximate rank order scales and then regressing district median income on this rank ordered data. The results of this effort, however, were disappointing. In our second procedure, we examined the intercorrelations among the rank ordered variables indicative of SES. This yielded a single variable, father's occupation, which was considered to be an appropriate indicator of all the SES variables and used as such in all further analyses.

* Use of the school as the basic unit of analysis created statistical problems such as the limited degrees of freedom. How were these problems overcome?

A series of strategies were employed to overcome the statistical problems associated with a large number of variables and limited degrees of freedom.

(1) At each step of the analysis, all available data were used. Allowances were made to include cases with missing data wherever possible. This procedure required a changing of sample size from one analysis to the next.

(2) Extensive analyses were conducted on the independent variables in the study prior to analyses of the relationships between independent and dependent variables. These preliminary analyses were conducted both within each set of data from a single data source (students, teachers, and principals) and across these

three data sets. In addition, the analyses within and across data sets were conducted using the conceptual clusters described earlier. These analyses served to identify the key variables to be included in the more advanced statistical analyses (with their accompanying more stringent limitations and assumptions). Then, using this limited set of variables, only a small series of multiple regression analyses were needed to conclude the analysis phase of the study.

(3) We do not believe that the analysis strategy described above represents an example of simply massaging existing data until it looked good. One important procedure which we believed helped to minimize evaluator bias in using existing data was the establishment of a set of criteria for determining statistical and educational significance of analysis results prior to undertaking the analysis. (For example, statistically significant correlations were considered to be educationally significant only if they appeared in both years or both grade levels.) These criteria also served the useful function of aiding in our data reduction efforts.

* What were the consequences of using the school as the unit of analysis?

The most important consequence of using the school as the primary unit of analysis was that information on within school variability was sacrificed in order to answer the major research questions in terms of between school differences only. The structure of the existing data required either sacrificing within school variability or, with the vast

majority of variables, assigning school average values to all students within a school. Because this latter procedure artificially multiplies the amount of data and inflates the degree of freedom, it was rejected, thus sacrificing within school analyses.

## Reporting of Results--Issues and Solutions

The presentation of the results of the SOS study to policy-makers within the State of Rhode Island raised one final issue regarding the use of existing data bases in evaluation studies.

* Given the methodological, conceptual, and statistical problems associated with the use of both new and existing data bases in evaluation studies, how should the results of such studies be presented?

Implicit in this issue is the recognition that the audiences of evaluation studies often include lay policy-makers (e.g., Boards of Education or Boards of Regents). The issue arises from the need to highlight the significant findings of the study while giving equal time to the limitations of the data, all within a non-technical framework.

In response to this issue, the authors of the SOS project attempted to focus on a limited number of key findings in presenting the results to state policy-makers. In addition, the authors presented the findings as indicators of possible relationships, trends, and patterns, rather than as conclusive findings about educational factors significantly related to achievement in Rhode Island's public elementary schools. Finally, the authors concluded the SOS report with a clear statement of the kind of research project (e.g., an experimental rather than a correlational

study) needed to answer the policy-makers' questions. Incorporated into this statement were several specific hypotheses worthy of future research, derived from the SOS study.

## Research with Existing Data--Is It Worth the Effort?

Given the methodological, conceptual, statistical, and reporting problems we encountered in the SOS study, we are forced to ask: Is it worth the effort? The combined effects of these problems detracted from the overall quality of the data set and seriously undermined the credibility of the results. Consequently, we felt obligated to present the results as tentative indicators of possible relationships, rather than as firm conclusions upon which policy could be based. That is, we were not able to meet the original expectations for the study.

Nonetheless, our own perceptions of the value of the SOS study were that it probably was worth the effort. These perceptions are based on four main factors. First, although the cost of the SOS project was sizeable, to have collected independent test results across a similar sample of Rhode Island students would have increased the cost factor six fold or more. Second, to successfully meet the challenge of a project like the SOS study produced great personal satisfaction among the researchers. Third, the study did contribute to the general body of knowledge on correlational relationships between educational factors and achievement. Finally and most importantly, the study did have a practical impact. This impact did not match original expectations, since the SOS results did not provide a firm foundation for policy decisions.

Rather, the practical impact of the SOS study was more subtly felt in the area of educating Rhode Island policy-makers about the limitations of correlational research. Just a few months ago, the Rhode Island State Board of Regents declared a one-year moratorium on statewide assessment, during which time they plan to sponsor a small set of more targeted, experimental studies. The SOS project was one among several related studies that significantly influenced this recent change of direction by the Board of Regents.

Although our response to the question, is it worth the effort, is "yes", it is a qualified "yes", and the reason for this qualification deserves special emphasis. The major problems we encountered in the SOS project stemmed from the fact that data originally collected to answer one set of questions were being combined with new data to answer new questions. Even minor discrepancies between the two sets of questions were sufficient to limit seriously the appropriateness and relevance of the existing data base to the new questions.

The best example of the point being made here is the unit of analysis issue. SAP data were originally collected (under a matrix sampling procedure) to provide <u>state level</u> achievement information, and deliberately <u>not</u> to provide school level achievement information. Yet, the very nature of the SOS study and the new teacher and principal data collected within it required that the overall questions be answered at the school level. So, we chose the school as the only appropriate unit of analysis for the SOS study, and, consequently, had to revise substantially the existing data base to fit analyses at the school level. (Such revisions included redefining the dependent variable, sacrificing within school variability, and working within significantly more limited degrees of freedom.)

In short, the SOS study was limited by the mismatch between the information contained in the existing data base and the information needed to answer the new questions. We can therefore suggest that the quality of an existing data base, innnthe context of using that data to answer new questions, is a direct funtion of the degree of similarity between the original and the new questions. Even in the SOS study, which attempted to use SAP data to answer SAP-like questions, there were enough discrepancies between the two sets of questions to lower significantly the quality of the existing SOS data. Yet, these discrepancies were not fully realized until after the study was in full operation.

Since use of existing bases is both sensible and cost-effective, researchers are likely to continue this practice. Based on our experience, we would strongly recommend that a careful and complete review of the suitability of the existing data base to answer new questions be conducted prior to beginning evaluation studies on such data. In the absence of such a review, researchers may find some helpful hints among the solutions wer presented above as they encounter the same kinds of problems.