UNITED STATES GENERAL ACCOUNTING OFFICE
WASHINGTON, D.C. 20548

INSTITUTE FOR PROGRAM EVALUATION

B-207737

JULY 6, 1982

The Honorable Bruce Chapman
Director, Bureau of the Census

Dear Mr. Chapman:

Subject: Data from the Retirement History Survey (GAO/IPE-82-5)

An analysis we initiated last year on retirement patterns among the elderly has been terminated because of problems we found in the data. These are data currently being disseminated by the Bureau of the Census under the title "The Retirement History Survey (RHS)." The purpose of this letter is to inform you of the problems we found and to suggest that you inform all requestors of the RHS data that they contain serious errors.

Our study, which we began in July 1980, was to focus on retirement patterns in the general workforce. Specifically, we were interested in why people retire, the overall distributions of their retirement ages, and how their life styles are affected by the decision to retire. We hypothesized that health, economic status, expenditure patterns, and levels of social activity are significant elements in the decision to retire, and we were interested in how these four elements are affected by retirement. We wanted to know, for example, whether health deteriorates after an individual has left the workforce.

In search of a data base, we approached the Social Security Administration (SSA) in hopes of using the Retirement History Survey, which seemed ideally suited to our objectives and interests. The RHS consists of data collected by the Bureau of the Census at the request of SSA, in an extensive effort to collect information on issues related to retirement. A panel of approximately 12,000 individuals was followed into and through the retirement years (59-72 years of age).

The Data Users Services Group at Census informed us that Census provides only computer tapes with data for each year--data for 1969, for example, are not on the same tape as data for 1977. Thus, merged tapes of selected items could not be provided to us. Since our analysis required cross-time comparisons, we subsequently arranged for the Center for Aging at Duke University to provide a merged data tape with 341 selected RHS variables. This step was taken at the suggestion of Dr. Lola Irelan, who headed the RHS project for SSA.

(973163)

Receiving the data in November 1981, we began to prepare it for analysis, but in the course of running some standard diagnostic checks, we discovered problems almost immediately. The deeper we delved into the data, the more problems we uncovered. Finally, after an expenditure of $12,000 in computer funds and many staff days, we were forced to abandon the project because of the data's poor quality. In all, we encountered nine kinds of problem, which we list below, with examples of each. We have not included detailed information on all problems with every variable. If you would like to have that information, we can forward it to you.

1. There are inconsistencies in items across years--for example, out of 91 items selected from the 1971 questionnaire, 21 items were not repeated in the three following waves (1973, 1975, and 1977). This prevents consistent analysis across years.

2. There are inconsistencies within individual questionnaires --for example, people who were married at the beginning of the instrument suddenly wind up in the "single" category for later questions on the same instrument.

3. Response categories are inconsistent across years--for example, annual salary is recorded on an interval scale in 1971 and 1975 but in categories (as $1,500-$2,000) in 1973.

4. There are unrealistic extreme values--for example, in each wave there are respondents who indicated that they have 44 or 99 children. Twenty-nine of 91 items chosen within 1971 have this problem.

5. There are frequent instances of missing values--for example, for one of the critical expenditure variables, 11,000 cases are listed as missing. Additionally, there are inconsistencies across years in the number of missing cases for the same item, so that there might be 3,000 missing in 1971 and 1977 but 9,000 missing in 1975.

6. Response categories are improperly defined--for example, if the industrial codes in the RHS manual are correct, the largest employer in this country is reported as forestry and fisheries.

7. There are negative values present that seem logically impossible--for example, earnings.

8. Internal errors exist across variables within years--for example, the maximum value for respondents of Social Security survivor benefits in 1975 was $14,784. The maximum value for respondent and spouse benefits from this same

source for the same year, however, was only $2,287. We found 15 such instances. Additionally, for fully 70 percent of the income variables, the maximum value reported was $50,000. Although this figure is certainly reasonable as a maximum for any single item (such as salary), our suspicions were aroused because of the frequency with which it occurs.

9. Inconsistencies exist in responses across years--for example, the mean for utility bills for 1971, 1973, 1975, and 1977 is given as $664, $525, $498, and $620, respectively. One could explain either a consistent increase (rising costs) or decrease (restrained consumption) across years, but the initial drop and dramatic increase in 1977 are hard to reconcile. Twenty-two of the 91 variables exhibited such cross-year inconsistencies.

These nine kinds of problem differ in the frequency with which they occur and their damage to any particular analysis. It is clear, then, that potential users may be better served if they know the condition of these data before committing significant funds and effort to acquire data tapes. The Bureau of the Census has a well-deserved reputation for the quality of its data. However, in the case of these tapes, investigators who will anticipate no more than the usual problems could be misled.

The General Government Division within GAO has been notified of the problems we encountered with the RHS data. We are also sending copies of this letter to Mr. Michael Garland, Chief of Data User Services within the Bureau of Census, and Mr. John Svahn, Director of the Social Security Administration.

Sincerely yours,

Eleanor Chelimsky
Director