# The Automation of Data Processing, Analysis, and Reporting in a Large Survey Time-Series Database

Christopher J. Gordon and Michael B. Zartman
Institute for Social Research
The University of Michigan

The May 1981 Survey will mark the 152nd Survey of Consumer Attitudes. Initiated in 1946, the purpose of the surveys is to measure changes in consumer attitudes and expectations, to understand why these changes occur, and to evaluate how they relate to consumer decisions to save, to borrow, or to make discretionary purchases under changing conditions.

Each survey contains approximately 40 core questions, each of which probes a different aspect of consumer confidence. Open-ended questions are asked concerning evaluations of expectations about personal finances, employment, price changes, and the national business situation. Additional questions probe for the respondents appraisal of present market conditions for houses, and other durables. Demographic data obtained in these surveys include income, age, sex, race, education, and occupation, among others. While many questions designed to measure change in attitudes and behavior are repeated in identical form in each survey, special questionnaire supplements are added to most surveys by outside sponsors on a time share basis. Supplements to the ongoing surveys give sponsors prompt turnaround to survey materials while taking advantage of shared field expenses. When the research task is first undertaken, a maximum amount of time and effort can be spent in developing these survey materials, not in establishing and setting in motion standard sampling and interviewing procedures, questionnaire and code development for standard demographic items, and so forth.

Although each survey task is unique in its time requirements, shared time participation on the ongoing Surveys of Consumer Attitudes is an effective and flexible approach for meeting many research needs. Current procedures include production of a fully documented computer data file available for analytic use within 48 hours of the close of interviewing. Within one week of the close of the survey, a report containing tabulations and charts of questions asked is sent to the sponsors.



## National Telephone Samples

The Survey Research Center conducts an ongoing nationally representative survey based on approximately 700 telephone interviews per month with adult men and women living in households in the coterminous United States (48 states and the District of Columbia). This sample is designed to maximize the study of change by incorporating a rotating panel sample design in an ongoing monthly survey program. For each monthly sample, an independent drawing of telephone numbers is made. The respondents chosen in this drawing are reinterviewed in six months. A rotating panel design results, and the total sample for any one survey is normally made up of 55% new respondents (RDD, and 45% of the respondents being interviewed for the second time (reinterview). This design permits the regular measurement of change in the aggregate and also allows the assessment of individual change at six-month intervals.

The rotating panel design of the Surveys of Consumer Attitudes has several distinct advantages over a simple random sample. The ability to gauge individual change expands the study of aggregate change by permitting a better assessment of the causes of that change. Quite complex research strategies can also be implemented using the recontact features of the sample. For example, respondents can be recontacted and follow-up questions on a reinterview can determine subsequent behavior on attitude change. In addition, the frequency of the surveys allows screening for supplemental samples to be easily accomplished.

The sampling and administrative designs of telephone surveys differ from those of most personal interview surveys simply because telephone surveys utilize a communication medium, both to identify sample households and to access them. Where many personal interview surveys use Census data to assign probabilities of selections to areas (e.g., counties, towns, blocks), national telephone surveys use the identifiers of telephones, ten-digit telephone sample designs of the Survey Research Center employ computer routines to randomly generate telephone numbers. In short, the sample is drawn so that every number in the coterminous

United States has an equal chance of being selected.

All surveys are subject to sampling error because not all members of the population are interviewed. Telephone interviewing fails to include the approximately 9-10% of the households in the United States that are not telephone subscribers, although the percentage of nonsubscribers is shrinking over time. Past analyses suggest that nonsubscribers are disproportionately poor, black, and rural and that they are more likely to rent and live alone than the rest of the population. Current studies of the bias which results from the exclusion of non-telephone subscribers indicates that it is not severe and probably is within the accuracy requirements for most, but not all, survey research projects.

Since only one adult per housing unit is selected, the results may need to be weighted by the number of eligible respondents in the household and inversely by the number of separate telephone numbers per household. Additional weights are introduced to adjust for variations in age and income distributions compared with a recent, self-weighting national personal interview sample, as well as the result of the incorporation of recontact respondents in the sample.

Most results on the total sample for this survey will differ by no more than 4 percentage points in either direction from what would have been obtained by using the same methods on the entire population. The sampling error from smaller subgroups is larger depending on the number of cases in the subgroup.

## Coding Section

Before any significant analysis of the verbal information gathered in surveys can be undertaken, these data must be converted to numeric codes for use in data processing. The Survey Research Center maintains a permanent staff of trained coders using highly evolved techniques. Coders are responsible not only for converting the data to machine-usable form, but also for providing the final checking function during which errors are detected and brought to the attention of the project staff.

It is in the coding section, however, that the highest level of non-automated human input occurs. This takes place in the coder translation of "open-ended" responses, or responses that may fall into any one of up to approximately eighty codes for a single question, as opposed to from five to seven codes for a "closed-ended" question.

## Data Cleaning

Data cleaning operations consist of the following steps; first, a check is made to assure that a complete set of valid data codes are present for each interview. Coded data are checked for logical consistency with current and previous coding decisions. Inappropriate decisions such as wild or inconsistent codes are fed back to the coder for correction. Following this, a number of checks are made for internal consistency of each interview. All inconsistencies are resolved under supervision of the project staff. A final check is then made between the data file and Field Section file to insure that all interviews have been processed and coded.

The Survey Research Center has also been developing a Direct Data Entry system named CATI (Computer Assisted Telephone Interviewing). CATI will eventually take over as an automatic coding and data cleaning system, allowing interviewers to enter data responses directly on a computer terminal. The interviewers will be able to check the entered codes, create data records for open-ended responses, and speed up the interviewing process because questionnaire branching will be accomplished automatically.

## Processing

The input to the data processing required to result in an archivable dataset is the raw data as obtained in the survey itself and appropriately coded. It currently comes in the form of a deck of punched cards and a codebook. In the process itself there are two main tasks involved. The first is the creation of recoded variables including the weight variables. The second is the merging of demographic and old core variables for reinterviewed respondents. The output from the process is three OSIRIS.IV datasets.

The full dataset is obtained and contains all variables ordered by category to ease data analysis. A standard dataset is obtained that contains a subset of the variables (core, demographic, and recoded), with variable numbers which coincide with standard datasets from past surveys. And finally, an archive dataset is obtained which contains all variables numbered sequentially in the order given by the questionnaire.

Since September 1980, a "master" dictionary has been used in creation of the full dataset; it has the advantage of keeping variable numbers, names, and other characteristics constant through time, allowing for consistency and easier time series analysis. Further, the processing setup files require many fewer changes from month to month, while easing the burden of maintaining a question history file, including information about variables such as question text, codes, chronology, and question wording changes. Standard datasets are made, to be compatible with past processes, and remain quite useful by allowing combinations to obtain larger datasets (e.g., quarterly or annual datasets).

## Data Analysis

The Institute for Social Research has developed excellent facilities for processing social science data. The Institute has developed and maintains several packages of statistical programs for processing data and provides a wide variety of computer services. The Survey Research Center utilizes the University of Michigan's AMDAHL Computer (MTS) for major processing and statistical analysis. The Survey Research Center also operates a PDP 11-70 for use with the CATI system.

Two program packages available to SCA staff--OSIRIS Version IV.5 and MIDAS--provide very comprehensive analytical capabilities. Specifically, they both include capabilities to perform routine univariate descriptive (raw frequency and/or percentages) spreads and bivariate (frequency and/or percentaged) tables. Both of these systems make available a wealth of summary statistics, appropriate to the level(s) of measurement of the variable(s) being analyzed, to support these descriptive data. In addition, both systems include capabilities to perform both bivariate and multivariate relational analysis, such as linear correlation and regression. OSIRIS contains multivariate analytic procedures developed by Institute staff especially for multivariate analyses of survey data.

## Reporting

Reporting may take two alternative forms: 1) the user may tap directly into the data base, either through MTS or by a tape sent according to the organizational specifications; or 2) preparation of standard data tables and an analytic report is conducted by staff researchers of the Monitoring Economic Change Program on a monthly basis after each new survey.

A major focus in output of the monthly surveys is the Index of Consumer Sentiment (ICS). The Index is created each month from five variables contained in every survey. The ICS is highly regarded in both private and public sectors as a reliable lead indicator of discretionary consumer expenditures and as a consumer based barometer of the economy. As a result of the dated nature of our data and in view of the fact that many of our clients utilize our data as an input to both short and long term planning decisions, there is necessarily a critical demand for the expedient accuracy in the processing and in reporting our survey data.

As recently as six years ago the "turnaround" time from raw survey data to publication of the results was 10 days, and required 10 people. In light of the fact that we only produced quarterly surveys at the time, and that the equivalent "turnaround" period for other survey research based programs was from 3 to 6 months, our situation was satisfactory. Satisfactory that is, unless an error was discovered in the final dataset (not uncommon given the high degree of human input); this would in turn require another 10 days of processing. A system decision was adopted at that time, with emphasis on ease of dataset correction and (inherently) de-emphasis of user input and interaction. And so, as the available software progressed so did our processing system.

In 1978, the Survey of Consumer Attitudes program began surveying on a monthly basis. Turnaround times that were sufficient to maintain a quarterly schedule were clearly not appropriate on a monthly basis. Moreover, tables and charts that were produced via hand editing and typing were simply too time consuming and inflexible for the production of both standard and nonstandard analysis runs.

From this need for better reporting methods arose our data depository in which the aggregated results of all of our surveys were to be stored in a standard format. Allowing for ease of access and analysis of as few or as many of our data sets as we specify at one time, the time series research implications are clear. With this data depository, the automation of the various steps for processing continue with emphasis now on a high degree of flexibility and the use of analytic research tools for an improved quality, variety, and expediency in our reporting system.

The standard datasets are the input to programs which add data to the data depository file, which is in turn the input for programs that create tables and charts for our monthly publications. The data in the depository file is indexed by date with IBM internal line numbers, allowing flexibility in data manipulation with FORTRAN and facilitating the creation of transformed data segments including three month moving averages and survey indexes. These segments themselves are stored in the same depository file as the raw data allowing plotting routines to selectively request the appropriate formats.

The Monthly and Quarterly publications each include 34 tables with two supporting charts, one highlights recent changes and the other one, historical patterns. The tables are automatically updated each month and are printed on a word processor for a "camera ready" copy. The chart setup files require several processing steps to result in machine readable (and plottable) commands. This output is also "camera ready" and is the endpoint of our monthly processing and reporting routines.

## Conclusion

We are currently at approximately 80 percent realization of our system design goals. In the short term we expect to bring our development to within 90 percent, as we more fully utilize our computing capabilities. In these last stages the developmental emphasis will be on the final processing step of reporting.

Whereas we now produce three standard types of publications (monthly, quarterly, and an archival time-series summary), we hope to to able shortly, to produce monthly and quarterly results that are representative of user (or sponsor) specified demographic subgroups. Thus, we could produce a publication that represented only those respondents 25 to 34 years old, or one that represented only respondents living in the Western region of the United States. We believe these could be powerful tools in the policy and planning decisions of our sponsors.

A major restriction hampering the automation of our system is the checking of the input data for wild codes. The scanning of the codes themselves is automated, resulting in a list of invalid codes, with case and variable numbers. Excessive user input is present at this point, forcing a check of the interview itself for the error, and then correcting the data file by recording the appropriate code. With the current wild code rate running between 5 and 6 percent of the processed surveys,

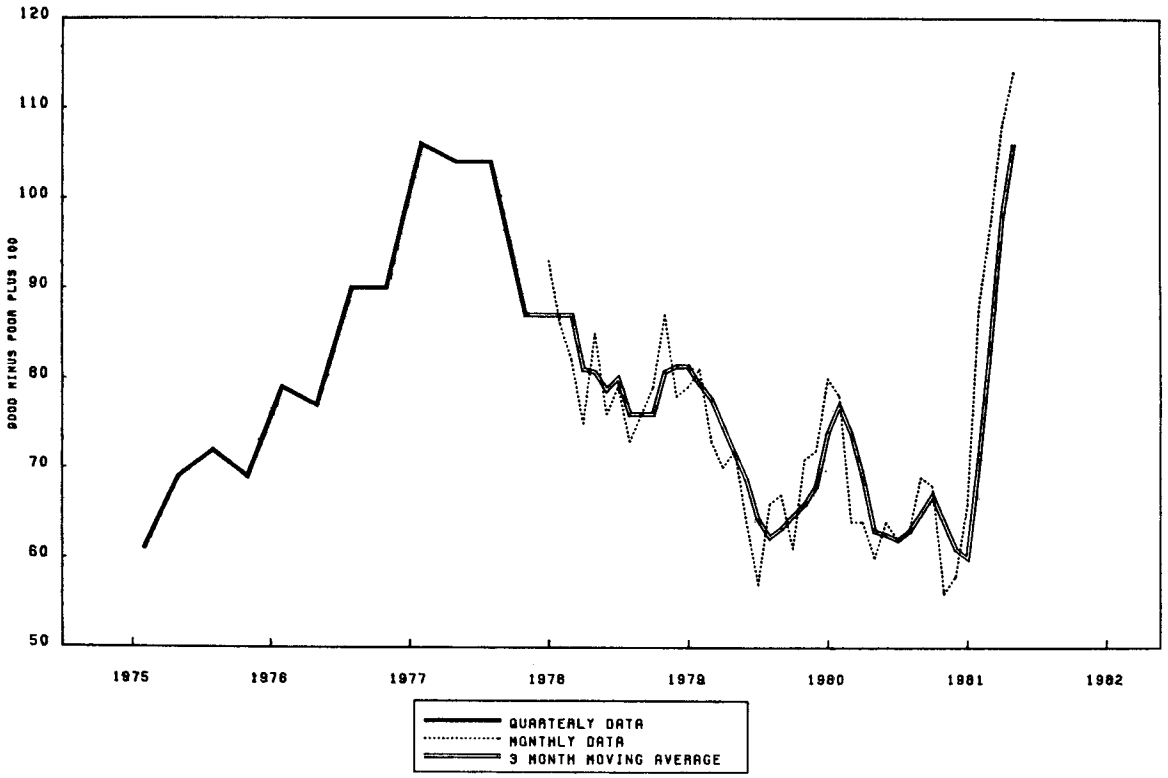
this particular processing step accounts for approximately 15 to 20 percent of the total survey processing time. The problem itself is a basic one, resulting from the input of user errors at the preliminary stages (precomputer processing) of the system, specifically between interviewing and coding.

The proposed CATI DDE system should eliminate this source of error and time loss. When using CATI the interview itself would be on line (versus on paper), prompting the interviewer for the appropriately coded answers for each question (variable), accepting only those codes listed as valid for each variable. Since the computer scans for wild codes as the interview is taking place, this system effectively eliminates the wild code problem.

In our efforts to more fully understand the complex relationship between attitudes, expectations, and the economic decision making process of consumers, greater emphasis has been placed on the research vehicle. Expediency and accuracy have become the measures by which a system is evaluated for its efficiency in the processing, analyzing and reporting of survey data. Our system was designed around the parameters of quick turnaround, limited user interaction, flexible analysis and accurate reporting. Full automation in these terms then, can only be realized through a marriage of a flexible system design and the adoption of the most current technology.

EXAMPLES

OPINIONS ABOUT THE GOVERNMENT'S ECONOMIC POLICY




| | May 1980 | June 1980 | July 1980 | Aug. 1980 | Sept 1980 | Oct. 1980 | Nov. 1980 | Dec. 1980 | Jan. 1981 | Feb. 1981 | Mar. 1981 | Apr. 1981 | May 1981 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A GOOD JOB | 5% | 10% | 9% | 8% | 10% | 8% | 6% | 7% | 10% | 19% | 22% | 28% | 30% |
| ONLY FAIR | 48 | 40 | 41 | 45 | 46 | 50 | 41 | 42 | 42 | 40 | 45 | 47 | 49 |
| A POOR JOB | 45 | 46 | 47 | 45 | 41 | 40 | 50 | 49 | 44 | 31 | 25 | 20 | 16 |
| DK,NA | 2 | 4 | 3 | 2 | 3 | 2 | 3 | 2 | 4 | 10 | 8 | 5 | 5 |
| TOTAL | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| CASES | 703 | 688 | 668 | 658 | 682 | 685 | 694 | 684 | 698 | 668 | 703 | 690 | 667 |

The question was: "As to the economic policy of the government--I mean steps taken to fight inflation or unemployment--would you say the government is doing a good job, only fair, or a poor job?"