Journal of Official Statistics
Vol. 5. No. 1, 1989, pp. 3–21
© Statistics Sweden

Information Technology and Survey Research: Where Do We Go From Here?

J. Merrill Shanks[1]

Abstract: This article reviews the areas in which information technology has had an impact on the cost, quality, or complexity of survey research, and discusses alternative strategies for integrating the computer-based activities which take place in different stages of the survey research process. Special attention is given to the continuing revolution in survey data collection, for it is the only area of applied computing that is unique to the survey field and is central to the eventual integration of data collection, analysis, and management.

Key words: Computer-assisted surveys; data collection; data management; data analysis; computer-assisted telephone interviewing (CATI); generalization of technical procedures; alternative computing environments; survey integration.

1. Introduction

Forty years have passed since the U.S. Bureau of the Census acquired the first UNIVAC to process data based on structured questionnaires. Since that time, the list of activities which depend on computers for the collection, processing, or management of survey-related information has grown steadily, to the point where nearly all aspects of the survey process are now at least partially dependent on computer-related technology. This article begins with an overview of survey activities where computers have already had an effect, with an emphasis on the current revolution in computer-assisted data collection. Most of the article, however, concerns alternative strategies for future research and development and the integration of separate techniques for data collection, analysis, and management.

Survey organizations must make increasingly complex (and expensive) decisions if they are to take advantage of continuing advances in computing and information technology. Survey projects that either develop or use computer-based techniques have a tendency to concentrate on short-term improvements or immediate objectives. The purpose of this article, however, is to re-emphasize long-term objectives in computer-assisted surveys and discuss alternative strategies for future development in that field. In cooperation with other survey organizations, the Computer-assisted Survey Methods

[1] Director, Computer-assisted Survey Methods Program, University of California, 2538 Channing Way, Berkeley, CA 94720, U.S.A.

Program (CSM) at the University of California is active in several aspects of the relationship between "computing" and "surveys." As a consequence, the observations which follow represent both a general commentary on our field's progress to date and specific recommendations for a research and development agenda that is shared by many other organizations.[2]

1.1. Computer-assisted surveys: Rationale and impact

For several years, the author has been a frequent visitor in other survey organizations, in order to discuss computer-based systems for various aspects of survey research. Most of these visits have been initiated by organizations that are interested in computer-assisted telephone interviewing (CATI), but the resulting discussions almost invariably cover a variety of other computer-related activities. At first, it seemed reasonable to assume that all participants in such discussions shared the same basic objectives in adopting new computer-assisted techniques. It quickly became clear, however, that the motivations and expectations which precede the adoption of a computer-based system can vary substantially within the survey organization. Our field is no different from many others in this respect, for potential users may be attracted to any computer-based system for one or more of the following (partially contradictory) reasons:

- the output from the process may be better, in that the resulting information will be more accurate or of higher quality;
- the entire process may be less expensive, even though computing equipment represents a cost that had not previously been involved;
- the entire process may be faster – whether or not it is less expensive, for time and money represent two different kinds of "resources;" or
- the process may be more powerful, for the task or the design may be so complex that it could not be done accurately without a computer-based system.

These kinds of expectations represent the intended consequences for survey researchers who consider the adoption of new computer-based techniques. Such adoptions, however, often lead to other (unintended) consequences, for significant changes can take place in the division of labor and structure of an organization after new technology is introduced.

To be sure, relatively few survey organizations adopt computer-based systems in order to change their bureaucratic structure or division of labor. As in other fields, however, new technology is often introduced for reasons that are partially in conflict, and groups with different objectives may compete for control of the new procedures. Frequently, one group or fraction seeks faster results or lower costs, while another emphasizes the possibility of greater quality or complexity – either of which may require additional funding or time.

In addition to conflicts between objectives, organizations that adopt computer-based systems frequently experience shifts in their inter-

[2] The opinions expressed in this article are entirely the author's, but they are based on cooperative projects with specialists in survey research from several institutions. In particular, the CSM Program has been working with the National Agricultural Statistical Service (NASS) since 1981 to develop, test, and utilize general-purpose systems for the collection and processing of survey data. The CSM software used in that project has been developed in an unusually collaborative fashion, involving staff members from the University of California, NASS, the Bureau of Labor Statistics (BLS), and several dozen other organizations that participate in the Association for Computer-assisted Surveys. The author is indebted to several colleagues for their ideas and criticism, but the views expressed in this article are not necessarily shared by other individuals in any of the organizations involved.

nal structure to fit the procedures that are defined by the new technology. Such changes were first evident during the 1960s and 1970s when survey researchers began to write their own statistical programs instead of relying on their centralized data processing staff. In the current period, the impact of technology on survey organizations has been particularly visible in the trend toward computer-assisted data collection.

Each organization is somewhat different in this respect, for survey projects vary enormously in their complexity and division of labor. Despite such differences, however, computer-based systems for data collection usually lead to some form of organizational change, for they combine survey activities that were previously carried out in separate offices or groups – i.e., in sub-organizations that specialize in instrument design, sampling, field, coding, or data processing. As additional survey activities are converted to rely on the same system or database, technical integration is required at the study or project level that simply did not exist before the shift to computer-assisted data collection. As emphasized below, this trend toward technical integration is likely to accelerate, as survey organizations improve linkages between separate systems for data collection, management, and analysis. Each step toward integration may suggest further changes in division of labor, as well as making some progress toward the objectives listed above – i.e., making surveys better, cheaper, faster, or more powerful.

The following section reviews the variety of information processing activities where computers have been used in survey research, with an emphasis on the changes that are still taking place in data collection – i.e., in the production, recording, and editing of survey information. The rest of the article then discusses alternative strategies for future development, with an emphasis on data management and the technical integration of the entire survey process.

2. Computer Applications and Survey Procedures

Many researchers can recall when survey data were stored on punched cards and processed by electromechanical devices instead of computers. It therefore seems natural to begin by identifying the range of traditional survey activities that involve some kind of "information processing" and the way in which computers were introduced in each of those areas. What follows is an oversimplification in some respects, but it suggests the variety of survey-related activities that may be affected by computing technology. (See Sonquist (1977) for a comprehensive review of computer applications in survey research.)

As shown in Fig. 1, most traditional survey activities have long been at least partially converted to computer-based procedures. The first area in which computers were widely used was the statistical summarization (or analysis) of coded data that had already been transferred from interview protocols to punched cards – which appears in Fig.1 as the penultimate activity in the traditional sequence from design through collection to post-survey data management. The first statistical programs were designed for a single type of computation (e.g., "means," "cross tabulation," or "correlations"), but software in this area quickly evolved into comprehensive packages, some of which now reflect over two decades of continuous development.

It is difficult to exaggerate the extent to which computers have improved the process of analyzing survey data. We now routinely expect results to be produced and displayed in a few seconds that once required an entire staff of technicians at desk calculators or unit record equipment. Survey researchers sometimes complain that our analytic strategies have not kept pace with improvements in the speed of computation. There can be no doubt, however, that survey analyses are now routinely completed that are far more accurate, cheaper, faster, and more complex than was possible before the introduction of digital computers. Specific capabilities in this area will not be discussed in this article for the systems involved are frequently reviewed in statistical or computer-related publications. For example, the *Statistical Software Newsletter*, published in connection with the International Statistical Institute, is entirely devoted to reviews and critiques of statistical software. For more comprehensive reviews, see Francis (1981) and Raskins (1989).

The traditional division of labor in survey organizations involves separate staffs for sampling, field, coding, data processing, and analysis. Because of these structures, computer applications for the other activities described in Fig. 1 were usually developed independently of statistical software, and were often based on different computers and data processing conventions. The piecemeal or uncoordinated nature of these initial computer applications can also be partially attributed to the sequence in which new computing capabilities became available. As shown in Fig. 2, survey researchers have worked with a series of "new" infor-

Fig. 1. Information-processing Activities in Traditional[3] Survey Research

Original Activity	*Initial Computer Utilization*
Preparation of Interview Schedules and Specifications	Text Processing and Document Preparation, for both Questionnaires and Interviewer Instructions
Sample Selection and Administration	Data Management, for Sample Selection and Preparation of Pre-interview Data
Interviewer Supervision and Management	Keeping Track of Field Outcomes and Measurement of Interviewer Performance
Content Analysis for Text or Verbatim Responses	Conversion of Coded Data to Machine-Readable Form (Keypunching or Direct Data Entry)
Data Preparation: Checking and Documenting the Resulting Data	Detection and Correction of Inconsistencies and Preparation of Machine Readable Codebooks
Survey Analysis: Producing Tables and Charts	Statistical Computation, including Graphic Displays as well as Descriptive and Inferential Statistics
Post-Survey Data Management: Storage and Retrieval	Generation of Composite Data Files, and the Development of Data Archives

[3] For this discussion, the term "traditional" is intended to suggest personal (or face to face) interviews in which questions are read from a structured questionnaire and answers are recorded on the same printed form. Survey data are of course also collected without computers through telephone interviews and self-administered questionnaires. The above list is designed to suggest the range of "information processing" activities that were originally carried out without computers, as well as the ways in which computers were first used for those activities.

mation processing technologies, most of which simply converted a narrow set of activities to computer-based processing without affecting other areas or stages in the survey process. For example, programs for data management and statistical analysis were well established before packages became widely available for text processing or document preparation. For this reason, procedures for handling survey-related text – e.g, questionnaires, instructions, or reports – were typically developed on computers and by individuals who worked independently of those with responsibility for collecting or processing the resulting data. Similarly, procedures for documenting and archiving survey data were not effectively linked to major systems for statistical analysis, so that most survey analysts still cannot easily generate printed results that include the full text of the questions or procedures that were used to collect the data.

The data management tools required for large scale survey operations were also developed independently of statistical software. As a result, the character-oriented formats used for the collection and storage of survey data often conflicted with the requirements for numeric representation imposed by early statistical packages. Subsequent developments in both areas have converged in many respects, including compatible data formats as well as overlapping capabilities. Thus, several statistical packages can now handle more complex structures, and "database" systems have acquired some statistical capabilities. As discussed below, however, essential differences persist between modern systems for data base management and statistical analysis – and both differ significantly from software developed for questionnaire-based data collection and archiving – so that new strategies will be needed for linking or combining these separate technologies.

The most important changes now taking place in computer-assisted surveys stem from the revolution still taking place in data collection. During the 1970s, experimental programs were introduced which displayed questions and accepted responses at terminals operated by telephone interviewers. As described below, systems for Computer-assisted Telephone Interviewing (or CATI) quickly grew to handle several other types of "information" associated with telephone surveys, and the resulting systems have been generalized to the point where

Fig. 2. Information Technologies Used in Survey Research (in Approximate Chronological Sequence)

Initial Technology	*Extensions and/or Generalizations*
Unit Record Equipment: Devices for Processing Data on Punched Cards	Replacement of Punched Cards by Magnetic Tape and Disk Storage
Statistical Programs: Faster Computation Than Unit Record Equipment	Comprehensive Statistical Packages (e.g., BMDP, OSIRIS, SPSS, SAS)
Data Management: Utilities for Updating and Manipulating Files	General Purpose Systems for Relational Database Management
Text Processing: Using Computers to Process Text as Well as Numbers	Utilization of the Same Text in Questionnaires, Codebooks, and Reports
Data Capture: Separate Systems for Telephone Interviewing and Data Entry	General Systems for Data Collection Based on Structured Questionnaires

they can be used for other types of questionnaire-based data collection – and in a variety of computing environments. For an earlier collection of essays on computer-assisted data collection, see Freeman and Shanks (1983). See Nicholls and Groves (1986 a and b) for a more recent review of computer-assisted telephone interviewing, and see Shanks and Tortora (1985) for a discussion of the specific approach to CATI and its generalization being followed by members of the Association for Computer-Assisted Surveys.

2.1. *Stages in the development of computer-assisted surveys*

The resulting systems for data collection have also reached a stage in which survey researchers are considering the possibility – and the potential advantages – of integrating all of the information processing activities involved in the survey research process. For example, a computer-assisted telephone survey may rely on the same computing environment for document preparation (to create the interview schedule and interviewer instructions), data management (to handle survey information that is collected outside the interview context), and statistical computation (to describe progress or problems in sample completion) – as well as production interviewing.

The range of activities carried out in the same computing environment has encouraged speculation about a unified – or comprehensive – approach to the entire survey process, in which unnecessary duplication of effort might be eliminated without sacrificing any existing capabilities for data management, collection, or analysis. The ultimate objective for integration of this sort is a reformulation of the entire survey process, in which researchers will be able to concentrate on the content and quality of the resulting information – rather than on complications arising from the information processing environment.

Survey researchers have only recently started to work on this kind of technical integration. As in other fields, computer-based development projects in survey research can be assigned to one of three distinct stages, depending on whether they concentrate on:

- the development of initial programs for a specific applications and computing environment;
- the generalization of systems to related activities and alternative computing environments; or
- the integration of multiple systems for different activities, including linkages between systems based on different approaches.

At the present, however, most developmental projects in computer-assisted surveys can be classified in the first or second of these categories, for much remains to be done to improve and generalize systems for data collection, analysis, and management.

The following paragraphs concentrate on the improvement – and generalization – of systems for data collection, because those systems represent the only computer application that is unique (or indigenous) to survey research – and because data collection procedures will have a pervasive influence on the technical integration of the entire survey process.

2.2. *Computer-assisted data collection*

As suggested above, computer programs that were originally developed for telephone interviewing (CATI) have evolved into systems that handle many telephone survey activities in addition to administering the questionnaire. Because of that evolution, the same systems are also being generalized to handle a variety of other forms of data collection. The reasons for that generalization, and for the growing acceptance of CATI-type technology, can be seen in the comprehensive nature of the activities involved. Figure 3 reviews the ways in which telephone surveys may be affected by a CATI system. Figure 3 is similar to several published lists of CATI-related activities. See Shanks and Tortora (1985) for an earlier summary of this sort as well as references to other discussions.

Fig. 3. Telephone Survey Activities Affected by CATI

Preparation of the Interviewer's Instrument – drafting complete specifications for question content, question sequence or branching, and interviewer instructions, and entering those specifications into the computer;

Translation and Checking of the Interviewer's Instrument – transforming the computer-based instrument into a format which maximizes efficiency in interviewing, and checking all specifications for syntax errors;

Creation of Sample File(s) and Scheduling Instructions – creating a computer-based data set which contains a record for each case with telephone numbers and/or other identifying information, data from previous interviews, random numbers to control assignment to alternative question sequences, and information to be used in scheduling calls;

Study Management – producing periodic reports on study progress, interviewer performance, and sample completion, as well as assignment of calls to specific interviewers;

Production Interviewing – includes repeated dialing using assigned search patterns to establish contact with eligible respondents and the routing of problematic cases to supervisors for special handling, as well as actual interviewing;

Interviewer Supervision – resolving cases where interview attempts have been unsuccessful (through reassignment to language or refusal specialists, or to a final non-interview status), monitoring interviewer performance, and assisting interviewers on request;

Specification of Coding and Cleaning Procedures – preparing instructions to editors (or coders) and to the computer to control any checking, cleaning, or supplementary data entry which should take place after each interview is complete;

Conversion and Checking of Coder's Instrument – a process which may resemble translation of the interviewer's instrument (above) if the instructions for cleaning (or coding) are stored in the same format;

Production Coding and Cleaning – assigning coded values to unstructured text associated with open-ended questions or "other specify" responses, and resolving any inconsistencies between recorded responses and the logic of the (coding and cleaning) instrument;

Certification and Output for Completed Cases – final checking for errors in the data and transferring satisfactory cases to an output file for analysis;

Data Analysis and Documentation – using the information in the interview (or coding) instrument to produce explanatory text for statistical reports and final survey documentation.

Since the first CATI systems were introduced, lists of this sort have suggested that the new technology might incorporate (and thereby integrate) activities that were traditionally handled by separate groups or staffs. In particular, the computer-based "instrument" that controls a given CATI application may include instructions for activities that were traditionally carried out separately by specialists in: questionnaire design, sampling, interviewing, coding, supervision, data preparation, analysis, and archiving. By incorporating instructions for several of these (previously separate) activities into a single computer-based instrument, CATI projects can make several information processing activities do double or triple duty. The following examples illustrate this (now familiar) potential for consolidating previously distinct activities:

- the same computer-based files may be used to define the sample, control the sequence in which cases are assigned to interviewers, and provide documentation concerning the progress or history of data collection for each case;
- the test and logic of the interviewers' instrument may be converted into a parallel instrument for controlling all post-interview data entry and definition, as well as documenting the final dataset;
- answers or response patterns that are defined as illegal need not be corrected after initial data collection, for such errors are detected (and resolved) during the interview;
- the same instrument may be used to ensure that all appropriate questions are answered, even if the interviewer (or coder) has deviated from the prescribed question order by skipping ahead, moving backward, or changing an answer;
- no separate process is needed to convert interview responses to machine-readable form, since all data (including precoded responses and verbatim text) are "captured" during the interview through direct keyboard entry.

The potential advantages of this kind of consolidation are responsible for the rapid growth and dissemination of CATI systems. Continuing growth in both capabilities and usage has also led to the generalization of such systems, both to multiple computing environments and to other forms of data collection. In addition to this process of generalization, however, CATI systems are still being changed frequently, for much remains to be done before telephone surveys make efficient use of current technology for all the activities mentioned above.

For example, the CSM program is currently concentrating on several CATI-related enhancements to the Computer Assisted Survey Execution System (or CASES), including: computation and storage for multiple types of variables (including floating point), additional kinds of screens and forms-type processing, automatic scheduling of telephone interviews, transfer of cases between computers (for distributed data collection), more efficient data storage, and "help" facilities to make it easier to use all of the programs involved. These new capabilities are sometimes released individually, to meet the needs of specific projects or users, but they are usually combined into major versions or releases. As of this writing, CASES users are testing Version 3.3E, and plans exist for three more (major) versions before all of the currently scheduled enhancements are completed. Informal reports suggest that other data collection systems (besides CASES) are going through a similar process of revision or enhancement, so that many survey organizations experience frequent changes in their computer-assisted data collection procedures.

2.3. Generalization to alternative types of data collection

The above kinds of changes represent important enhancements for many CATI users, but they have had to compete for developmental resources with a quite different set of objectives based on the general nature of the activities involved. The breadth or diversity of any system's user community is an important determinant of the resources it can devote to development and maintenance. For that reason, and at the request of specific users, several CATI systems have been revised so that the same program can be used in an wider variety of contexts.

The first of these types of generalization (and revision) stems from the understandable desire of many survey organizations to use the same kind of procedures for projects which use different types (or modes) of data collection, and for single projects that must use more than one of those modes, including:

- Computer-assisted Personal Interviewing (CAPI),
- Self-Administered Questionnaires (SAQ), for Respondent-Entered Data, and
- Direct Data Entry (DDE), for Paper-and-Pencil Forms, as well as
- Computer-assisted Telephone Interviewing (CATI).

Two of these extensions (CAPI and SAQ) are currently limited because of their requirement that respondents (or subjects) be brought into direct contact with a computer.[4] With the continuing improvement in portable computers and communications, however, self-administered options may become much more important for several types of research. For example, computer-based questionnaires are already being administered on the telephone without an interviewer. In this approach, questions are presented through voice reproduction to respondents who call a designated number on a touch-tone phone. The respondent then answers the (voice reproduction-based) questions by entering numeric codes on the phone. This technique is called touch-tone data entry (or TDE) by researchers at the Bureau of Labor Statistics, where it is being used as an alternative form of data capture (to be combined with telephone interviews conducted in CASES) for the Current Employment Survey. (See Working, Tupek, and Clayton (1988).) Also, voice- and graphics-oriented options will soon be available for applications in which the respondent (or subject) can interact directly with a computer, instead of over the telephone. By simply "calling" other programs or devices, structured questionnaires may soon take on a very different character, as the concepts of "question" and "response" expand to include both images and sounds.

2.4. Generalization for distributed data collection

Most CATI systems were originally developed for a single (multi-user) computer, in which interviewers sat in front of terminals connected to the computer by direct lines or telephone.

[4] The generalization of CATI to Computer-assisted Personal Interviews (or CAPI) is still a moderately recent development, and several approaches are being explored to integrate the (computer-assisted) questionnaries in projects which call for both CATI and CAPI. Organizations working with CASES are developing similarly structured (but separate) instruments for each data collection method, based on the assumption that differences between modes (in instructions or logic) cannot be handled in a fixed or system-prescribed fashion. In contrast, the BLAISE system being developed by the Central Bureau of Statistics in the Netherlands can be used to produce a single instrument that is processed in a different way for CATI, CAPI, or direct data entry (DDE) based on paper forms. See Denteneer, Bethlehem, Hundepool, and Keller (1987).

For some time, the only exception to this rule was the Wisconsin system for micro computers, but PC-oriented survey organizations can now choose between many systems or approaches.[5] Increasingly, however, many survey projects require that data collection facilities be "distributed" across several computers, in one or more of the following ways:

- personal (or single user) computers are used for interviewing, but all of the data is maintained by a single file server over a local area network;
- a multi-user system serves as a satellite to a master (or host) computer (i.e., hierarchical relationships may exist between multi-user computers within a single facility); or
- computing facilities that are geographically (and organizationally) separate must be centrally coordinated for a specific project.

Within the CSM user community, each of these approaches to distributed data collection is already in use. The computer programs involved, however, need substantial changes to more effectively carry out (and check) the inter-system communications involved. The general problems of maintaining study-level integrity during data transfer between computers can be particularly severe when the two (linked) systems have different hardware and operating systems. An early requirement, therefore, for some applications has been that the programs involved function the same in several computing environments.[6]

The most important developmental tasks, however, have only begun, for data collection systems need more sophisticated protocols for transferring information from study-level databases between computers – regardless of the hardware and operating systems involved.[7]

3. Data Management and Survey Integration

Since the mid 1970s, survey organizations have concentrated most of their resources for computer-related development on general-purpose systems for collecting data based on structured questionnaires. The resulting programs are no longer new or experimental, for they are routinely used for "production" data collection in projects based on self-administered questionnaires as well as telephone and personal interviewing. As indicated in the previous section, however, much remains to be done in improving and generalizing those systems before they satisfy all of the objectives which have been identified by their user communities.

While those systems are still being improved, survey organizations have also become interested in the capabilities for handling large and complex data structures offered by systems that were developed for management – rather than the collection or analysis – of survey-type information. Survey activities which may call for a separate data management system include:

- questionnaires with more complex relationships than simple hierarchies (including "many-to-many" relationships like those between multiple patients and doctors);

[5] See Palit and Sharp (1983) for a statement of objectives for the Wisconsin system. Other data collection systems for the PC environment include those produced by Sawtooth, Inc. and Computers for Marketing, Inc., as well as CASES and BLAISE.

[6] For example, CASES programs have been converted for use in VMS (for VAX systems produced by the Digital Equipment Corporation) and a variety of UNIX systems in addition to personal computers that use PC- or MS-DOS. Work has also begun on an MVS version for IBM-compatible mainframe systems, and a version is planned for MacIntosh (Apple) computers.

[7] See Statistics Sweden (1989) for a description of their approach to distributed data collection which involves personal computers in interviewers' homes.

- interlocking surveys or "studies" in which more than one questionnaire may be administered in one interview with the same respondent;
- the management or allocation of data collection resources between multiple (simultaneous) surveys, and the measurement of staff performance across survey projects;
- the creation of large or complex datasets by combining information from multiple (survey and non-survey) sources; and
- the creation and maintenance of data archives, or comprehensive collections of datasets and documentation for a large number of surveys in a general area.

For these and other reasons, survey researchers are now exploring the potential benefits of "database" technology in managing complex data structures and integrating information from multiple sources – while continuing to rely on existing systems for data collection, analysis, and documentation.

3.1. Distinguishing data management from data collection and analysis

The number of different sources of survey-related "information," and the relationships between those sources, present a classic illustration of the circumstances in which an organization may benefit from using a relational database management system (or RDBMS). For example, a survey organization may already be maintaining separate computer-based files containing information about the following kinds of "entities" in addition to the data being collected or analyzed:

- past instruments (including both question wording and interviewer instructions);
- staff members (including employment history, hourly costs, and previous performance, as well as hours spent on each current project);
- sample elements (including information about unused cases as well as those assigned to current studies);
- multiple projects (including administrative information such as planned expenses vs. actual costs, as well as personnel plans and time schedules); and
- completed datasets (including documentation concerning data type and location of variables, time period, access permissions, and relevant publications).

A single data base application could include files for each of the above types of entities, so that users in one area (or project) would have access to information that was originally collected for other purposes. In applications of this sort, the data management system must support all of the linkages or relationships involved (e.g., between such entities as projects, instruments, staff members, sample elements, datasets, and variables), and it must permit users to define their own reports for retrieving and displaying information. The central concept in relational data base technology is the decomposition of any application into a series of simple or rectangular datasets (one for each type of entity), each of which is linked to other datasets through relationships between the entities involved, such as membership in the same family or data collection project.[8]

While survey researchers explore a variety of RDBMS applications, basic systems will continue to improve for data collection and analy-

[8] See Codd (1970) for an influential summary of design principles for relational database management systems. See Baker (1987) for a lucid account of the ways in which these concepts can be used to improve the management of traditional survey operations (based on household samples and face to face interviews).

sis, as well as database management. Most of the gains in survey-related computing will therefore continue to be extensions of single-purpose systems, i.e., as additional features in packages that were originally designed for data collection, analysis, *or* management – rather than entirely new systems which carry out all three kinds of activities.

To some extent, the boundaries between systems for data collection, management and analysis are becoming less distinct, for the major packages in each area have acquired capabilities in other areas without sacrificing the integrity of their original applications. Thus, statistical (or analysis) systems have acquired options for data entry and handling non-rectangular stuctures, and database management systems can be used for data entry and statistical calculations as well as displaying characteristics of individual cases. Similarly, as discussed above, systems for computer-assisted telephone interviewing (CATI) have been adapted for other (non-telephone) forms of data collection, and may include capabilities for statistical analysis and data management. This expansion of existing systems across the three basic "stages" (collection, management, and analysis) will continue for some time, but it is unlikely to produce a satisfactory computer-based integration of the entire survey research process. Experience to date suggests that the combined set of information-processing activities in projects based on structured questionnaires is extremely diverse. No single system (for data collection, analysis, *or* management) will soon reach the point where it provides all of the capabilities required.

As suggested in the introduction to this essay, each stage in the survey process is characterized by its own information processing requirements and complexities, many of which have been "handled" by simply ignoring information that is essential at other stages. For example, data collection procedures rest on complex instructions concerning the sequence in which steps are to be taken (or repeated) as well as voluminous instructions to staff members involved in the data collection process. Current practices in the management and documentation of survey data incorporate only a portion of those instructions, and almost all of that detail is discarded when creating analysis files for most statistical packages. Similarly, data management systems emphasize the relationships between entities and fields in different files, but such systems are usually intensive to the sequence in which data values should be entered or (re-)calculated – and they retain very little of the information about data content or the collection process that is typically included in survey data documentation (or codebooks).

In effect, each type of system has concentrated on a distinctive type or aspect of survey data processing, while disregarding information and logic which may be crucial at other stages of the research process. These differences, or simplifications, have made it easier to develop our existing systems for data collection, analysis, and management. The resulting differences, however, can make the databases produced by these separate technologies very difficult to integrate or combine.

3.2. Barriers to integration: Differences between stages and systems

Within any survey organization, one of the most frequent kinds of computer-related irritation arises from the utilization of several different systems for data collection, analysis, or management. As specific systems are generalized to the point where they can be used in different operating systems (and computing hardware), some difficulties in transferring data between systems may disappear – for the same computing environment can be used for different activities or stages. Barriers to data transfer or integration, however, can still arise when all activities (i.e., data collection, analysis, and management) rely on the same computing environment, because of dissimilarities between the data structures used by packages that were originally developed for different purposes. In particular, systems for data collection, management, and analysis rest on quite different strategies for representing the instructions used to collect survey information and the content (or structure) of the resulting data.

This problem is compounded by the existence of alternative systems within each of the major stages in the survey research process. Survey organizations now rely on a substantial number of alternative systems for data collection and analysis, and several relational database management systems are now being evaluated for handling large or complex data structures. Any comprehensive list of systems now in use in the United States would include more than a dozen packages for CATI-type operations and many more than that for statistical analysis. In the foreseeable future, it is unlikely that any single system (for collection, analysis, *or* management) will grow to the point where it includes all of the capabilities needed in the other two areas, and multiple alternatives are certain to exist within each area for many years to come. Given that prospect, the inherent problems of integrating survey activities between the three major stages of the research process are compounded by the sheer number of system-to-system linkages involved.

To be sure, a single survey project may use only one system for data collection, one for statistical analysis, and a third for database management, so that it might need only two or three (bilateral) conversions of data from one system to another. Many organizations, however, use more than one package in some areas, and each system-to-system linkage (or conversion) can involve a substantial investment in software development. As a consequence, the developmental effort required for a single organization to transfer information between systems may be quite large – and must be repeated in organizations that use other combinations of packages – unless a more general solution to the problem can be found.

3.3. Alternative strategies for technical integration

Attempts to transfer computer-based survey information between systems encounter a variety of difficulties, based on differences in data structure between alternative systems in the same general category (or stage) as well as general differences between stages in the type of survey information involved. Most such problems can of course be "solved" through additional programming, but such projects can be very time-consuming and expensive. For this reason, researchers in several organizations have expressed interest in a general-purpose approach, so that solutions developed in one context can be used by other projects or organizations.

The barriers to linking or combining computer-based information maintained by different systems (for data collection, analysis, and management) represent what might be called the "last frontier" in developing a comprehensive computer-based environment for survey research. During the next several years, research and development projects will explore a variety of approaches to coordinating or integrating information-processing activities across the major stages of the survey research process. All such efforts can be associated with one of three basic strategies, with fairly obvious differences in costs and risks:

- Select a single system for each major phase and build linkages or translation programs to move information from each specific system (and stage) to the others (a *one-to-one* strategy for integration);
- Develop a single *comprehensive system* for all stages of the research process (a goal whose

scope has sometimes been described as analogous to an "airlines reservation system" for survey research);

- Develop general-purpose or system-neutral procedures for *data description*, so that users could move data from one system to any other system that uses the same external structure for data description (a *many-to-many* strategy for integration).

The high cost of the first of these strategies has already been discussed above. Until some other plan is successful, however, bilateral (one-to-one) linkages will continue to be the only solution.

At this point, it seems highly unlikely that any project or group will be successful in pursuing the second strategy, i.e., developing a single system which offers all of the services required for the collection, analysis, and management of survey-related information. Current systems in each area are based on internal structures that will be very difficult (or time-consuming) to reproduce within a system that also covers the other stages. Despite those obstacles, a comprehensive system for all survey-related activities represents an important long-term challenge, so that some researchers should seek the (substantial) resources that will be required to *design* such a package. If and when such a design is completed, it should be carefully evaluated before any actual development takes place, for the costs associated with a false start could be extremely large. In the meantime, however, it seems safe to assume that no comprehensive system will soon emerge which covers all three aspects or stages of the survey research process.

3.4. Cross-system integration through data description

That assumption, coupled with the recurrent need for transferring data between alternative systems for data collection, management, and analysis, suggests that our field might benefit from a common (or neutral) standard for storing and documenting the data produced by survey procedures. In such an approach, all cooperating systems for collection, management, or analysis would accept input data (and documentation) that have been stored in a common (or standard) format and could produce output data (and documentation) in that same format. Each such system would then only need to convert data to and from that common (neutral) format, rather than developing a different conversion program for each other system. The survey field has not agreed on a format standard for "system-neutral" data description, but discussions take place from time to time concerning potential alternatives.

As a first step in this direction, the CSM Program is developing a Data Description Language (DDL) which could become a common format for transmitting survey data and documentation between the several types of systems discussed above. Aspects of the intended language are already used to describe input and output data for CSM programs for Conversational Survey Analysis (CSA),[9] and procedures are now being completed to automatically convert data and Q instruments from a CASES project to the same (DDL) format, and to automatically generate setups from DDL for other statistical packages, (e.g., SPSS and SAS). This approach to data conversion between systems is summarized in Fig. 4.

The current DDL includes only those data elements required for CSA, but a comprehensive language of this sort should include documentation for survey-based information produced by all of the activities described at the beginning of this essay. Several survey organizations (including NASS) are considering a "database ap-

[9] The objectives and current status of Conversational Survey Analysis are beyond the scope of this paper. For a brief discussion of CSA's design and capabilities, see the CSA User's Guide (CSM Staff, 1989).

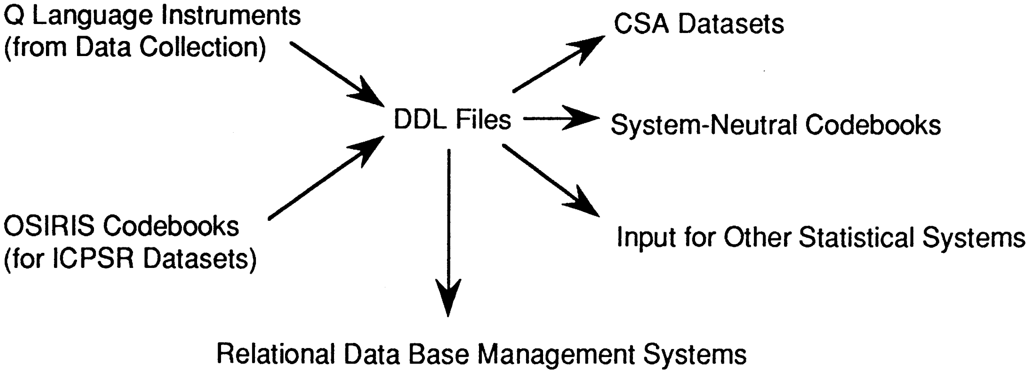


Fig 4. A "Neutral" Format for Transferring Data Between Systems Sources and Uses for DDL Files

proach" to managing all of their surveys, but it will be some time before specifications are completed for all of the data elements and structures for a demonstration project of that sort. See Tortora, Vogel, and Shanks (1985).

3.5. Alternative approaches to integration: Combining data collection and management

The above approach – based on a standard format for data description – represents only one of several strategies for overcoming the incompatibilities between alternative systems for data collection, analysis, and management. As suggested above, at least one group of specialists (in surveys and computing) should begin the process of designing a single (or comprehensive) system which provides all of the services required. A variant on that approach is now receiving increasing attention within organizations that must manage large and complex collections of survey data. In that approach, a relational database management system provides a common database environment (and computational capabilities) within which other programs can be accessed for either data collection or analysis.

Specifically, survey specialists from several organizations have advocated that data collection packages be revised to read and write all of their files in the internal format required by ORACLE (or SYBASE, INGRES, DBII, etc.), so that these study-level files can be integrated with other types of information and large collections of data sets. This strategy may be represented by the following (quite different) relationship between the various stages in the survey process:

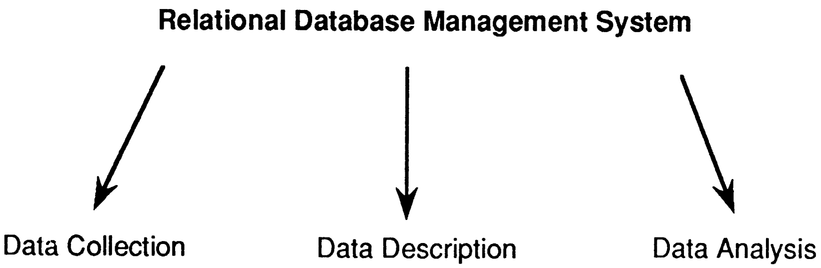


Fig. 5. Relational Database Management System

In such a design, the dominant status of a database management system should not be mandatory, for users must still be able to combine or link existing programs for collection, description, and analysis – as described above. Several organizations, however, are experimenting with a comprehensive "database approach" of this sort, in which all survey information is managed by a single relational system. NORC is already working on such a system (private communication with R. Baker 1989), and CSM is discussing a "relational" version of CASES with NASS and BLS.

This essay is neutral with respect to the comparative difficulty or long-term effectiveness of these alternative strategies, and encourages a variety of organizations to at least design the projects that will be required to answer such questions. No matter which approach emerges as the most effective strategy for integrating the survey process, however, the concepts and techniques used in RDBM systems are likely to play an important role. This article agrees with "database" theorists who have argued that complex data structures are best handled by decomposing those structures into a series of simpler (rectangular) files, and by representing the complexity involved in terms of relationships between those files. That perspective suggests that survey researchers will be more successful in integrating their diverse computer-related activities if all of their survey information is represented by (multiple) rectangular files and relationships (or linkages) between files. Development projects that deviate from this principle should be less successful in the long run, and this expectation should become increasingly important as the survey activities involved become more complex and comprehensive.

3.6. *The impact of new technology*

Since the late 1970s, many survey organizations have changed their internal division of labor in response to new systems for computer-assisted data collection. As emphasized in previous sections of this article those changes have involved a steady increase in the concentration of technical responsibility, as computer-based "instruments" specify more and more of the survey procedures involved. Thus, the interview schedule in a computer-assisted survey often defines sample elements and outcomes, and the same instrument may be augmented to control post-interview data entry or revision and document the resulting data. The substantial overlap in machine-readable information between stages and activities, coupled with improved procedures for transferring data between different systems (for data collection, analysis and management) have encouraged the integration of all computer-related activities for a given survey under a single study director. After more than a decade of experimentation and development, this trend toward computer-based integration (of previously separate survey activities) is now well underway, and recent developments in computing technology will almost certainly facilitate that process.

In particular, survey professionals are now discovering the possibility of visual integration of previously separate activities, based on high performance workstations. In this new computing environment, individuals who are responsible for coordinating survey activities across previously separate stages can control those tasks as simultaneous "windows" on a large screen attached to a computer that resides on their desk. These workstations offer a noticeably large display (at least 19 inch) with much greater resolution (more than a million pixels), a larger memory (several million bytes), and much more computing power (several million instructions per second) than in the environment that most of us have used until quite recently. These capabilities will almost certainly accelerate the integration of activities previously carried out by separate individuals, for a single study director can now move quickly between windows in which individuals can carry out a variety of simultaneous operations. This kind of "workstation integration" is suggested in the screen shown on the following page, which contains a separate window for questionnaire administration, questionnaire development (or modification), statistical analysis, and identifying cases with specific characteristics.

The technical integration of survey activities is also being encouraged by the changes now taking place in communications between computers. Until fairly recently, survey information could only be processed on a given computer by physically moving all of the files to that system. With the growth of high-speed networks and distributed file systems, however, survey researchers can use their own (local) computer to process information that is stored on different (remote) machines. Survey researchers will soon be using high-speed communications and distributed file systems in a variety of contexts, including:

- immediate access to large databases from previous surveys, for re-analysis or comparison with current results;
- rapid movement of information between geographically separate systems for data collection and analysis;
- use of inexpensive workstations for data collection, so that larger computers in the same local network can be dedicated to data storage and retrieval; and
- "online" access to other computers during data collection, for circumstances in which information must be retrieved from large (external) databases before determining the next appropriate question.

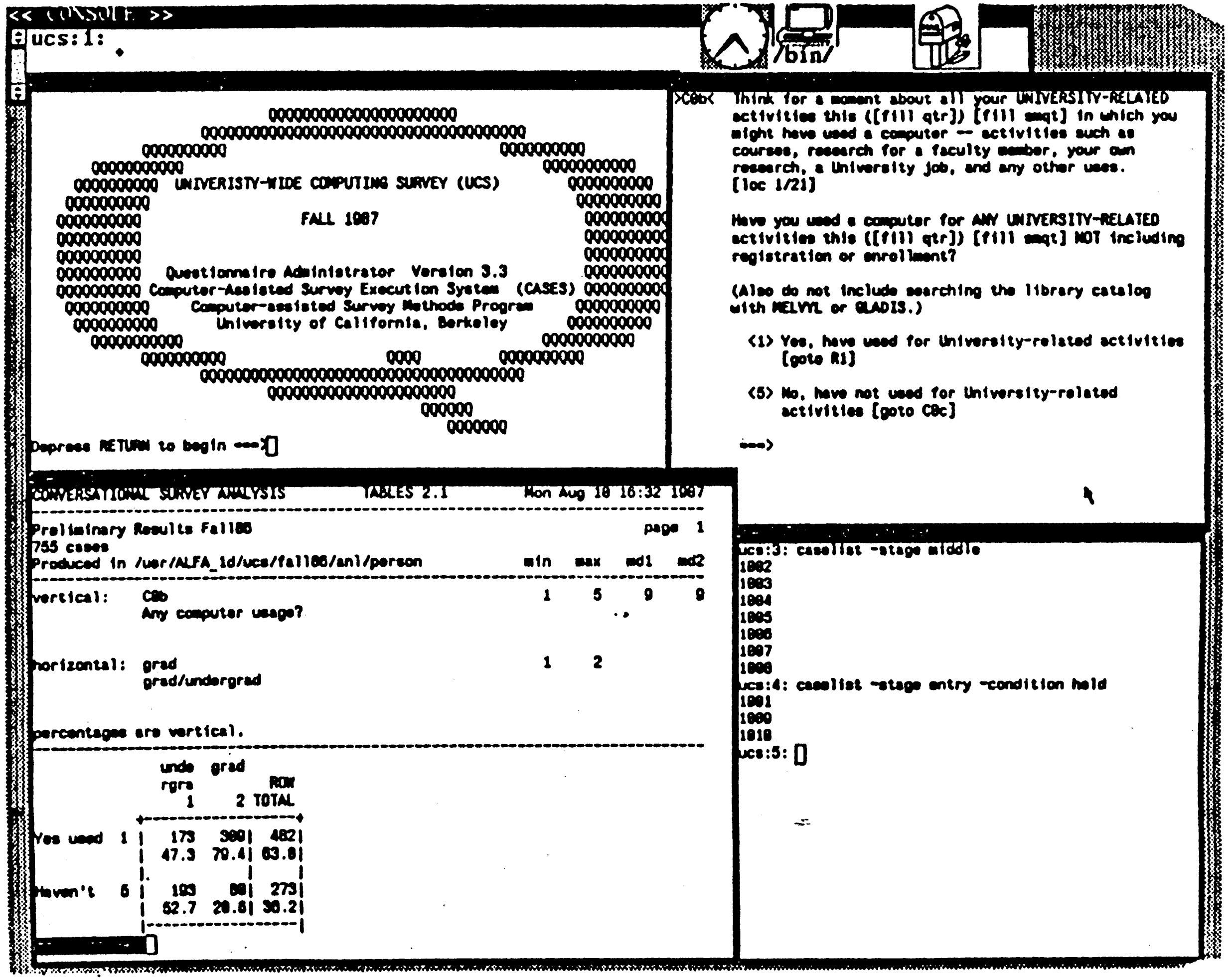
As emphasized in previous sections of this essay, the concept of a "structured questionnaire" is already changing, because of the personal computer's new capabilities for input and output (including images and sound). When coupled with advanced function workstations and highspeed communications, those capabilities will also contribute to the general trend toward survey integration, for new forms of data collection will present corresponding challenges for data management and documentation.

3.7 Summary

During the next few years, research and development in the survey field will continue to involve the improvement and generalization of separate systems for data collection, analysis, management, and documentation. Much remains to be done in each of these areas in order to take advantage of the developments now taking place in information technology, including workstations, graphics, large scale databases, and high-speed communications. In addition to these new capabilities, however, survey researchers will be exploring alternative approaches to integrating the entire research process. The so-called "database approach" has been enormously successful in providing integrated systems for collecting, managing, and displaying information in many other fields. The process of collecting survey-type data, however, is different from the activities handled by existing database management systems, and our combined requirements for "information processing" exceed those likely to be provided by current systems for data collection, analysis, or management.

As a consequence, survey researchers are now designing, developing, and testing a variety of ways to combine or coordinate their use of all three types of systems. At this point, we can only speculate about the comparative merit of those approaches. Hopefully, a sequel to this essay will identify strategies which have been successful in providing a unified (and simplified) approach to "information processing" in survey research.

THE SURVEY RESEARCHER'S WORKSTATION



```
ucs:1:
```

```
                  00000000000000000000000
              0000000000000000000000000000000000000000
       0000000000                                 0000000000
    00000000000                                       00000000000
   0000000000  UNIVERISTY-WIDE COMPUTING SURVEY (UCS)      0000000000
  0000000000                                                0000000000
 0000000000                  FALL 1987                       0000000000
 0000000000                                                  0000000000
 0000000000                                                  0000000000
 0000000000   Questionnaire Administrator  Version 3.3       0000000000
 0000000000 Computer-Assisted Survey Execution System  (CASES) 0000000000
  0000000000     Computer-assisted Survey Methods Program     0000000000
   0000000000      University of California, Berkeley        0000000000
    00000000000                                       00000000000
       0000000000                 0000            0000000000
              0000000000000000000000000000000000000000
                  00000000000000000000000
                                     000000
                                        0000000
Depress RETURN to begin --->[]
```

```
>C8b<  Think for a moment about all your UNIVERSITY-RELATED
       activities this ([fill qtr]) [fill smqt] in which you
       might have used a computer -- activities such as
       courses, research for a faculty member, your own
       research, a University job, and any other uses.
       [loc 1/21]

       Have you used a computer for ANY UNIVERSITY-RELATED
       activities this ([fill qtr]) [fill smqt] NOT including
       registration or enrollment?

       (Also do not include searching the library catalog
       with MELVYL or GLADIS.)

         <1> Yes, have used for University-related activities
             [goto R1]

         <5> No, have not used for University-related
             activities [goto C8c]

         --->
```

```
CONVERSATIONAL SURVEY ANALYSIS        TABLES 2.1       Mon Aug 10 16:32 1987
-----------------------------------------------------------------------------
Preliminary Results Fall86                                            page 1
755 cases
Produced in /usr/ALFA_1d/ucs/fall86/anl/person          min   max   md1   md2
-----------------------------------------------------------------------------
vertical:     C8b                                         1     5     9     9
              Any computer usage?


horizontal:   grad                                        1     2
              grad/undergrad


percentages are vertical.
-----------------------------------------------------------------------------

                  unde  grad
                  rgra          ROW
                     1     2  TOTAL
                +---------------+
Yes used     1  |   173   309|  482|
                |  47.3  79.4| 63.8|
                |.           |     |
Haven't      5  |   193    80|  273|
                |  52.7  20.6| 36.2|
                |------------------|
```

```
ucs:3: caselist -stage middle
1002
1003
1004
1005
1006
1007
1008
ucs:4: caselist -stage entry -condition held
1001
1009
1010
ucs:5: []
```

4. References

Baker, R. P. (1987): Information Systems in Survey Research. Proceedings of the Bureau of the Census Third Annual Research Conference, Baltimore, Maryland, March 29–April 1, pp. 166–177.

Codd, E. F. (1970): A Relational Model of Data for Large Shared Data Banks. Communications of ACM, 13(6).

CSM Staff (1987): User's Guide to CASES; The Computer-Assisted Survey Execution System. University of California, Berkeley, Computer-assisted Survey Methods Program (CSM).

CSM Staff (1988): User's Guide to Conversational Survey Analysis (CSA). University of California, Berkeley, Computer-assisted Survey Methods Program (CSM).

Denteneer, D., Bethlehem, J. G., Hundepool, A. J., and Keller, W. J. (1978): The BLAISE System for Computer-assisted Survey Processing. Proceedings of the Bureau of the Census Third Annual Research Conference, Baltimore, Maryland, March 29–April 1, pp. 112–127.

Francis, I. (1981): Statistical Software: A Comparative Review. Elsevier, North Holland.

Freeman, H. E. and Shanks, J. M., eds. (1983): The Emergence of Computer-assisted Survey Research. Sociological Methods and Research, 12(2), pp. 115–118.

Nicholls, W. L., II and Groves, R. M. (1986a): The Status of Computer-assisted Telephone Interviewing: Part I – Introduction and Impact on Cost and Timeliness of Survey Data. Journal of Official Statistics, 2(2), pp. 93–115.

Nicholls, W. L., II and Groves, R. M. (1986b): The Status of Computer-assisted Telephone Interviewing: Part II – Data Quality Issues. Journal of Official Statistics, 2(2), pp. 117–134.

Palit, C. and Sharp, H. (1983): Microcomputer-assisted Telephone Interviewing. Sociological Methods and Research, 12(2), pp. 169–189.

Raskins, R. (1989): Statistical Software for the PC: Testing for Significance. PC Magazine, 8(5), March 14, pp. 103–116.

Shanks, J. M. (1983): The Current Status of Computer-assisted Telephone Interviewing: Recent Progress and Future Prospects. Sociological Methods and Research, 12(2), pp. 119–142.

Shanks, J. M. and Tortora, R. D. (1985): Beyond CATI: Generalized and Distributed Systems for Computer-assisted Surveys. Proceedings of the First Annual Research Conference of the Bureau of the Census, (March).

Sonquist, J. (1977): Computing and Surveys. Prentice Hall, Englewood Cliffs, New Jersey.

Statistics Sweden (1989): Computer-assisted Data Collection in the Labour Force Survey. A Report of Some Technical Tests. Technical report.

Tortora, R. D. (1984): CATI in an Agricultural Statistics Agency. U. S. Department of Agriculture, Statistical Reporting Service.

Tortora, R. D., Vogel, F. A., and Shanks, J. M. (1985): Computer-Aided Survey Methods. U. S. Department of Agriculture, Statistical Reporting Service.

Werking, G., Tupek, A., and Clayton, R. (1988): CATI and Touchtone Self-Response Applications for Establishment Surveys. U. S. Bureau of Labor Statistics. Paper presented at the Fourth Annual Census Bureau Research Conference, Arlington, Virginia (March 20–23).

Received August 1988
Revised April 1989